UNIVERSITY OF CALIFORNIA SAN DIEGO

Systems Biology of Protein Secretory Pathway: from Improving Recombinant Protein Production to Dysregulation in Alzhiemer's Disease

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Chih-Chung Kuo

Committee in charge:

Professor Nathan E. Lewis, Chair
Professor Xiaohua Huang, Co-Chair
Professor Ludmil Alexandrov
Professor Hannah Carter
Professor Maho Niwa

2021

The Dissertation of Chih-Chung Kuo is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am grateful to the many people with whom I get to cowork and befriend during my time as a PhD student. This work would not have been possible without their support and input.

I would like to acknowledge the guidance and support of my research advisor Professor Nathan Lewis. His patience, feedback on my work, and depth of knowledge have been instrumental to my continuing development into a scientist. I couldn't have asked for a better advisor.

I would also like to thank my close collaborators, Austin Chiang, Magdalena Malm, Helen Masson, Caressa Robinson and Mojtaba Samoudi for the stimulating conversations and their willingness to share their expertise. I am grateful to be a part of a wonderful lab. To all the CHOmigos I've crossed path with-- Albert, Alex, Anne, Ben, Bokan, Chintan, Erik, Hooman, Hratch, Isaac, Jahir, James, Joanne, Julie, Matt, Phil, Sarat, Shangzhong and Sharon, thank you all for making my life in the lab memorable.

To all my friends, old and new, thank you for your companionship. My journey as a PhD student definitely would not have been as enjoyable without every single one of you.

Last but not least, I am greatly indebted to my parents. I would not have made it this far if it were not for their unconditional love and support.

Chapter 1, in full, is a reprint of the material as it appears in Kuo CC, Chiang AW, Shamie I, Samoudi M, Gutierrez JM, Lewis NE. "The emerging role of systems biology for engineering protein production in CHO cells". Current opinion in biotechnology, 2018. The dissertation author was the primary investigator and author of this material.

Chapter 2, in part, is currently being prepared for submission for publication of the material as it may appear in Malm M, Kuo CC, Barzadd MM, Mebrahtu A, Wistbacka N, Razavi R, Volk AL, Lundqvist M, Kotol D, Edfors F, Gräslund T, Chotteau V, Field R, Varley PG, Roth RG, Lewis NE, Hatton D, Rockberg J. "Harnessing secretory pathway differences between HEK293 and

CHO to rescue production of difficult to express proteins" (Submitted). The dissertation author was the one of the two primary investigators and authors of this material.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Kuo CC, Masson HO, Lewis NE. "Transcriptomic analysis of recombinant protein-producing CHO cells reveal product-dependent host response" The dissertation author was the co-first investigator and author of this material.

Chapter 4, in full, is a reprint of the material as it appears in Samoudi M, Kuo CC, Robinson CM, Shams-Ud-Doha K, Schinn SM, Kol S, Weiss L, Bjorn SP, Voldborg BG, Campos AR, Lewis NE. "In situ detection of protein interactions for recombinant therapeutic enzymes". *Biotechnology and Bioengineering*, 2020. The dissertation author was the co-first investigator and author of this material.

Chapter 5, in part, is currently being prepared for submission for publication of the material. Samoudi M, Kuo CC, Robinson CM, Lewis NE. "Essential components of the secretory pathway correlating with high productivity in antibody-producing CHO cells" The dissertation author was the co-first investigator and author of this material.

Chapter 6, in full, is a reprint of the material as it appears in Kuo CC, Chiang AW, Baghdassarian HM, Lewis NE. "Dysregulation of the secretory pathway connects Alzheimer's disease genetics to aggregate formation". *Cell Systems*, 2021. The dissertation author was the co-first investigator and author of this material.

VITA

2013 Bachelor of Science, Electrical Engineering, National Taiwan University

2021 Doctor of Philosophy, Bioengineering, University of California San Diego

PUBLICATIONS

1. M. Samoudi*, **C.-C. Kuo***, C. M. Robinson*, N. E. Lewis. Essential components of the secretory pathway correlating with high productivity in antibody-producing CHO cells (In preparation).
2. **C.-C. Kuo***, H. O. Masson*, N. E. Lewis. Transcriptomic analysis of recombinant protein-producing CHO cells reveal product-dependent host response (In preparation).
3. M. Malm*, **C.-C. Kuo***, M. M. Barzadd, A. Mebrahtu, N. Wistbacka, R. Razavi, A.-L. Volk, M. Lundqvist, D. Kotol, F. Edfors, T. Gräslund; V. Chotteau, R. Field, P. G. Varley, R. G. Roth, N. E. Lewis, D. Hatton, J. Rockberg. Harnessing secretory pathway differences between HEK293 and CHO to rescue production of difficult to express proteins (Under review at Cell Systems).
4. **C.-C. Kuo**, A. W. T. Chiang, H. M. Baghdassarian, N. E. Lewis. Dysregulation of the secretory pathway connects Alzheimer's disease genetics to aggregate formation. Cell Systems, 2021. doi:10.1016/j.cels.2021.06.001
5. M. Samoudi, H. O. Masson, **C.-C. Kuo**, C. M. Robinson, N. E. Lewis. From omics to cellular mechanisms in mammalian cell factory development. Current Opinion in Chemical Engineering, 2021. doi:10.1016/j.coche.2021.100688
6. L. Y. Hsieh, A. W. T. Chiang, L. D. Duong, **C.-C. Kuo**, S. X. Dong, R. Dohil, R. Kurten, N. E. Lewis, S. S. Aceves. A Unique Esophageal Extracellular Matrix Proteome Alters Normal Fibroblast Function in Severe Eosinophilic Esophagitis. Journal of Allergy and Clinical Immunology, 2021. doi:10.1016/j.jaci.2021.01.023
7. A. W. T. Chiang, H. M. Baghdassarian, B. P. Kellman, B. Bao. J. T. Sorrentino, C. Liang, **C.-C. Kuo**, H. O. Masson, N. E. Lewis. Systems glycobiology for discovering drug targets, biomarkers, and rational designs for glyco-immunotherapy. Journal of Biomedical Science, 2021. doi:10.1186/s12929-021-00746-2
8. P. N. Spahn, X. Zhang, Q. Hu, N. K. Hamaker, H. Hefzi, S. Li, **C.-C. Kuo**, Y. Huang, J. C. Lee, P. Ly, K. H. Lee, N. E. Lewis. Restoration of deficient DNA Repair Genes Mitigates Genome Instability and Increases Productivity of Chinese Hamster Ovary Cells. bioRxiv, 2021. doi:10.1101/2021.01.07.425558
9. M. Samoudi*, **C.-C. Kuo***, C. M. Robinson*, K. Shams-Ud-Doha, S.-M. Schinn, S. Kol, L. Weiss, S. P. Bjorn, B. G. Voldborg, A. R. Campos, N. E. Lewis. In situ detection of protein interactions for recombinant therapeutic enzymes. Biotechnology and Bioengineering, 2020. doi:10.1002/bit.27621
10. H. Dahodwala, P. Kaushik, V. Tejwani, **C.-C. Kuo**, P. Menard, M. Henry, B. G. Voldborg, N. E. Lewis, P. Meleady, S. T. Sharfstein. Increased mAb production in amplified CHO cell lines is associated with increased interaction of CREB1 with transgene promoter. Current research in biotechnology, 2019. doi:10.1016/j.crbiot.2019.09.001
11. A. W. T. Chiang, S. Li, B. P. Kellman, G. Chattopadhyay, Y. Zhang, **C.-C. Kuo**, J. M. Gutierrez, F. Ghazi, H. Schmeisser, P. Ménard, S. P. Bjørn, B. G. Voldborg, A. S. Rosenberg, M. Puig, N. E. Lewis. Combating viral contaminants in CHO cells by engineering innate immunity. Scientific reports, 2019. doi:10.1038/s41598-019-45126-x

12. A. Richelle, A. W. T. Chiang, **C.-C. Kuo**, N. E. Lewis. Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. PLoS computational biology, 2019. doi:10.1371/journal.pcbi.1006867
13. **C.-C. Kuo**, A. W. Chiang, I. Shamie, M. Samoudi, J. M. Gutierrez, N. E. Lewis. The emerging role of systems biology for engineering protein production in CHO cells. Current opinion in biotechnology, 2018. doi:10.1016/j.copbio.2017.11.015
14. N. C. Yeo, A. Chavez, A. Lance-Byrne, Y. Chan, D. Menn, D. Milanova, **C.-C. Kuo**, X. Guo, S. Sharma, A. Tung, R. J. Cecchi, M. Tuttle, S. Pradhan, E. T. Lim, N. Davidsohn, M. R. Ebrahimkhani, J. J. Collins, N. E. Lewis, S. Kiani, G. M. Church. An enhanced CRISPR repressor for targeted mammalian gene regulation. Nature methods, 2018. doi:10.1038/s41592-018-0048-5
15. M. Uhlen, H. Tegel, Å. Sivertsson, **C.-C. Kuo**, J. M. Gutierrez, N. E. Lewis et al. The human secretome – the proteins secreted from human cells. bioRxiv , 2018. doi:10.1101/465815
16. J. P. Shen*, D. Zhao*, R. Sasik*, J. Luebeck, A. Birmingham, A. Bojorquez-Gomez, K. Licon, K. Klepper, D. Pekin, A. N. Beckett, K. S. Sanchez, A. Thomas, **C.-C. Kuo**, D. Du, A. Roguev, N. E. Lewis, A. N. Chang, J. F. Kreisberg, N. Krogan, L. Qi, T. Ideker, P. Mali. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nature methods, 2017. doi:10.1038/nmeth.4225
17. A. W. Chiang, S. Li, P. N. Spahn, A. Richelle, **C.-C. Kuo**, M. Samoudi, N. E. Lewis. Modulating carbohydrate-protein interactions through glycoengineering of monoclonal antibodies to impact cancer physiology. Current opinion in structural biology, 2016. doi:10.1016/j.sbi.2016.08.008

* These authors contributed equally

ABSTRACT OF THE DISSERTATION

Systems Biology of Protein Secretory Pathway: from Improving Recombinant Protein Production to Dysregulation in Alzhiemer's Disease

by

Chih-Chung Kuo

Doctor of Philosophy in Bioengineering

University of California San Diego, 2021

Professor Nathan E. Lewis, Chair
Professor Xiaohua Huang, Co-Chair

The secretory pathway is made up of a collection of mostly ER- and Golgi-resident proteins that work together to synthesize, post-translationally modify, transport and quality control the secreted proteins (secPs). This complex system can be leveraged to produce a variety of recombinant proteins such as antibodies, growth factors, and many other biotherapeutics. While the production yield of many recombinant proteins has seen a several-fold increase; many proteins still fail to express recombinantly despite bioprocess optimization. We start this doctoral dissertation by investigating the properties limiting production of difficult-to-express rProteins via analysis of proteogenomic data from popular mammalian cell hosts expressing different human proteins.

Decades of research has now better charted the secretory pathway, and the functional roles of individual proteins are better understood. However, many secPs experience production bottlenecks within the secretory pathway to various degrees. Incidentally, the nature of secP synthesis has been shown to be highly product-specific. As protein-protein interactions (PPIs) are one of the major modalities through which machinery proteins in the secretory pathway assist protein secretion, we charted the transient interaction partners of key secPs using proximity-based biotinylation and mass spectrometry analysis to better understand the rate-limiting steps unique to each secP within the secretory pathway. Additionally, we determined and verified the interactions that contribute to high recombinant protein yields via structural interaction modeling.

