Beyond the ribosome: proteome-wide secretability studies with SECRiFY

Boone, M.[1,2,7]; Ramasamy, P.[1,3,4,5,6], Maddelein, D. [1,3], Turan, D. [1,3], Vandermarliere, E.[1,3], Vranken, W.[4,5,6], Callewaert, N.[1,2]

1. Center for Medical Biotechnology, VIB, Zwijnaarde, Belgium

2. L-ProBE, Department of Biochemistry and Microbiology, Faculty of Sciences, UGent, Ghent, Belgium

3. Department of Biochemistry, Faculty of Medicine and Health Sciences, UGent, Ghent, Belgium

4. Structural Biology Brussels, VUB, Brussels, Belgium

5. Structural Biology Research Center, VIB, Brussels, Belgium

6. Interuniversity Institute of Bioinformatics in Brussels (IB)^2, ULB-VUB, Brussels, Belgium

7. Present address: Department of Biochemistry and Biophysics, UCSF, San Francisco, CA, USA

Abstract

While transcriptome- and proteome-wide technologies to assess processes in protein biogenesis up to the ribosome-associated stages are now widely available, we still lack global approaches to assay post-ribosomal processes, in particular those occurring in the eukaryotic secretory system. We developed SECRiFY to simultaneously assess the secretability of >10^5 heterologous protein fragments by two yeast species, *S. cerevisiae* and *P. pastoris,* using custom fragment libraries, surface display and a sequencing-based readout. SECRiFY generates datasets that enable datamining into protein features underlying secretability, and it will enable studies of the impact of secretory system perturbation on the secretable proteome.

Main text

The eukaryotic secretory system processes roughly a quarter of the proteome, ensuring correct folding, assembly, and delivery of proteins to the extracellular environment, the plasma membrane, or membrane-bound organelles[1–3]. Many proteins and processes involved in these pathways have been identified through studies using just a few model secretory cargos, but an integrative understanding of their role in enabling production of the thousands of secretory proteins is still lacking. For example, it is generally unknown which chaperones are critical for assisting the folding of which types of secretory protein domains. Unfortunately, most current approaches are unsuited to study a comprehensive range of protein folds after entry in the ER. In practice, the absence of such an integrated picture of the secretory system is most apparent in the unpredictability of heterologous protein secretion. Indeed, four decades after the recombinant DNA revolution, obtaining detectable levels of functional protein in a particular heterologous host system (secretory or not) is still principally a process of trial and error. This slows down progress in the many fields of basic and applied life science where recombinant protein production comes into play. Although expression screening of parallel constructs or variant libraries of a protein of interest has gained momentum[4–6] to increase heterologous protein expression success rates, it has to be repeated for each new target. Additionally, alternative comprehensive strategies to assess heterologous expression across entire proteomes have generally been limited to intracellular expression in *E. coli*, small proteomes, or clone-by-clone strategies[7–10].

To assay the secretory potential of (human) protein fragments in yeast on a proteome-wide scale, we combined yeast surface display of a novel type of protein fragment libraries and deep sequencing, and dubbed the platform SECRiFY (SECretability screening of Recombinant Fragments in Yeast) (**Fig. 1a** and **e**). By focusing on domain-sized protein fragments rather than full-length proteins, we avoid missing detection of secretability of parts of multi-domain proteins that fail to express or secrete in their entirety due to local issues with misfolding of particular protein areas, translation inhibitory sequences, protease susceptibility, the absence of stabilizing interaction partners or modifications, or toxicity. Chopping up difficult proteins in experimentally tractable fragments has been exploited by structural biologists for years, both in rational target design as well as in random library screens for soluble expression[11–15]. Considering the notorious inaccuracies of domain boundary prediction[16], and the knowledge that, even with a reliable estimate, small variations in the exact N- and C-terminus of the fragment can lead to dramatic differences in expressability[16], we opted for a random fragmentation approach. We designed and built directional, randomly fragmented cDNA libraries covering the human transcriptome with fragments coding for approx. 50-100 amino acids, which is the median domain size of human proteins (**Fig. 1b, Suppl. Fig 1a**). To reduce fragment abundance differences inherent to the large dynamic range of mRNA transcripts in human cells, we normalized this library using the Kamchatka crab duplex-specific nuclease[17,18] (**Suppl. Fig. 1b**). Careful optimization of the random primer tag sequence for compatibility with normalization effectively minimized the required library size by approx. 1000-fold, to within a range that is feasibly reconcilable with the diversity that can be routinely obtained in cDNA library construction and yeast transformation (+/- $5*10^6$ - $5*10^7$) (**Fig. 1c-d, Suppl. Results**). To our knowledge, this is the first time that an effective method for normalization of tagged random-primed cDNA fragment libraries is reported, and required substantial development (see also **Suppl. Fig. 1c**). This new method should also find many applications in areas where the protein-coding potential of a cell needs to be effectively covered in expression libraries.