Perturbations to the secretory pathway result in misfolded proteins. Amyloid disorders, such as Alzheimer's disease (AD), involve aggregation of secPs. However, it is largely unclear how secretory pathway proteins contribute to amyloid formation. We integrated expression data with PPI networks to estimate a tissue's fitness for producing specific secreted proteins, and analyzed the fitness of the secretory pathway of various brain regions and cell types for synthesizing the AD-associated amyloid-precursor protein (APP). We associated Aβ aggregation with systemic dysregulation of the secretory pathway components proximal to APP and amyloidogenic secretases in AD. Our analyses suggest that perturbations from AD risk loci cascade through the APP secretory support network and into the endocytosis pathway, connecting amyloidogenesis to dysregulation of secretory pathway components supporting APP and suggesting novel therapeutic targets for AD.

CHAPTER 1: SYSTEMS BIOLOGY OF THE SECRETORY PATHWAY

Abstract

To meet the ever-growing demand for effective, safe, and affordable protein therapeutics, decades of intense efforts have aimed to maximize the quantity and quality of recombinant proteins produced in CHO cells. Bioprocessing innovations and cell engineering efforts have improved product titer; however, uncharacterized cellular processes and gene regulatory mechanisms still hinder cell growth, specific productivity, and protein quality. Herein we summarize recent advances in systems biology and data-driven approaches aiming to unravel how molecular pathways, cellular processes, and extrinsic factors (e.g. media supplementation) influence recombinant protein production. In particular, as the available omics data for CHO cells continue to grow, predictive models and screens will be increasingly used to unravel the biological drivers of protein production, which can be used with emerging genome editing technologies to rationally engineer cells to further control the quantity, quality and affordability of many biologic drugs.

Introduction

Over the past few decades, Chinese hamster ovary (CHO) cells have emerged as the primary host for the biopharmaceutical industry. CHO cell lines were derived from the same hamster in 1957, and variants of the cell line have (e.g., CHO-K1, CHO-S and DG44) been further developed to meet different production requirements (see [1,2] for detailed genetic and phenotypic differences across the common CHO cell lines). These cell lines have been adopted by industry for various reasons, including the development of DHFR-deficient CHO cell line including which enable efficient transgene transfection and amplification. They also exhibit excellent capabilities to perform human-compatible post-translational modifications (PTMs), and they are highly adaptable to suspension-growth culture conditions in chemically-defined media. They also exhibit favorable safety profiles, e.g., being less prone to virus infection [3–5]. Over the past several

decades, extensive efforts have aimed to increase the productivity of these cells to reduce the costs associated with culturing the cells and purifying products. Thus, innovations have effectively increased the yield of recombinant proteins (e.g., monoclonal antibodies) by three orders of magnitude, from 10-50 mg/L in the 1980s to >10g/L in the 2010s [6,7].

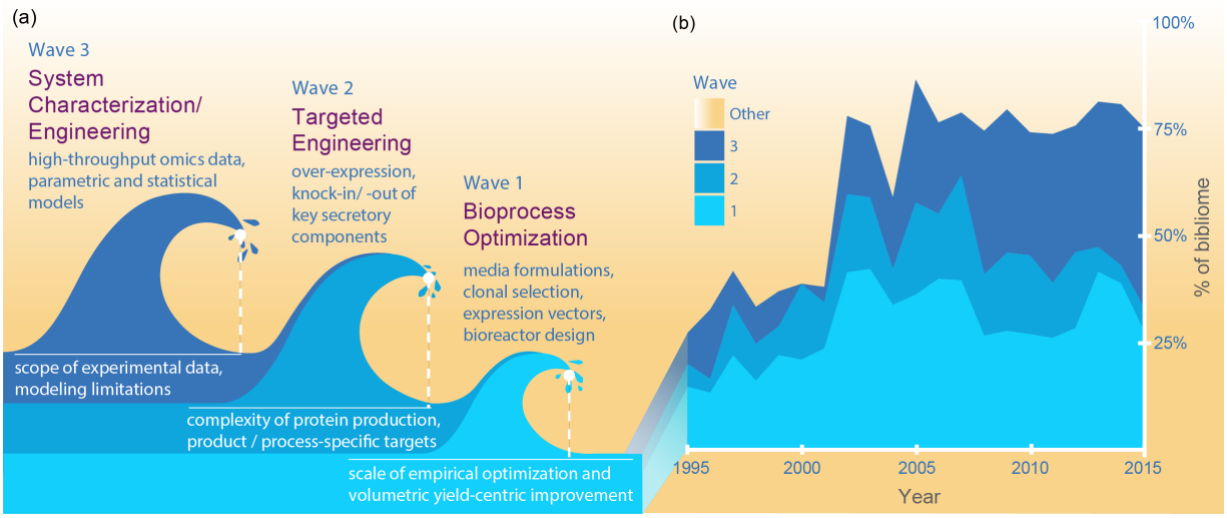


Figure 1. Three waves of different technologies have enabled continued improvement of recombinant protein production in CHO cells.
(a) Recombinant protein production has steadily improved over the past few decades thanks to innovations in bioprocessing, targeted genetic manipulation of cells, and systems biology approaches. Together, novel technologies, approaches and discoveries in each field have been of great importance. (b) A comprehensive survey of the CHO bioprocessing literature [8] highlights the historical development of the field in CHO cell research. The first wave-bioprocess development has been driving most of the earlier studies, while the targeted genetic manipulation, omics studies, and modeling efforts have become increasingly important after the mid-2000s with the increased prevalence of genomic resources, genome editing technologies, and development of novel computational models and algorithms.

Historically, at least three waves of innovations have offered additional toolboxes to further enhance biotherapeutic production (Fig. 1). The first wave significantly improved the volumetric yield, and leveraged innovations in bioprocessing techniques, media optimization [9] and tools that improve production by engineering the transgene and vectors (e.g., to optimize mRNA copy number and codon usage). The second wave involves targeted engineering of the host cells to enhance productivity and per-cell yield [10]. The third wave is beginning to use systems-level

engineering to boost protein productivity by modulating cellular pathways to optimize cellular processes (e.g., metabolism [11]). It is being enabled by systems biology models [12,13], large-scale omics datasets [14,15]), and combinatorial genome editing [16,17], which are discovering and leveraging more comprehensive knowledge about the cell pathways influencing protein quantity and quality. Each wave continues to contribute novel innovations, and are resulting in improved protein production (Fig. 2).

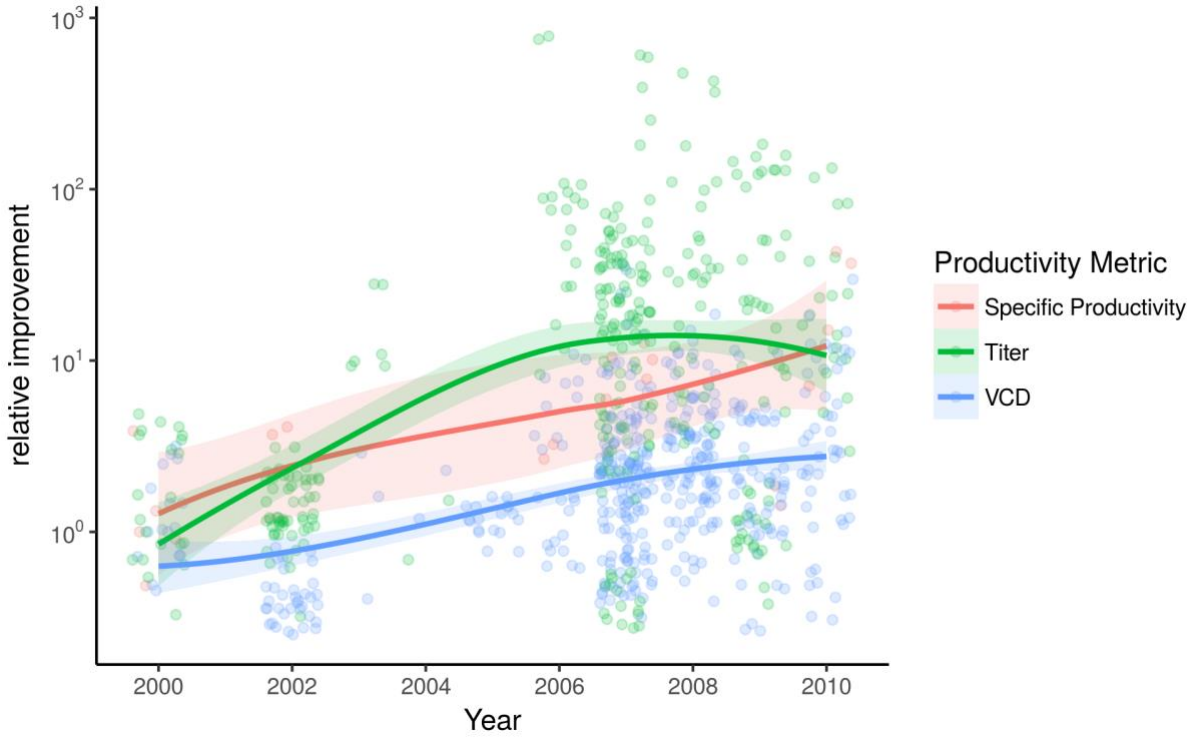


Figure 2. Published specific productivity, cell density and total product titer has improved steadily over the years.
The trend for three major productivity metrics reported by literature from 2000- 2010 [8]. As a result of development in bioprocessing and feeding strategies, the volumetric yield has been greatly improved. The introduction of cell engineering to CHO has further improved the per-cell productivity since the mid-2000s.

In this review, we mention a few important innovations in each wave, and focus primarily on emerging efforts in systems biology and data-driven approaches that can advance our understanding of the cellular mechanisms contributing to recombinant protein production in CHO

cells. These techniques are starting to guide efforts to engineer the cellular pathways and improve the product quality and protein productivity. These emerging efforts are ushering in an era of rational cell factory design in mammalian cell bioprocessing.

Wave 1: Bioprocess and transgene expression optimization

Bioprocess and transgene expression optimization has improved recombinant protein titer in CHO cells by ~100-fold over the past few decades. This increase in volumetric yield has been primarily achieved through media optimization [18], clonal selection processes [19], expression vectors [20], genetic elements [21], bioprocess controls [22], and bioreactor design [23]. Recent innovations further enhance production through high-throughput assays to test everything from genetic elements to media conditions [24], leveraging tools from robotics to microfluidics [25].

Wave 2: Targeted engineering of CHO cells

Optimizing these extrinsic factors has improved titers, often by achieving higher cell densities; therefore, opportunities remain to further enhance the per-cell yield by directly engineering the cells [26]. Several cellular processes are associated with protein production, such as metabolism [27] and the secretory pathway [28]. Thus, following the success from bioprocess optimization came the second wave of strategies to engineer host cell lines. The advent of targeted genetic modification technologies, including knock-in strategies, have enabled the study of genes that improve protein production [29]. For example, overexpression of secretory pathway elements has been used to locate the faulty step in protein secretion while comparing the expression of easy- and difficult-to-express proteins [30]. Additional tools, such as ZFNs, TALENs [31] and CRISPR/Cas9 [16] enable efforts to edit individual host cell genes to fine-tune cell physiology, and precisely control product quality, such as glycosylation [32,33]. Further improvements to protein production could be achieved as additional emerging technologies are applied to CHO to activate or repress host cell genes with the CRISPRa/i system [34] and targeted epigenetic changes [35].

Wave 3: Characterizing and Engineering the CHO Protein Secretion System

Genome-wide analysis of protein secretion through omics technologies

The advances in the first two waves have provided powerful tools to enhance protein production. However, the synthesis and secretion of a single protein depend on the concerted function of hundreds or thousands of other proteins. Thus, truly effective engineering strategies may require multiple genetic changes to the host cell. To achieve this, efforts have been made to comprehensively study the molecular changes that occur to enable high rates of protein secretion, thus shedding light on molecular and physiological factors making certain cells high producers.

Omics data have been used extensively to study productive clones. For example, a differential proteomic analysis identified the up-regulation of glutathione biosynthesis and the down-regulation of DNA replication to be characteristic of high-producing CHO cells [36]. Likewise, transcriptomic profiling of various CHO cell lines indicated that certain favorable metabolic and glycosylation patterns are associated with differential expression of key genes [37]. Ribosome profiling and polysome profiling have also been used to quantify translation of recombinant proteins and the endogenous mRNA in antibody-producing CHO cells [38,39].

Metabolite profiling of CHO can improve production by measuring metabolite accumulation and nutrient consumption. Indeed, several studies have profiled both extracellular and intracellular metabolites in CHO cell cultures with different growth media to connect cell culture media, productivity and growth rate [40–42]. Metabolomics has also successfully identified novel apoptosis-inducing metabolites that accumulate in the culture media [43].

These and many additional studies, show that omics data have emerged as valuable assays that provide insights into which genes, proteins, metabolites are associated with desired traits in protein production in CHO cells. Furthermore, they are helping to identify potential targets for cell engineering and bioprocess optimization for enhanced protein production.

Mapping out the CHO secretory pathway

The aforementioned high-throughput omics experiments often provide many differentially expressed genes, and it can be unclear which genes are most responsible for the improvements in production. Since bottlenecks in the secretory pathway frequently limit recombinant protein production [44], analysis of omics data in the context of this pathway can be informative. Recent progress in high-throughput omics technologies now allow researchers to systematically map out and dissect portions of the secretory pathway, such as protein synthesis, the unfolded protein response, glycosylation, and metabolism.

Various omics technologies are helping identify components of the secretory machinery. For example, a systematic discovery of genes involved in protein folding was carried out in yeast with synthetic genetic arrays [45]. More recently, a similar screen conducted at the single-cell level with combinatorial CRISPR interference revealed the bifurcation of unfolded protein response in unprecedented detail [46].

Such studies are fueling efforts to connect the known secretory machinery components. A network reconstruction of the CHO secretory pathway characterized the functional roles and localizations of the secretory machinery components, allowing better integration of omics data in the context of the secretory pathway [47]. Similarly, the machinery required for protein synthesis, post-translational modification, and secretion of individual recombinant proteins has been mapped out for mammals, enabling insights into product-specific needs [48].

Another component of the CHO secretory pathway required for most biotherapeutics is human-compatible glycosylation [49]. Recent advances in glycomics have enabled the profiling of glycan structures under various glycosyltransferase genes knockouts [32] and lectin binding preference [50] in CHO.

Developing predictive models for elevating cell productivity and product quality

Efforts to map out the protein secretion pathway are enabling more informative analyses of omic datasets. Such resources provide a platform for systems biology and machine learning algorithms to understand cell mechanisms for the production of recombinant proteins in CHO cells. Modeling efforts centered around the mechanisms in CHO protein production usually fall into one of the two frameworks: knowledge-based parametric models, and data-driven statistical models.

The knowledge-based modeling paradigm links the genotype to phenotype on a mechanistic basis. With careful curation, the models could help distill biological causation from observed data correlation. Genome scale models (GEMs) directly couple cellular functions such as cell growth and protein synthesis to enzyme activities [51]. The most comprehensive genome scale metabolic reconstruction in CHO [13] has provided recent insights into changes in lipid metabolism in antibody-producing CHO cells [52]. Apart from the stoichiometrically motivated GEMs, kinetic models characterize the dynamics of the cellular processes. These models have provided valuable insights in smaller-scale systems such as glycolysis and the pentose phosphate pathways [53]. N- and O-linked glycosylation profiles can also be modelled [54] through rule-based kinetic [55] and Markov models [56,57]. In addition, specific productivity was found to influence mAb glycosylation through an integrated model that couples glycosylation with cellular metabolism and secretory capacity [58].

Data-driven models do not rely on labor-intensive human curation, and they make fewer assumptions about the host cells. Therefore, these methods are particularly valuable in poorly characterized systems. Such models have been deployed to study the productivity of recombinant proteins and antibodies using CHO gene expression [59], product sequence features [60] and measurements of various bioprocess parameters [8]. Other bioprocess variables such as lactate consumption can also be accurately predicted [61]. One dilemma facing data-driven models is the shortage of high-coverage experimental data, used as training sets [62]. As biological data can be

difficult or expensive to obtain, having a community-driven repository for various types of omics data can be one way to mitigate the shortage of training data [63,64].

Both of these powerful modeling frameworks are enabling the simulation and analysis of cellular responses influencing recombinant protein production in CHO cells. Furthermore, they are facilitating detailed analysis and integration of multiple omics data types. With the rather recent introduction of systems biology and machine learning methods to recombinant production in CHO cells, we expect to see a more widespread adoption of these tools for guiding rational design of CHO cell factories.

Conclusion

While innovations have driven a 1000-fold increase in protein titer in CHO cells, many challenges remain surrounding the production of many therapeutic proteins at high specific cellular productivity and high quality. Thus, further innovations in bioprocess optimization are needed to optimize expression conditions. Similarly, to speed up screening efforts, we need higher efficiency in genome editing strategies and high expression targeted integration sites for transgenes. Finally, omics studies and model-guided approaches will continue to map out the cellular pathways influencing the quantity and quality of secreted proteins. Fundamentally, a better general understanding of CHO cells is needed. For example, clonal variation and genomic instability in CHO lead to variable protein production over time. A recent multi-omics study showed that ~40% of differentially expressed genes in a producer cell line contained different copy number variations, suggesting CNVs as a driver of transcriptional activation as opposed to epigenetic or regulatory changes [52]. Thus, to unravel which genetic and epigenetic changes underlie desired protein production traits, large scale genetic screens coupled with multi-omics data and computational models [46,65,66] will be invaluable to understand and engineer desired characteristics such as specific productivity, viability, morphology, and growth rate for large-scale bioprocesses.

We anticipate that such data and novel computational tools will be increasingly valuable to therapeutic protein production.

Chapter 1, in full, is a reprint of the material as it appears in Kuo CC, Chiang AW, Shamie I, Samoudi M, Gutierrez JM, Lewis NE. “The emerging role of systems biology for engineering protein production in CHO cells”. Current opinion in biotechnology, 2018. The dissertation author was the primary investigator and author of this material.

CHAPTER 2: TARGETED ENGINEERING OF THE SECRETORY PATHWAY TO BOOST DIFFICULT-TO-EXPRESS RECOMBINANT PROTEIN PRODUCTION ENABLED BY SYSTEMS ANALYSIS OF PROTEOGENOMIC DATA

Abstract

Biologics represent the fastest growing group of therapeutics, but many advanced recombinant protein moieties remain difficult to produce. Here, we identify bottlenecks limiting expression of recombinant human proteins through a systems biology analysis of the transcriptomes of CHO and HEK293 during recombinant overexpression. Surprisingly, one third of the challenging human proteins displayed improved secretion upon host cell swapping from CHO to HEK293. While most components of the secretory machinery showed comparable expression levels in both expression hosts, genes with significant expression variation were identified. Among these, ATF4, SRP9, JUN, PDIA3 and HSPA8 were validated as productivity boosters in CHO. Further, more heavily glycosylated products benefited more from the elevated activities of the N- and O-glycosyltransferases found in HEK293. Collectively, our results demonstrate the utilization of HEK293 for expression rescue of human proteins and suggest a methodology for identification of secretory pathway components improving recombinant protein yield in HEK293 and CHO.

Introduction

The Chinese hamster ovary (CHO) cell line is commonly used for producing recombinant proteins (r-proteins) as it enables efficient expression of proteins with the need for human-like post-translational modifications. The CHO cell line provides several attractive properties for large-scale production of biopharmaceuticals, such as the ability to be cultivated at high cell densities in serum-free and chemically defined media and low risk of infection of human viruses[67]. Currently, CHO cell lines are the biopharmaceutical mammalian workhorses, producing 84% of recently

approved monoclonal antibodies[68]. However, with the boom of biologics within the pharma industry combined with more complex pharmaceutical proteins reaching the market, there is a demand for bioproduction platforms that can produce more difficult to express proteins. Data from the ongoing human secretome project[69,70], a comprehensive research study that includes the secreted production of >1500 human proteins, suggests approximately 35% of proteins are challenging to efficiently produce in secreted form by the CHO expression system. For such difficult to express proteins, alternative expression systems based on other expression hosts or engineered cell lines may provide improved expression titers. As cells of different origins can have tissue-specific expression patterns of secretory pathway genes[48], such differences can bring about variation in expression and processing of r-proteins depending on expression host, which in turn can impact the protein's stability, function, activity, immunogenicity and production titer. Moreover, poor expression of human proteins in CHO has previously been overcome by exogenous expression of human endoplasmic reticulum (ER)-associated proteins in the CHO production cells[29,71]. Thus one could speculate, when focusing on improving expression of many challenging human proteins, that the CHO cell line may have a secretory systems disadvantage compared to human production hosts. In particular, the human embryonic kidney 293 (HEK293) cell line has traditionally been a popular alternative for bioproduction and is commonly used for transient production of proteins for research and preclinical studies. HEK293 is considered easy to transfect, it adapts well to suspension cultivation, grows rapidly in serum-free media and is capable of producing r-proteins at high titers[72,73]. In particular, glycosylation profiles of r-proteins produced in either CHO or HEK293 have been shown to differ[74–76]. In the context of biopharmaceutical production, specific demands for human post-translational modifications for certain r-proteins have made HEK293 successful alternatives to the conventional CHO cells and in 2018 five approved protein therapeutics were produced in HEK293 cell lines[68,77].

Here, we evaluated alternative expression systems for production of a set of human difficult to express secreted proteins or extracellular domains of single-pass plasma membrane-

anchored proteins[69,70]. Expression systems were evaluated both based on r-protein expression levels but also systemic differences based on transcription of secretory pathway components between cell lines. Initially, a comparison of two CHO-based expression systems was carried out followed by a comparison of the secreted expression between CHO and HEK293. Moreover, transcriptome data from CHO and HEK293 cells transiently expressing proteins was analyzed to map differences in secretory pathway components between HEK293 and CHO. Based on the most profound differences in the expression of secretory pathway components between the cell lines, secretory pathway genes with significant impact on secreted productivity of human difficult-to-express proteins were identified. These include genes that assist protein secretion in a product-independent fashion. Additionally, we coupled specific post-translational modifications (PTMs) of different r-proteins with the protein titer improvements from CHO to HEK293 cells and showed that differences in titer improvements can be jointly explained by PTM features of the r-proteins and the activities of the enzymes responsible for these PTMs. These highly product-specific genes enable bespoke cell line designs that cater to the unique secretory requirements of different r-proteins, and allows for a more rational selection of cell hosts for a given r-protein.