Relying on the sophisticated quality control (QC) machinery of the eukaryotic secretory system, which ensures efficient degradation of unstable or misfolded proteins[19,20], we further reasoned that surface display could be used as a proxy for productive secretion. As such, once cloned into the surface display vector and transferred to yeast, library polypeptides are directed to the secretory system by an N-terminal secretory leader sequence (MFα1 prepro), and furthermore on the yeast cell wall via C-terminal fusion to the GPI-anchoring region of the *S. cerevisiae* cell wall protein Sag1 (**Fig. 1a**). Fragments for which the fragment-Sag1 fusion successfully passes (or escapes) secretory system QC without proteolytic degradation are recognized through their N- and C-terminal epitope tags (FLAG and V5, resp.), and are segregated from the rest using iterations of high-efficiency magnetic- and fluorescence-activated cell sorting (MACS/FACS) (**Fig. 1e**). Finally, fragment identification and classification is achieved by deep sequencing of recovered fragment amplicons from the unsorted and sorted cell populations (**Fig. 1e and Suppl. Fig 2**).

We first benchmarked the method in *S. cerevisiae*. Triplicate secretion screening of a 1.96*10^6 clone fragment library of the HEK293T transcriptome in this yeast revealed that on average 1.76% ± 0.12% of library cells displayed a fragment with an intact N-terminus (FLAG-tag) and intact C-terminus (V5-tag) (**Fig. 1e**, **Suppl. Fig 3**). Accounting for a 1/9 chance of up- and downstream in-frame cloning, this means that approximately 15.8% of in-frame fragments are detectably displayed and hence, potentially secretable. After a 32-fold enrichment of these double positive cells through a single round of MACS and two subsequent rounds of FACS (**Fig. 1e**, **Suppl. Fig. 3**), followed by sequencing of the final sorted population, on average, 1.12*10^6 unique fragments/replicate were detected, covering on average 26.45% ± 0.86% of the human canonical transcriptome with at least three reads (**Supp. Table 1-3**). To assess the secretion-predictive value of the method, we picked random clones from the sorted population of a single experiment (**Suppl. Fig. 4, Suppl. Table 4**) and tested the secretion of their encoded fragments when not fused to the anchor protein Sag1. The N- or C-terminal tags of 18/20 (90%) fragments could be reliably detected on western blot from the growth medium, and for 16/20 (80%) fragments, both tags were recognized (**Fig. 1f, Supp. Table 5**). As such, fragments displayed by sorted cells are indeed 'secretable' with a high probability. We further classified fragments into those that were enriched (also referred to as secretable) and those that were passively depleted (hence, not detected as secretable) by sorting, setting an arbitrary cut-off on the enrichment factor ($E\ factor = log2\frac{FPTM_{sorted}}{FPTM_{unsorted}}$) at 1 and -1, respectively, reflecting a minimal 2-fold increase and decrease in normalized sequence read counts after sorting. Indeed, as our MACS/FACS based sorting scheme was operated as a binary classifier (sorted/non-sorted), a binary threshold-based classifier is more appropriate in data processing than generalized linear models. Of 170,226 in-frame fragments commonly detected in the three experiments, 6.83% were consistently enriched in all three replicates, and 80.21% consistently depleted (**Supp. Table 6, Suppl. Fig. 5**). Thus, using this metric, these screens were reproducible with an 87.03% concordance between replicates. Even this partial-coverage dataset now represents by far the largest resource on eukaryotic secretability of human protein fragments. To enable researchers to easily access and analyze the data from the screens in this study, we built an interactive database (http://iomics.ugent.be/secrify/search) allowing visualization of the protein fragments detected in these screens and their mapping on available PDB structures (**Suppl. Fig. 11,** Figshare links to data in **Online Methods**).