Results

The two CHO platforms ExpiCHO and QMCF provide different benefits for difficult to express proteins

Due to differences in expression platform protocols, performance in productivity of r-proteins may vary even between hosts of the same origin. To shine light on differences between platforms for a range of difficult to express proteins, we compared expression levels in various systems. Initially, we evaluated two CHO-based expression systems by expressing a set of 23 challenging human proteins in the ExpiCHO system and compared secreted productivities to Human Secretome Project production data, wherein expression was performed using the episomal stable QMCF technology[78]. Briefly, this system utilizes CHOEBNALT85 cells stably expressing the Epstein-Barr virus EBNA-1 protein and the mouse polyomavirus (Py) large T

antigen, which facilitates nuclear retention and replication of the pQMCF expression vector. The ExpiCHO platform is a fully transient system that enables protein production at very high cell densities. In this comparison, transient expression of the 23 human genes in ExpiCHO was performed using the expression vector pKTH16_dPur, developed in-house. Results from this comparison showed different expression profiles depending on r-protein expressed and expression platform used (Fig. 3A). Even though purified secreted protein titers varied dramatically between the two platforms for some proteins, no platform provided an overall improved expression profile compared to the other. Instead, each platform provided improved expression in a protein-feature specific manner. Notably, a significant correlation between titer fold changes between the ExpiCHO and QMCF platforms and both protein size ($R=-0.5$, $p=0.015$) and glycosylation ($R=0.44$, $p=0.044$) was observed (Fig. 3B). This suggested that each platform provided improved expression in a protein-feature specific manner, where larger and less glycosylated proteins tended to have an expression advantage in the QMCF system and vice versa in case of the ExpiCHO platform. We performed gene expression profiling of the ExpiCHO and CHOEBNALT85 cell lines and noticed comparable transcriptional and secretory pathway activities across the two platforms. Interestingly, the stable QMCF system showed significantly higher translational utilization than the ExpiCHO system. The protein-feature specific expression profiles observed between the two CHO expression platforms suggested that neither of the CHO expression hosts or cultivation-protocols provided optimal conditions in a protein-independent manner.

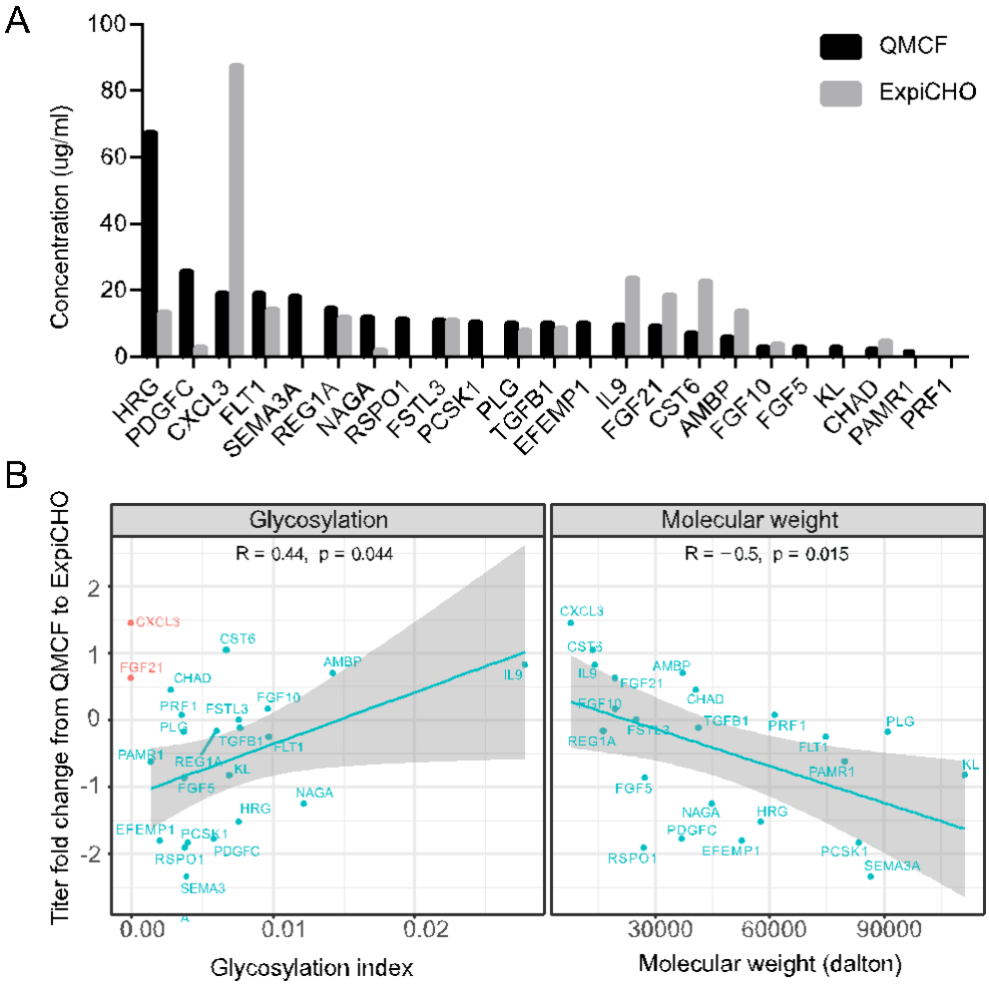


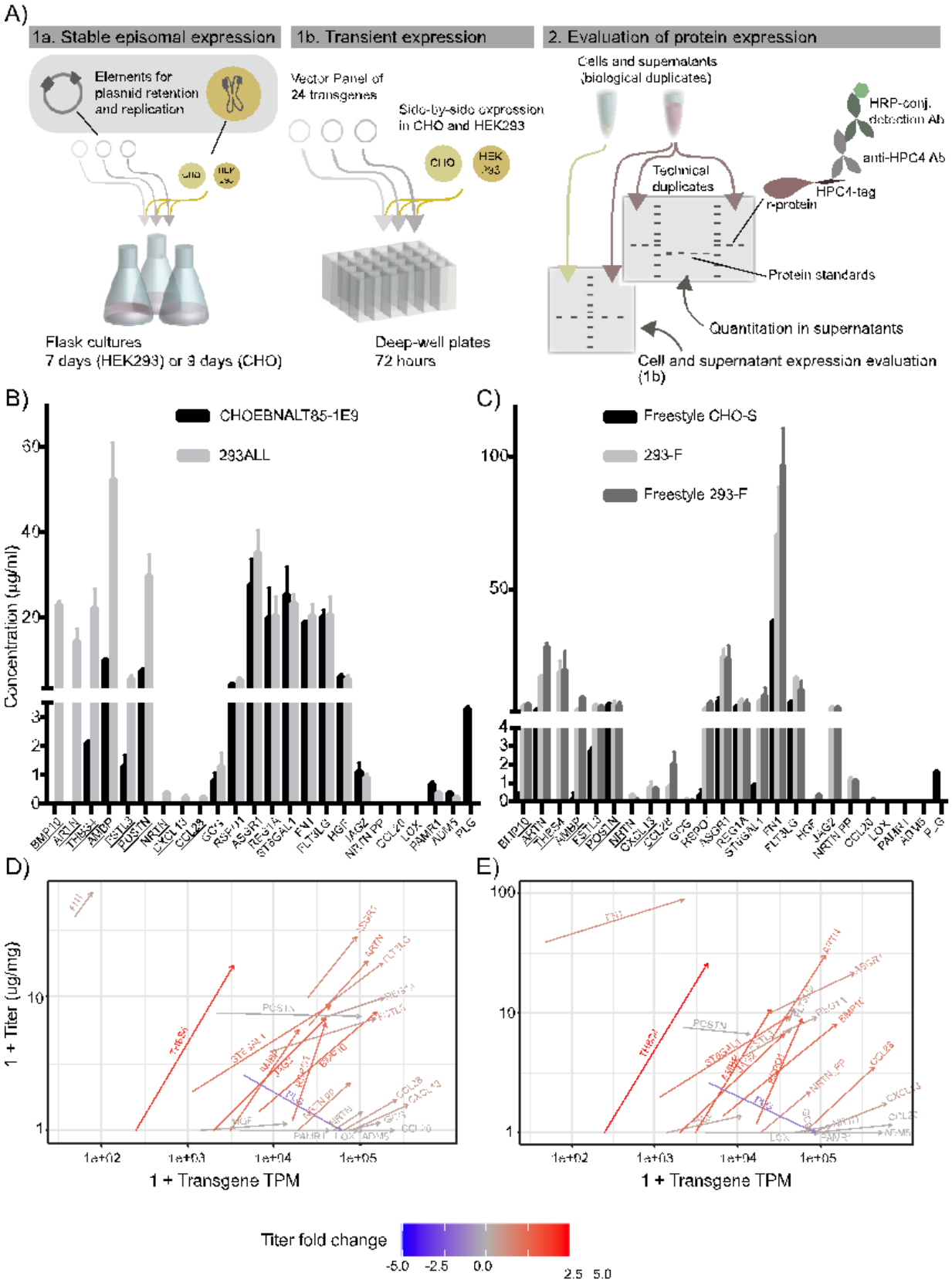
Figure 3. The expression of human secreted proteins in CHO cells.
(A) The secreted and purified titers of 23 difficult to express proteins in the QMCF technology versus the ExpiCHO system. (B) Correlation between the molecular weight and r-protein titer fold-change in ExpiCHO compared to the QMCF system. Non-glycosylated proteins are colored red and excluded from the correlation calculation.

Expression of challenging human proteins in HEK293 resulted in overall improved secreted titers compared to CHO

We hypothesized that a human expression host may have a more compatible secretory pathway for human secreted proteins and hence provide benefits for the expression of those that are particularly difficult in CHO. Thus, we also sought to compare r-protein production between CHO and HEK293. A panel of 24 difficult to express proteins from the Human Secretome Project was expressed side-by-side in CHO and HEK293 cells using a standardized expression- and evaluation-pipeline (Fig. 4A). We evaluated both an optimized version of the semi-stable QMCF

technology, using the CHOEBNALT85-1E9 and 293ALL cell lines, and a fully transient expression setup with 293-F, Freestyle 293-F and Freestyle CHO-S cell lines. In the episomal stable QMCF-system, 9 out of 24 exogenously expressed genes (THBS4, ARTN, BMP10, POSTN, FSTL3, AMBP, CCL28, CXCL13 and NRTN) showed more than 2-fold improved expression in HEK293 (293ALL) compared to CHO cells (CHOEBNALT85-1E9) (Fig. 4B). On the other hand, only one gene (PLG) showed at least 2-fold improved expression in CHO compared to HEK293. In addition, for six of the genes (NRTN, NRTN pp CXCL13, CCL28, CCL20 and LOX) no detectable secreted expression could be observed in the CHO cell line, whereas 3 genes (CCL20, LOX and PLG) resulted in no detectable protein expression in case of 293ALL.

Figure 4. HEK293 provides improved secreted titers of difficult to express proteins compared to CHO in two different expression systems.
Comparison of expression titers from supernatants of HEK293 and CHO cell lines from (A) a vector panel of the 24 transgenes was expressed side-by-side in HEK293 and CHO cells in both medium-scale (shake flasks) in the stable episomal long-term expression system QMCF in CHOEBNALT85-1E9 and 293ALL cell lines (1a) and in small-scale (deep-well plates) transient cultivations of Freestyle CHO-S, 293-F and Freestyle 293-F (1b). The protein expression was analyzed by western blot where r-protein was detected by targeting the C-terminal HPC4-tag using an anti-HPC4 antibody. Mean secreted titers ± SD of difficult to express proteins expressed in (B) the stable episomal expression system pQMCF using the CHOEBNALT-85-1E9 and 293ALL cell lines or (C) the transient cultivation protocol of Freestyle CHO, 293-F and Freestyle-293-F. The underlined gene names indicate genes with more than two-fold improved expression in HEK293 cell lines compared to CHO. Differences in protein titer and mRNA levels were quantified for each expressed transgene in CHO and HEK293 cells. The transgene RNA and protein amounts for Freestyle CHO-S and 293-F cultures (D) and Freestyle 293-F (E) cultures are plotted on the X and Y-axis, respectively. For each protein, the changes in transgene and protein levels from CHO to HEK293 are represented by an arrow. The fold changes in protein levels in HEK293 cells are color-coded.


A)
1a. Stable episomal expression
1b. Transient expression
2. Evaluation of protein expression
Elements for plasmid retention and replication
Vector Panel of 24 transgenes
Side-by-side expression in CHO and HEK293
CHO
HEK
Cells and supernatants (biological duplicates)
HRP-conj. detection Ab
anti-HPC4 Ab
Technical duplicates
HPC4-tag
r-protein
Protein standards
Quantitation in supernatants
Flask cultures
7 days (HEK293) or 3 days (CHO)
Deep-well plates
72 hours
Cell and supernatant expression evaluation (1b)
B)
CHOEBNALT85-1E9
293ALL
Concentration (µg/ml)
C)
Freestyle CHO-S
293-F
Freestyle 293-F
D)
E)
1 + Titer (ug/mg)
1 + Transgene TPM
Titer fold change
-5.0 -2.5 0.0 2.5 5.0

In the small-scale fully transient expression setting the improvement in secreted expression using HEK293 compared to CHO was even more profound. Protein titer estimations of cell culture supernatants are shown in Fig. 4C. The results showed more than two-fold higher secreted titers of 16 of the 24 difficult to express proteins in both of the closely related HEK293 cell lines compared to the CHO cell line. Only one protein, PLG, showed at least two-fold higher expression in CHO cells compared to HEK293. For half of the evaluated genes (12 of 24), no or only traces of protein could be detected in supernatants from CHO cells. Moreover, for nine of the investigated r-proteins (THBS4, CCL28, CXCL13, CCL20, HGF, PAMR1, LOX, NRTN and NRTN pp) no or only traces of protein could be detected in both the culture supernatant and cell lysate of CHO cells. In the case of expression in HEK293 cells, only four genes (PAMR1, LOX, ADM5 and PLG) resulted in no or only traces of secreted protein in both HEK293 cell lines. However, three of these proteins (PAMR1, LOX and ADM5) could be detected in cell lysates of both HEK293 cell lines, suggesting inefficient secretion from HEK293 cells. Strikingly, only one of the 24 investigated proteins, PLG, could not be detected in either cell lysates or cell supernatants in any of the HEK293 cell lines. For a subset of supernatant samples, the relative titer change between cell lines were confirmed using liquid chromatography tandem mass spectrometry (LC-MS/MS) combined with protein quantification based on the SIS PrEST technology[79]. Results from the MS/MS analysis showed the same expression trends between HEK293 and CHO cells for each r-protein investigated compared to the western blotting data. As the two methodologies in this case depend on different parts of the polypeptide sequences for r-protein detection, depending on r-protein processing in the samples and differences in r-protein processing between cell lines, protein titers of either method reported here should be considered estimates and not absolute quantities.

Taken together, 8 out of 24 genes (BMP10, ARTN, THBS4, AMBP, FSLT3, NRTN, CXCL13 and CCL28) consistently expressed better in HEK293 cells in both the semi-stable and

the transient expression comparison. For CHO, the only gene that expressed better compared to HEK293 in both setups was PLG.

Transcriptome profiling showed variation in secretory pathway utilization between HEK293 and CHO driven by limited set of gene outliers

Based on the observed improvement of the expression of several human r-proteins in HEK293 compared to CHO, a transcriptomic comparison between the two cell lines was performed with emphasis on r-protein transcript levels and host gene expression patterns. Initially, transcript levels for each transgene in the panel were quantified and compared to the r-protein titers in supernatants. Overall, both HEK293 cell lines showed elevated transgene transcript levels compared to CHO cells, consistent with the generally higher transfection efficiency observed for HEK293 compared to CHO (Fig. 4D-E). However, neither the transgene transcript abundance nor its fold changes correlated with secreted protein titers. Notably, some of the proteins with the highest transgene abundances (CCL20, CCL28, CXCL13 and ADM5) displayed among the lowest titers of secreted protein, or no secretion at all. This suggests that such extreme transgene mRNA levels may come at the cost of endogenous gene expression. Alternatively, this could indicate inefficient translation of the transgene mRNAs resulting in an accumulation of untranslated mRNAs in the cell. Moreover, among the proteins with the highest secreted titers in HEK293 (FN1 and THBS4), mRNA levels of the transgene were among the lowest in the data set (between 40 – 4200 TPM).

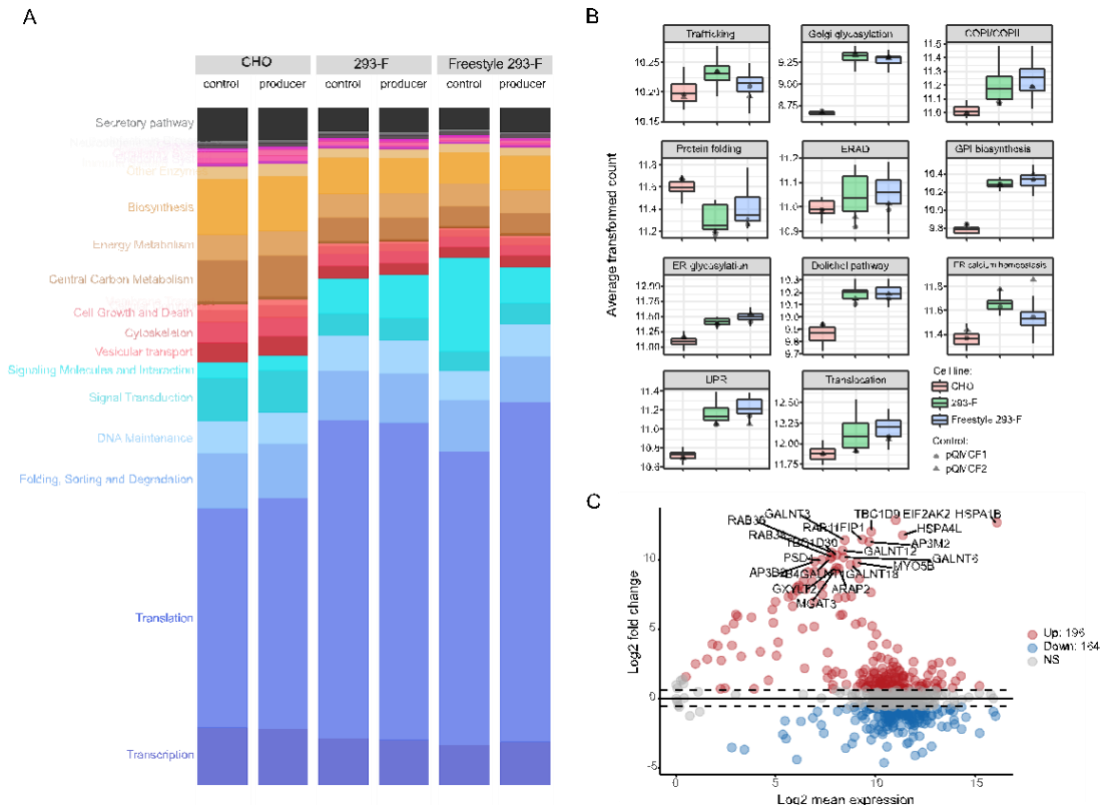


Figure 5. The utilization of the secretory pathway is distinctly different between CHO and HEK293 as a result of a limited set of gene outliers.
(A) Comparison of overall transcriptome usage across cell lines and producers. The fraction of the transcriptome dedicated to various cellular functions is represented by the height of each bar, where related pathways are colored similarly. The transgene mRNA levels were excluded. (B) Overall gene expression of 11 functional secretory pathway subgroups in CHO and HEK293 upon overexpression of target genes. For each group, the average gene expression level for the sample was computed and plotted for all samples. Black dots represent average expression in each secretory pathway group for control samples with an empty vector (pQMCF-plasmid). (C) Differential expression MA-plot showing the mean expression and fold-changes for the secretory pathway genes between HEK293 and CHO. Positive fold-changes denote higher expression in HEK293 and vice versa. The top 20 most differentially expressed secretory pathway components are labeled.

Furthermore, the mRNA expression levels across major molecular processes in CHO and HEK293 cells were quantified, both for cells expressing the panel of r-proteins (expressing cells) and cells transfected with empty plasmid (non-expressing cells) (Fig. 5**A**). Between expressing and non-expressing cells within each cell line, the transcriptome usage was similar for both CHO and HEK293, with the exception of transcription of genes related to translation and signaling molecules in the Freestyle 293-F cell line. Overall, HEK293 and CHO cells showed great variation

in the utilization of their respective transcriptomes. Most evidently, genes associated with translation showed lower mRNA expression on average in CHO cells, compared to HEK293 cell lines, whether a transgene is being expressed or not. In addition, the CHO cells showed a higher proportion of their transcriptome expression focused on biosynthesis, central carbon metabolism and signal transduction. Interestingly, HEK293 cells showed an overall less active secretory pathway compared to CHO (Fig. 5A), despite secreting higher amounts of r-proteins. However, the overall gene expression levels were generally higher in HEK293 compared to CHO of all secretory pathway subgroups, with the exception of protein folding that was significantly higher in CHO (Fig. 5B). Instead, the higher fraction of the transcriptome devoted to the secretory pathway in CHO cells compared to HEK293 was due to a very small subset of highly expressed secretory pathway genes whereas HEK293 cells, on the other hand, expressed a more diverse set of secretory pathway genes. While the fraction of the transcriptome, across all samples, that was utilized for secretory pathway genes did not correlate with transgene expression levels nor estimated secreted r-protein titers, there was a linear relationship between expression of several secretory pathway genes and transgene mRNA abundances in CHO producers. In HEK293 cells however, peak secretory pathway activities occur in clones with low to medium transgene load. This suggested a saturation of the secretory pathway in HEK293 cells, which may be a result of the exceptionally high transgene mRNA loads observed in case of several transgenes. Across all samples, a significant negative correlation was observed between r-protein titer and gene expression within the protein folding and ER glycosylation functional groups, respectively. Focusing on individual secretory pathway genes, overall similar expression levels were observed between the two cell lines with the exception of a limited set of extreme gene outliers (Fig. 5C). Amongst these outliers, there was a substantial group of genes with very low or no expression in CHO whereas the expression in HEK293 was moderate to high (Fig. 6A). Within this group of genes we observed several genes with previous support of impact on r-protein secreted titers[71,80–89].