Just as cytosolic protein expression is influenced by a variety of DNA, mRNA, and protein sequence or structural features and their complex interplays[21–24], fragment secretion will depend on a combination of multiple parameters, some of which are related to the unique environment and QC machinery of the ER and beyond. Even already at the simple level of general averaged parameters over the entire secretable vs. non-secretable protein fragment collections, several intriguing observations emerged from our data. Based on predictions of biophysical propensities and on PDB structure mapping (see also **Suppl. Fig. 12**), secretable fragments not only tend to have a lower α-helical content than depleted fragments (**Fig. 2a-b, Suppl. Fig. 13-14**), but are also distinctly more flexible and intrinsically disordered (**Fig. 2c-d, Suppl. Fig. 15a-16**), as well as slightly compositionally biased (**Suppl. Fig. 17-18**, see also **Suppl. Results**). Possibly, this reflects how unstructured sequences that lack typical exposed hydrophobic amino acids are missed by ER chaperones and can subsequently travel downstream. In contrast, no clear differences in number of Cys or N-glycosylation sequons were observed (**Suppl. Fig. 17f-g, Suppl. Fig. 18f-g**). Increasing the fidelity of these findings, all of these patterns were reproduced in screens of a different library of slightly larger fragments in a different yeast species (*P. pastoris*) (**Suppl. Results, Suppl. Table 7-10, Suppl. Figs. 6-19**). As for the early folding probability (**Suppl. Fig. 14b**), in cases where depleted and enriched fragments overlap in sequence on the same protein, the presence or absence of regions that are most likely to fold rapidly often correlates with secretability (**Fig. 2e**). Of those fragments that mapped to known structures (roughly 50% of representative fragments), secretable fragments are enriched in distorted sandwich and β complex folds compared to depleted fragments, suggesting that these folds are potentially more stable in the secretory environment, while the opposite is true for proteins with, for example, an α horseshoe architecture (**Suppl. Table 11**, Figshare links to data in **Online Methods**). Similarly, certain Pfam domains, such as the AAA18-domain (PF13238), are more prominent in enriched fragments than depleted fragments, while many typically cytoplasmic domains such as ribosomal protein domains or the tetratricopeptide repeat (TPR, PF13181) are found exclusively in depleted fragments (Figshare links to data in **Online Methods,** see also **Suppl.**

Fig. 19). This illustrates that sequence- and fold-contextual patterns of features still contain much information that is not apparent from averaged parameters. For the first time, our SECRiFY method generates secretability data at a scale at which training of predictive machine learning classifiers becomes feasible and this will be a subject of future work.

With SECRiFY, we thus demonstrate that the secretability of protein fragments across entire proteomes can be verified experimentally in an efficient, high-throughput and reproducible manner. Already, the databases we have generated in this work constitute by far the largest resources of such yeast-secretable human protein segments, which will be useful in structural studies (where high levels of proteins are required for crystallization), immunological experiments (where recombinant production is needed for vaccine development or antigen discovery), and biochemical characterizations of particular proteins (which also necessitates a minimal protein amount). Furthermore, it is likely that SECRiFY will provide a means to characterize the substrate scope of secretory system processes that regulate secretory protein passage through the eukaryotic secretory system in a proteome-wide manner. This complements existing methods, such as ribosome profiling[25], which deal with protein biogenesis prior to passage through the secretory system.

Acknowledgements

The authors thank the VIB Nucleomics Core for Illumina sequencing, as well as M. Vuylsteke for statistics discussions, K. Vandewalle for help with western blot sample preparation, and Y. Dondelinger for help with plasmid construction. We thank Lennart Martens for freeing up time to discuss the project and to allow D.M, D.T., E.V. and P.R. to spend time on this project away from their main tasks. This work was supported by a Ghent University BOF PhD Fellowship (M.B.), a PhD Fellowship from the Research Foundation Flanders (FWO) (M.B.), an FWO research grant (G.0276.13N) (N.C.) and an ERC Consolidator grant no. 616966 (N.C.). E.V. is a postdoctoral research fellow of the Research Foundation Flanders. P.R. and W.V. acknowledge support from the Research Foundation Flanders through grant number G.0328.16N.

Author Contributions

M.B. performed library constructions, methods development and optimization, screenings, sequencing data processing and analysis, and wrote the manuscript. P.R. performed structure-based data-interpretation under the supervision of E.V. and W.V. W.V. ran, analyzed and visualized the biophysical predictions. D.M. and D.T. developed the website interface and the underlying database. N.C. conceived the project, and assisted in experimental design, interpretation, and manuscript writing. All authors read, revised and approved the final version of the manuscript.