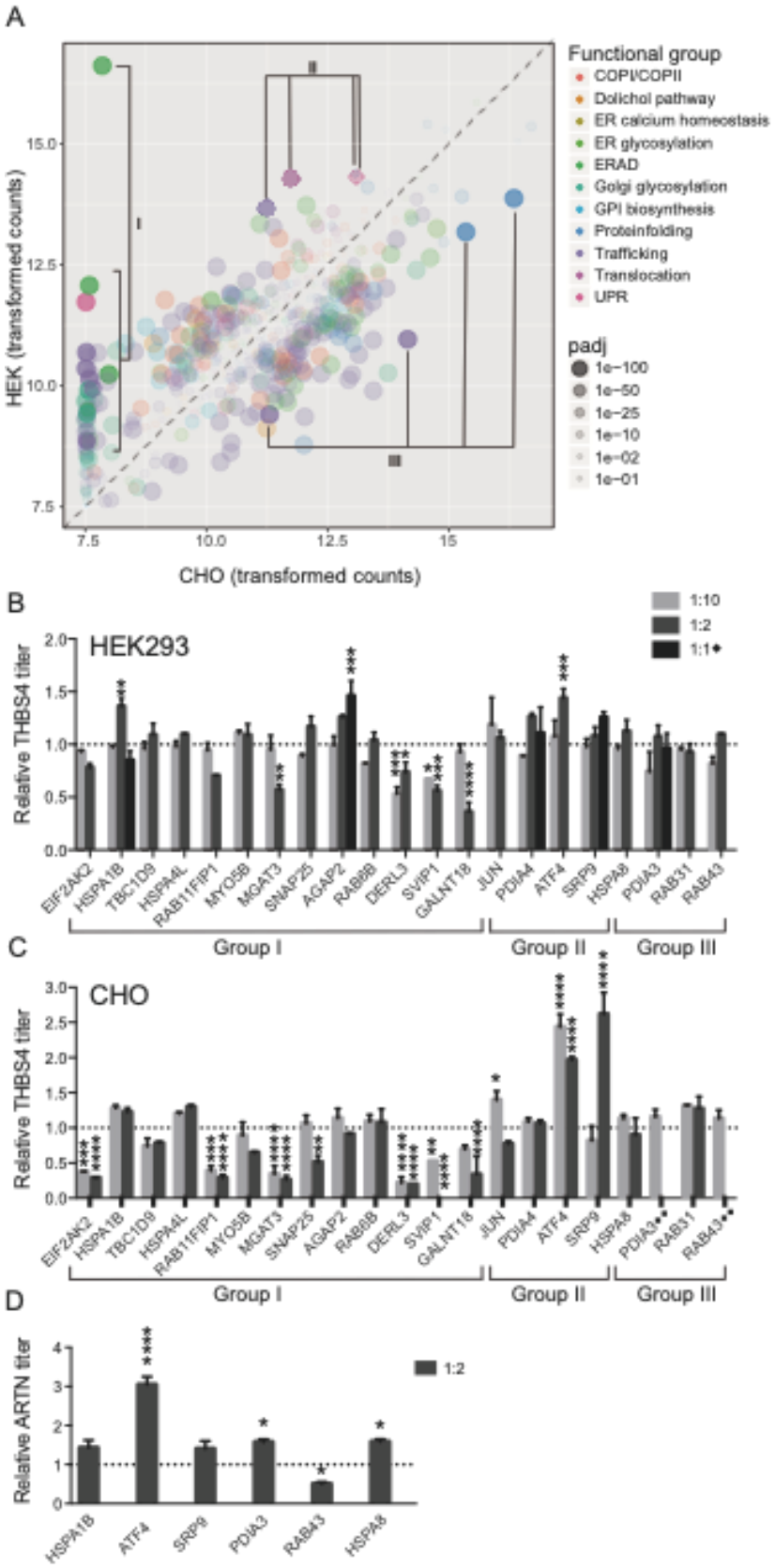
Highly expressed helper proteins in HEK293 compared to CHO have a positive impact on secretion of difficult to express proteins when co-expressed also in CHO cells

We hypothesized that the secretory pathway genes with extreme expression variation between the cell lines may contribute to profound differences in productivity observed between the cell lines for some human difficult to express proteins. Consequently, a set of secretory pathway genes was co-expressed with the difficult to express protein THBS4 in CHO and HEK293 to evaluate their impact on secreted productivities. The selection of gene outliers to evaluate was based on the highest differential expression between the cell lines but also on previous literature supporting potential roles in protein secretion or demonstrated effects on bioproductivity[71,80–89]. The selected genes were divided into three groups (I, II and III) based on expression levels in the two cell lines (Fig. 6A). Three genes (HSPA1B, AGAP2 and ATF4) had a small, albeit significant, positive effect on the secreted titer of THBS4 compared to cells only expressing THBS4 (Fig. 6B). For CHO cells a more profound titer improvement was observed with genes in group II (Fig. 6C), including SRP9, ATF4 and JUN that had moderate endogenous expression in CHO but higher expression in HEK293. More than two-fold improvements in secreted THBS4 titers were observed in CHO when co-expressed with ATF4 or SRP9, and a 1.5-fold titer increase was observed when co-expressed with JUN. In addition, slight improvements (however not statistically significant) of THBS4 titers in CHO cells were observed when co-expressed with some secretory pathway genes from groups I and III such as HSPA1B, HSPA4L and RAB31, depending on the plasmid ratio between the co-expressed transgenes. Validation of a subset of the gene outliers in combination with ARTN in CHO cells (Fig. 6D) showed a significant positive impact of ATF4, PDIA3 and HSPA8 on ARTN secretion. Moreover, HSPA1B and SRP9 overexpression generated a small increase, albeit not significant, in ARTN titers. Higher secreted ARTN titers, but not THBS4, in CHO cells were associated with lower viable cell densities and viability at harvest compared to controls. This may relate to differential effects of the two r-proteins on cells. In

addition, several outlier secretory pathway genes of group I (EIF2AK2, RAB11FIP1, MGAT3, DERL3, SVIP1 and GALNT18), with low or no expression in CHO, but moderate to high expression in HEK293, significantly decreased secreted THBS4 titers dramatically when added exogenously in CHO. A similar trend was observed for some of these genes upon co-expression in HEK293 even though the negative impact on secreted THBS4 titers was not as profound. In the case of some of these co-expressed genes (DERL3, MGAT3 and EIF2AK2), these effects were likely the result of a negative impact on cell growth in CHO but not in HEK293, suggesting that these genes are not compatible with the CHO cell machinery of maintaining cellular growth and productivity.

Since both THBS4 and ARTN were profoundly better expressed in HEK293 compared to CHO and the low expression of these proteins in CHO could be rescued by the overexpression of secretory pathway components expressed at higher levels in HEK293, we hypothesize that such differences in secretory pathway components may be beneficial for the secretion of difficult to express proteins in HEK293.

Figure 6. Evaluation of outlier secretory pathway genes on secreted protein titers.
(A) Comparison of expression of individual secretory pathway genes, with transformed counts for CHO and HEK293 shown on the X- and Y-axis respectively. Most secretory pathway genes lie within the shaded region drawn around the identity line (x=y), showing an overall conserved pattern of expression. Differentially expressed secretory pathway genes evaluated based on effects of the expression of a difficult-to-express protein are highlighted and divided into three groups (I, II and III) based on expression levels in the two cell lines. All other genes are shaded. Mean relative secreted titers ± SD (N = 2) of THBS4 determined by ELISA in HEK293 (B) and CHO (C) when co-expressed with differentially expressed secretory pathway genes (x axis) compared to the expression level of THBS4 alone (empty vector). Plasmid ratios of 1:1, 1:2 or 1:10 (secretory pathway gene:THBS4) upon transfection were evaluated. ◆The 1:1 plasmid ratio was only evaluated for HSPA1B, AGAP2, ATF4, SRP9 and PDIA3 in HEK293 cells. ◆◆Only the 1:10 plasmid ratio evaluated were evaluated for PDIA3 and RAB43 in CHO cells. D) Co-expression results of ARTN expression levels (determined by western blot) when combined with a subset of the differentially expressed secretory pathway genes in B and C in CHO with plasmid ratios 1:2 (secretory pathway gene:ARTN). Mean relative ARTN titers ± SD (N = 2) between supernatants of cells co-transfected with secretory pathway genes versus no transgene (empty vector). Significant different expression of THBS4 (B and C) or ARTN (D) compared to the co-expression with an empty vector control (determined by one-way ANOVA and Dunnett's test) are indicated by the asterisk sign (* P adj ≤ 0.05; ** P adj ≤ 0.01; *** P adj ≤ 0.001; **** P adj ≤ 0.0001).


A
Functional group
COPI/COPII
Dolichol pathway
ER calcium homeostasis
ER glycosylation
ERAD
Golgi glycosylation
GPI biosynthesis
Proteinfolding
Trafficking
Translocation
UPR
padj
1e−100
1e−50
1e−25
1e−10
1e−02
1e−01
HEK (transformed counts)
CHO (transformed counts)
B
HEK293
1:10
1:2
1:1
Relative THBS4 titer
EIF2AK2
HSPA1B
TBC1D9
HSPA4L
RAB11FIP1
MYO5B
MGAT3
SNAP25
AGAP2
RAB8B
DERL3
SVIP1
GALNT18
JUN
PDIA4
ATF4
SRP9
HSPA8
PDIA3
RAB31
RAB43
Group I
Group II
Group III
C
CHO
Relative THBS4 titer
D
Relative ARTN titer
1:2

Proteasomal and propeptide convertase genes were differentially expressed between CHO and HEK293

To obtain a more fine-grained understanding of the cell line differences, differential expression analysis comparing HEK293 to CHO cells was performed. Due to drastic organismal differences, many genes outside of the secretion machinery showed distinct expression between CHO and HEK293 cells. In fact, more than 80% of the genes were significantly differentially expressed between the two organisms. Few canonical pathways were consistently up- or down-regulated in one cell line compared to another, as shown by gene-set enrichment analysis. However, among them the proteasome protein family, which degrades misfolded proteins in a controlled fashion, was expressed at significantly higher levels in CHO compared to HEK293 cells. Another protein family that showed significantly different expression profiles between HEK293 and CHO is the propeptide convertase family. HEK293 showed significant upregulation of PCSK2, PCSK4, PSCK5, PCSK6 and PCSK8 compared to CHO. On the other hand, CHO expressed higher levels of PCSK1, PCSK3 and PCSK7.

Differentially activated secretory pathway genes between HEK293 and CHO upon transgene expression

To evaluate overall differences in cellular dynamics between HEK293 and CHO cells when producing r-proteins and to better account for organismal disparity, we calculated the average activation of genes by determining the differential expression upon transgene expression across cell lines during r-protein production with non-producer cells as reference (Fig. 7A). While the majority of the genes were not differentially activated in either of the cell lines, several genes showed significantly opposite trends in HEK293 and CHO cells. The top 20 genes identified as differentially activated in producers between CHO and HEK293 strongly enriched for members of the secretory pathway (hypergeometric p-value < 0.0005). Among them, four genes (SEC61A1, SEC63, DNAJC3 and ERO1B) are directly involved in posttranslational functions of the secretory pathway. Co-expression evaluation of SEC61A1, DNAJC3 or ERO1B between HEK293 and

CHO, did however not result in improved secreted expression of THBS4 in either cell line. On the contrary, overexpression of all these genes had a significant detrimental effect on secreted titers in CHO cells. This may suggest that the activation of these genes is likely a result of, rather than the cause of a systemic response upon transgene expression and that further activation of these genes has no, or detrimental, effects on protein production.

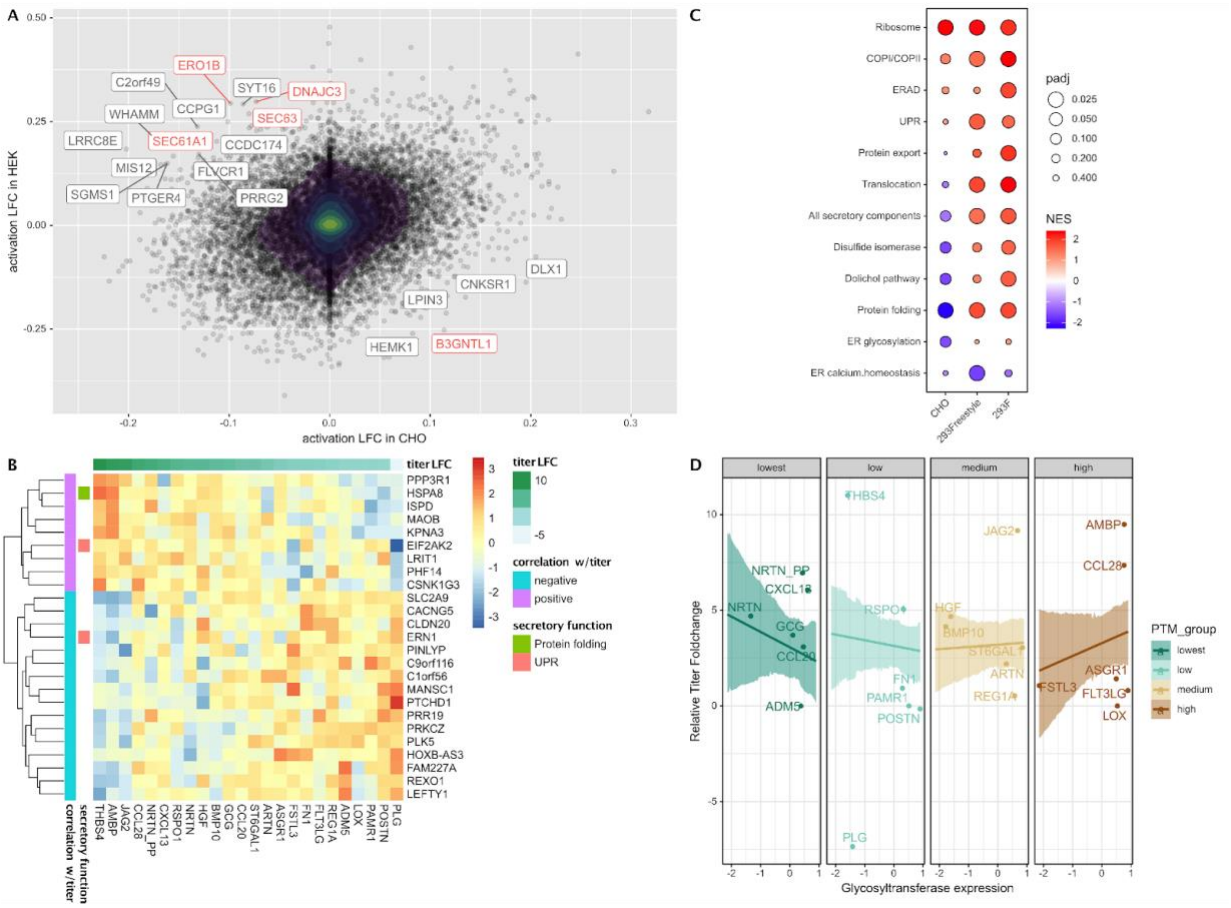


Figure 7. Transgene expression induces differential activation of genes between cell lines and between different r-proteins.

(A) Comparison of gene activation upon transgene expression between HEK293 and CHO. Log2 fold-changes (LFCs) are calculated based on the differential expression between producers and controls for CHO and HEK293 cells, and are plotted on the X- and Y-axis respectively. Genes showing the most divergent activation patterns between HEK293 and CHO are labeled. The top differentially activated genes are enriched for secretory pathway genes (colored red, hypergeometric p-value = 0.004). (B) Heatmap of endogenous secretory pathway genes with the highest positive or negative correlation between differential activation and r-protein titer change from CHO to HEK293. In each entry, a more positive fold-change of the endogenous secretory pathway gene (Y-axis label) for a given r-protein (X-axis label) indicates stronger activation in HEK293 compared to CHO when producing that protein. (C) Overall activation of secretory pathway subsystems. The dots in the scatter heatmap show pathway activities (y-axis) across cell lines (x-axis), with colors indicating activation/ suppression and sizes indicating the corresponding significance. (D) Differences in activation of O- and N-linked glycosyltransferases between HEK293 and CHO cells correlate with titer improvement of moderately- to heavily-glycosylated r-proteins. R-proteins harbouring more frequent N- and O-glycosylation sites per AA residue (rightmost panel) tend to benefit more from increased expression of glycosyltransferases from CHO to HEK, while a reversed trend was observed for non- and lightly-glycosylated r-proteins (leftmost panel).

To see pathways that were preferentially activated in each cell line, gene set enrichment analysis (GSEA) was performed using the fold changes of differential activation for each cell line (Fig. 7C). The producers in all three cell lines significantly upregulated genes involved in translation. However, HEK293 cells showed higher expression for genes related to protein secretion, compared to CHO cells. For example, translocation and protein export were much more strongly activated in both 293-F and Freestyle 293-F. Genes involved in protein folding such as molecular chaperones, were significantly downregulated in CHO cells, while significantly upregulated in the HEK293 cell lines. Similar pathway activation was observed between the two HEK293 cell lines variants, with the only notable exception being proteasomal function, whose activations were more strongly in Freestyle 293-F than in 293-F.

To systematically identify genes with potential impact on productivity of the cell lines, we calculated for each gene the correlation between its differential activation and the r-protein titer changes from HEK293 to CHO across clones. Overall, only a small number of the genes displayed r-protein titer change-dependent differential activation. The top genes with the highest positive or negative correlation are given in Fig. 7B. Results showed that three secretory pathway genes (EIF2AK2, HSPA8 and ERN1) correlated in titer and activation change between HEK293 and CHO. Interestingly, HSPA8 and EIF2AK2 were also found amongst the genes with the highest expression fold-change between HEK293 and CHO cells.

Beyond the differences in titer improvements, the panel of r-proteins are diverse in their PTM compositions, utilizing distinct sets of enzymes. With this, we explored whether the differential expression of the enzymes responsible for some of the PTMs between CHO and HEK293 can explain the variation in titer improvements seen by different r-proteins. We posited that proteins with more frequent PTM sites may be more sensitive to changes in the expression of the enzyme responsible for the PTM in question. To quantify the degrees to which certain PTMs are overrepresented in each r-protein, we calculated a “PTM-index” for each r-protein - PTM combination based on the PTM site densities in each r-protein. Among the three most ubiquitous

PTMs taking place within the secretory pathway- disulfide bond, GPI anchor and N-/ O-linked glycosylation, we saw significant interaction between glycosylation-index and enzyme expression in determining titer improvement. More specifically, the titer improvement for more heavily glycosylated proteins showed a strong positive correlation with the differential expression of glycosyltransferases, whereas for non- and lightly-glycosylated r-proteins, the changes in titer were negatively correlated with the glycosyltransferases fold change (Fig. 7D).

Discussion

The CHO cell line is a well-established bioproduction host with readily available expression protocols both in transient and stable settings. However, with the increasing number of new biologics approaching the market, including next-generation biologics such as engineered scaffold proteins and antibody-fusion proteins, the pharmaceutical industry faces new challenges for efficient protein production. But even natural human proteins can pose challenges for bioproduction both in transient and stable expression systems and require laborious process optimization. Due to the clonal divergence of different immortalized cell lines[90–95], the expression of r-proteins may vary considerably between hosts even of the same origin. Indeed, this was observed in this study, where the two CHO-S based expression platforms, QMCF and ExpiCHO, showed protein-dependent differences in secreted titers. The higher secreted titers of smaller r-proteins and/or more heavily glycosylated r-proteins observed in ExpiCHO compared to the QMCF platform (Fig. 3C) may relate to the high cell densities of cultivation and high amount of plasmid DNA added upon transfection in the ExpiCHO system. We speculate that this may put growth pressure on the cells, making this platform better suited for producing smaller proteins. On the contrary, the lower cell densities of the QMCF system in combination with the reduced cultivation temperature (30°C instead of 37°C) and longer cultivation times may result in lower cellular stress compared to the ExpiCHO system, enabling production of larger proteins. Changes in titer between the platforms could also be related to clonal differences between cell lines, even

though the overall transcriptome utilization is comparable between the CHOEBNALT85 and ExpiCHO cell lines. Thus, the data suggest that each system can provide protein-specific advantages possibly related to platform differences. On the other hand, a r-protein-independent improved secreted expression of challenging human proteins was observed when changing expression host from CHO to HEK293 (Fig. 4). One third of the proteins were expressed at more than 2-fold higher titers in HEK293 compared to CHO in both systems.

Overall, a part of the increased expression in transient cultivations of HEK293 can be explained by an increased transfection efficiency and plasmid uptake compared to CHO reflected in the overall higher mRNA abundances observed in HEK293 (Fig. 4), in line with previous observations that HEK293 tend to perform well as a transient expression host and are easier to transfect compared to CHO[73,96]. However, for some r-proteins extremely high transcript levels resulted in very low or no secreted product, suggesting that such extreme transgene mRNA levels could come at the cost of endogenous gene expression. Moreover, since one third of r-proteins showed increased titers in the HEK293 cell lines compared to CHO also in the semi-stable expression system of QMCF, which provides comparable transfection efficiencies between HEK293 and CHO cells (personal communication with Icosagen), we argue that a subset of human difficult to express proteins that consistently express better in HEK293 compared to CHO in both transient and episomal stable expression settings were likely also a result of differences related to the secretory machinery of the two cell lines.