Competing Interests

The authors declare that no competing financial interests exist.

References

1. Barlowe, C. K. & Miller, E. A. Secretory protein biogenesis and traffic in the early secretory pathway. *Genetics* **193,** 383–410 (2013).

2. Braakman, I. & Hebert, D. N. Protein folding in the endoplasmic reticulum. *Cold Spring Harb. Perspect. Biol.* **5,** (2013).

3. Aviram, N. & Schuldiner, M. Embracing the void-how much do we really know about targeting and translocation to the endoplasmic reticulum? *Curr. Opin. Cell Biol.* **29C,** 8–17 (2014).

4. Cornvik, T. *et al.* Colony filtration blot: a new screening method for soluble protein expression in Escherichia coli. *Nat. Methods* **2,** 507–509 (2005).

5. Seitz, T. *et al.* Enhancing the stability and solubility of the glucocorticoid receptor ligand-binding domain by high-throughput library screening. *J. Mol. Biol.* **403,** 562–577 (2010).

6. Lockard, M. A. *et al.* A high-throughput immobilized bead screen for stable proteins and multi-protein complexes. *Protein Eng. Des. Sel. PEDS* **24,** 565–578 (2011).

7. Martinez Molina, D. *et al.* Engineering membrane protein overproduction in Escherichia coli. *Protein Sci.* **17,** 673–680 (2008).

8. Luan, C.-H. *et al.* High-throughput expression of C. elegans proteins. *Genome Res.* **14,** 2102–2110 (2004).

9. D'Angelo, S. *et al.* Filtering 'genic' open reading frames from genomic DNA samples for advanced annotation. *BMC Genomics* **12 Suppl 1,** S5 (2011).

10. Gupta, A. *et al.* A novel helper phage enabling construction of genome-scale ORF-enriched phage display libraries. *PloS One* **8,** e75212 (2013).

11. Reich, S. *et al.* Combinatorial Domain Hunting: An effective approach for the identification of soluble protein domains adaptable to high-throughput applications. *Protein Sci.* **15,** 2356–2365 (2006).

12. Yumerefendi, H., Tarendeau, F., Mas, P. J. & Hart, D. J. ESPRIT: an automated, library-based method for mapping and soluble expression of protein domains from challenging targets. *J. Struct. Biol.* **172,** 66–74 (2010).

13. An, Y., Yumerefendi, H., Mas, P. J., Chesneau, A. & Hart, D. J. ORF-selector ESPRIT: a second generation library screen for soluble protein expression employing precise open reading frame selection. *J. Struct. Biol.* **175,** 189–197 (2011).

14. Pedelacq, J.-D. *et al.* Experimental mapping of soluble protein domains using a hierarchical approach. *Nucleic Acids Res.* **39,** e125 (2011).

15. Hart, D. J. & Waldo, G. S. Library methods for structural biology of challenging proteins and their complexes. *Curr. Opin. Struct. Biol.* **23,** 403–408 (2013).

16. Prodromou, C., Savva, R. & Driscoll, P. C. DNA fragmentation-based combinatorial approaches to soluble protein expression Part I. Generating DNA fragment libraries. *Drug Discov. Today* **12**, 931–938 (2007).

17. Zhulidov, P. A. *et al.* Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* **32,** e37 (2004).

18. Bogdanov, E. A. *et al.* Normalizing cDNA libraries. *Curr. Protoc. Mol. Biol.* **Chapter 5,** Unit 5.12.1-27 (2010).

19. Thibault, G. & Ng, D. T. W. The endoplasmic reticulum-associated degradation pathways of budding yeast. *Cold Spring Harb. Perspect. Biol.* **4,** (2012).

20. Xu, C. & Ng, D. T. W. Glycosylation-directed quality control of protein folding. *Nat. Rev. Mol. Cell Biol.* **16**, 742-752 (2015).

21. Kliman, R. M., Irving, N. & Santiago, M. Selection conflicts, gene expression, and codon usage trends in yeast. *J. Mol. Evol.* **57,** 98–109 (2003).

22. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).

23. Zur, H. & Tuller, T. Strong association between mRNA folding strength and protein abundance in S. cerevisiae. *EMBO Rep.* **13,** 272–277 (2012).

24. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529,** 358–363 (2016).

25. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).

Figures

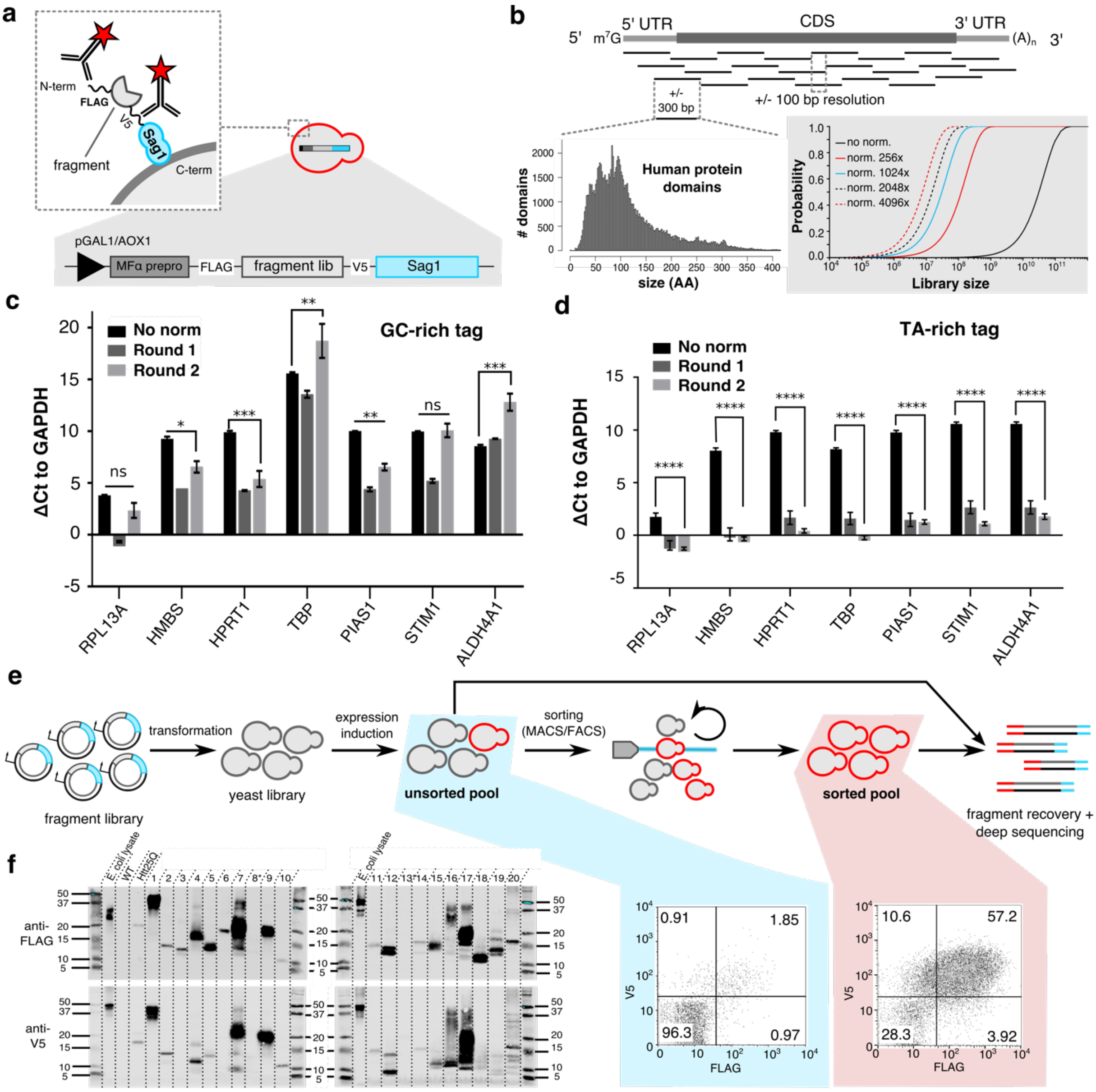

Fig. 1. The SECRiFY surface display platform for secretability screening of fragment libraries in yeast. (**a**) Surface display as a proxy for secretion. Productive passage through the yeast secretory system leads to antibody-based labeling of displaying clones through FLAG and V5 epitope tags. (**b**) Human mRNA is fragmented to mimic the size distribution of human protein domains (lower left, Gene3D, n= 104,734). Lower right in grey: estimated relationship between library size and the probability of sampling any fragment, depending on the efficiency of fragment abundance normalization. (**c** and **d**) Normalization of abundance differences compared to *GAPDH*, as ΔCt +/- SEM, is effective only when using a TA-rich (d), not a GC-rich (c) tag in the random primer. Two-way ANOVA with Tukey post-hoc, ns: non-significant, * p<0.05, ** p <0.01, *** p<0.001, **** p<0.0001. n= 9. (**e**) After fragment expression

induction, displaying FLAG$^+$V5$^+$ clones are sorted in multiple rounds of MACS/FACS. Fragments are identified by deep sequencing of both pools. (**f**) The majority of protein fragments from sorted cells are detected as secreted when expressed without the Sag1 display anchor.

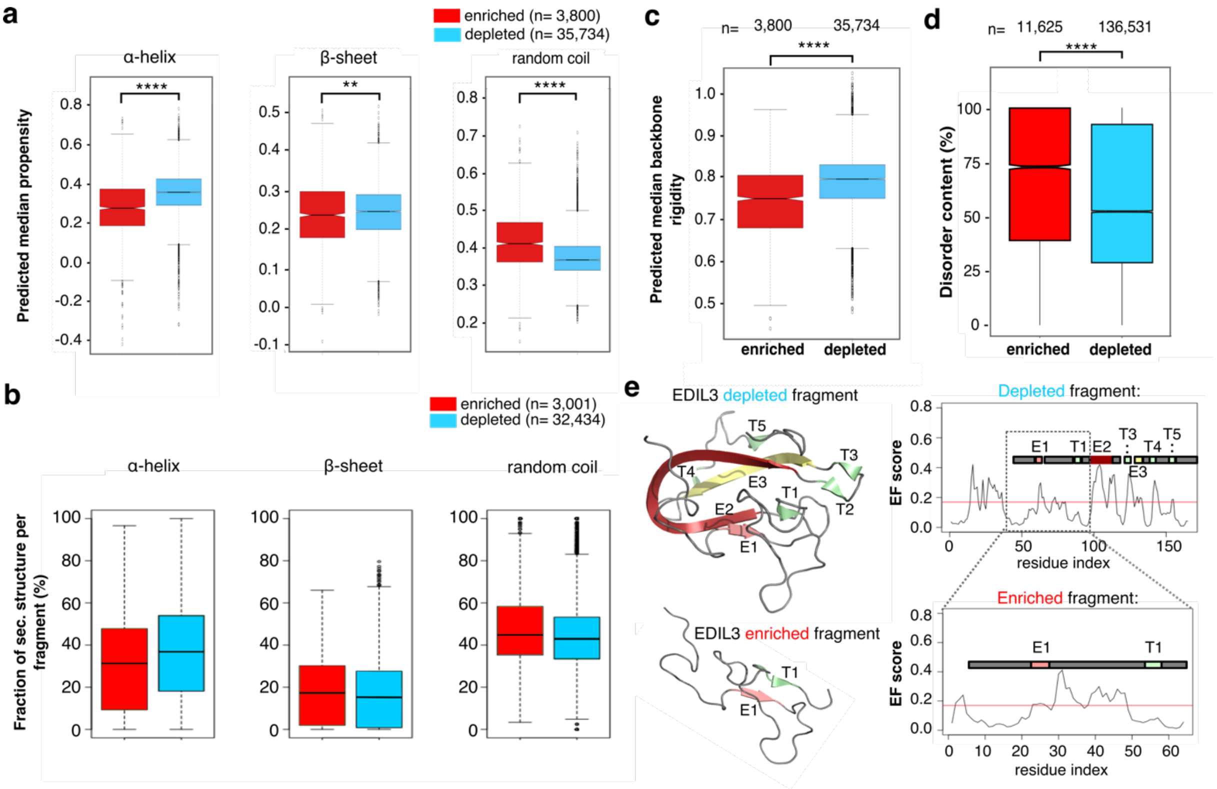

Fig. 2. Patterns in secretable fragments. (a). Predictions of secondary structure propensity in enriched and depleted fragments in *S. cerevisiae* shows that enriched fragments have a lower helical content and a higher random coil propensity, which is confirmed further by mapping fragments to known structures in PDB (**b**). Enriched fragments are also predicted to be more dynamic (**c**) and disordered (**d**) than depleted fragments. (**e**). Two overlapping fragments of the human protein EDIL3 differ in secretability outcome. Early folding (EF) propensity predictions suggest that for the depleted fragment regions E2, T3/E3 and R4 are likely the regions driving folding of the depleted fragment, and lack of these regions in the enriched fragment result in a change in secretability.