Transcriptomic analysis of transiently expressing HEK293 and CHO cells showed a profound difference in the overall utilization of the transcriptomes between CHO and HEK293 (Fig. 5A), which is expected due to the different origins of the cell lines. The higher translational machinery activities in HEK293 cells may afford them increased capacity for translating mRNA, although the level of transgene mRNAs seen in this study should not saturate the ribosomal capacity, as transgene with mRNA levels ranging upwards of 20% of total mRNA content has been shown to translate efficiently[38]. As the secretory pathway is a major determinant of the titers

of secreted proteins[97], comparisons between the cell lines focused on secretory pathway components. Higher activities were observed in the protein quality control pathways of UPR and ERAD in HEK293 cells compared to CHO (Fig. 5B), which may impact protein secretion as upregulated transcription of genes associated with these pathways can increase secretory capacities of host cells[98,99]. Moreover, genes involved in protein folding were more highly expressed in CHO cells compared to HEK293. Notably, protein folding showed a significant association with decreased protein titer (FDR p-value = 0.0023). This could be a cellular response to increased difficult-to-express protein load, especially if the native machinery for r-protein folding is lacking. At the gene level, most secretory pathway genes have similar expression between the two cell lines (Fig. 5C and Fig. 6A). However, a limited set of genes showed extreme variation in expression between the cell lines. For instance, CHO cells do not express many of the moderately expressed secretory machinery genes expressed in HEK293 cells. Since the r-proteins in our screen are all human proteins, it is possible that the lack of compatible secretory components forced the CHO cells to utilize a smaller subset of more generic machinery components, and this lack of specialization could possibly impact secreted titers. Alternatively, the absence of expression of such genes in CHO may be compensated by the expression of other genes without a human ortholog and hence not included in our analysis. Notably, amongst the genes more highly expressed in HEK293 compared to CHO, we identified several examples of genes that have previously been associated with improved protein production. For instance, both ATF4 and SRP9 have previously been associated with improved recombinant expression in CHO cells either alone or in combination with other ER components[71,86,87]. In this study we could show a profoundly positive effect of ATF4 also on secreted production of two difficult to express proteins in CHO cells (Fig. 6), suggesting that this protein act in a more universal way of improving yield. Overexpression of SRP9 significantly increased THBS4 titers dramatically and had a slightly positive, but not statistically verified, impact on ARTN secretion. The already high endogenous expression of these two genes in HEK293 compared to CHO may explain why the positive effect

on secreted titers was not as profound in this cell line and moreover, we hypothesize that such differences in secretory pathway components may be beneficial for the secretion of difficult to express proteins in HEK293. In addition, several genes with more moderate impact on the secreted expression of r-proteins in either CHO or HEK293 were identified in this study, supporting the usage of transcriptomic data to shine light upon secretory pathway differences that impact bioproductivity between cell lines. In HEK293 cells, AGAP2, HSPA1B and ATF4 significantly boosted THBS4 secretion, whereas in CHO cells, SRP9, JUN, PDIA3 and HSPA8 had significant positive impact on secretion of either THBS4 or ARTN. As summarized in Fig. 8, such genes could indicate productivity bottlenecks in CHO cells when expressing a difficult to express protein.

Among the top 20 most differentially activated genes, four ER-associated genes (SEC63, SEC61A1, DNAJC3 and ERO1B) were found upregulated in HEK293 but not in CHO upon transgene expression (Fig. 7A). Even though SEC61A1 along with the other subunits of the translocon has previously been shown to improve the specific productivity of a difficult-to-express antibody in CHO cells[71], the individual overexpression of ERO1B, DNAJC3 or SEC61A1 along with THBS4 showed a significantly detrimental effect on productivity in CHO cells in this study. This may support a less active role of these proteins in the CHO secretory pathway. On the other hand, the significant upregulation of these genes in HEK293 may be well tuned by the cells without further improvements added by exogenous overexpression. Interestingly, differential activation between CHO and HEK293 were overall independent of protein identity or r-protein titer change between the cell lines. This observation suggested that the differential activation upon r-protein expression was mainly driven by cell line differences rather than protein identity or load and highlights cell-line dependent variation in utilization of cellular machinery upon protein production. However, a subset of differentially activated genes correlated in activation fold-change and titer change between the cell lines (Fig. 7B). Three of these (EIF2AK2, HSPA8 and ERN1) are part of the secretory machinery, out of which HSPA8 has previously been recognized as a marker for

protein productivity in CHO cells[81]. Notably, EIF2AK2 and HSPA8 were also recognized as extremely differentially expressed between CHO and HEK293. Overexpression of HSPA8 had a slightly positive effect on THBS4 secretion in HEK293 and CHO and a significantly positive effect on ARTN secretion from CHO cells. It's worth noting that HSPA8 showed stronger activation in HEK293 compared to CHO in THBS4 producing clones, and a reversed activation trend was observed in ARTN producing clones (Fig. 7B). Hence, HSPA8 expression could be part of the secretory machinery orchestrating differences in protein secretion between the cell lines. On the other hand overexpression of EIF2AK2 showed a negative effect on THBS4 secretion, especially in CHO cells where endogenous EIF2AK2 levels were much lower compared to HEK293. Notably, the correlation between EIF2AK2 activation and titer fold-change between the cell lines was mainly driven by the strong negative activation fold-change of EIF2AK2 in HEK293 producing PLG compared to all other transgenes in the panel. As EIF2AK2 encodes a protein involved in ER stress and the UPR, causing inhibition of mRNA translation, this remarkably lower differential expression activation of EIF2AK2 in HEK293 cells expressing PLG, along with non-detectable secretion of the PLG r-protein, may suggest that PLG expression is limited already at steps prior to ER processing in HEK293 cells. On the other hand, as HSPA8 and ERN1 showed a more consistent correlation across transgenes these genes may be universal secretory pathway markers for secreted titer differences between the cell lines across a variety of r-proteins (Fig. 8).

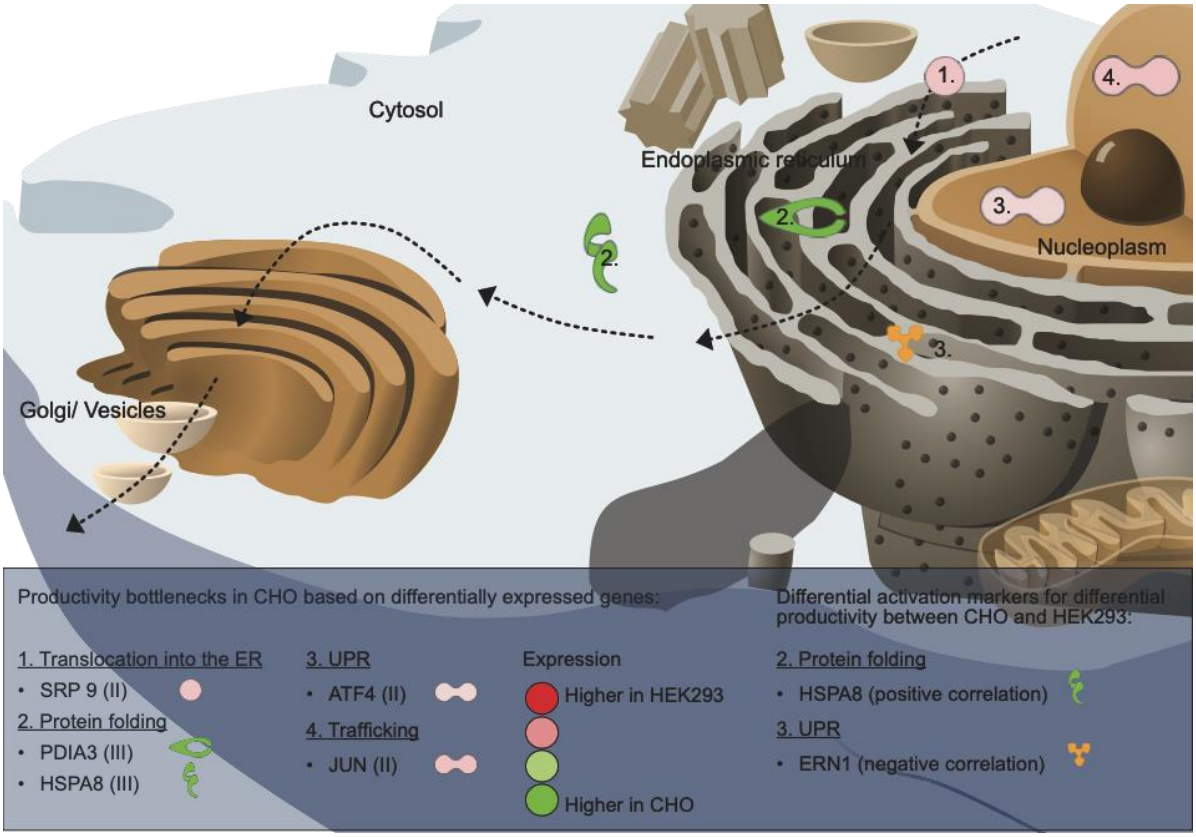


Figure 8. Overview of secretory pathway components with significantly positive impact on productivity in CHO and differential activation markers correlating with differential productivity between CHO and HEK293.
The arrows indicate the secretion path of secretory proteins in the cell. Differentially expressed genes between CHO and HEK293 with significantly positive impact (P adj. ≤ 0.05) on THBS4 and/or ARTN titers upon overexpression in CHO are indicated as productivity bottlenecks in CHO cells. The color scale from red to green indicates gene expression fold-change between HEK293 and CHO, where genes of similar expression change between the two cell lines are also grouped into one of three groups (indicated by the roman numerals I, II and III, further described in Figure 4). Differentially activated secretory pathway genes correlating with titer fold-changes between HEK293 and CHO serve as activation markers for titer fold-change of difficult to express protein between cell lines. HSPA8 showed a positive correlation between differential activation upon transgene expression in HEK293 vs. CHO and titer-fold change in HEK293 vs. CHO, whereas ERN1 had a negative correlation. All gene product symbols are mapped into their respective cellular compartments and the numbers indicate the secretory pathway subgroup of these gene products.

At pathway levels, all cell lines responded to transgene expression by up-regulating ribosomal activities (Fig. 7C). However, CHO cells struggled with activating the components responsible for the downstream processing and export of the r-proteins. which may explain why

overexpressing a single gene in most cases only had minor positive effects or failed to boost the expression of THBS4, as the genes in pathways showing deficient expression in CHO cells likely work in tandem, and the overexpression of just one gene is often not sufficient to activate these under-expressed pathways.

While some secretory pathway genes identified to have profound positive effects on r-protein secretion seemed to assist r-protein expression indiscriminately, mounting evidence has suggested a product-specific role for many of the components within the secretory pathway. Genetic perturbation studies targeting the secretory pathway revealed that different secreted proteins utilize distinct sets of secretory pathway components during their production[10,100–102]. Further lending credence to this product-specific nature of the secretory pathway was the secreted protein-dependent expression of the secretory pathway components. For example, protein disulfide isomerase (PDI) expression and disulfide-rich protein secretion rates are correlated across human tissues[48] and mouse lymphocytes[103]. While our results show no significant PDI-dependent titer improvements, the availability of N- and O-linked glycosyltransferases can restrict titer improvements of glycosylation-enriched r-proteins (Fig. 7D). Notably, increased glycosyltransferase activities seemed to decrease the expression of lightly- and non-glycosylated r-proteins, suggesting that when underutilized, the metabolic costs of certain cellular resources can easily outweigh their benefits in protein secretion. This cautions against the pursuit of an omnipotent host cell line and highlights the importance of customizing engineering strategies according to the properties of the r-proteins. Other key factors that can have protein-specific impacts on the secreted protein titers are e.g. RNA instability[104–106], the choice of signal peptide within the transgene sequence[107,108] and proneness to proteolytic degradation[109–111]. Such issues should be addressed for each specific r-protein for thorough bioproduction optimization.

In summary, we show that HEK293 can serve as a valuable fall-back expression strategy, for difficult or non-secreting proteins expressed in CHO cells, or that comparisons between the different host cells can guide efforts to rescue poor expression in CHO. Taken together the results of this study shine light on the variation in expression and activation of secretory pathway related genes between HEK293 and CHO. Such cell line-specific variations could have an impact on the optimal choice of bioproduction host for specific r-proteins depending on the requirement for specific secretory pathway processing. We hypothesize that the existence of a collection of secretory machinery that better conforms to our panel of human proteins in the HEK293 cell lines is key to their improvements in protein titers. Indeed, amongst the most profound differences in expression between HEK293 and CHO secretory pathways, genes with especially positive impact on protein secretion in CHO were found. Although many of the secretory machinery components promiscuously assist the secretion of different proteins[112], there are reports of more product-specific improvements to the secretory machinery[100,113,114]. Supporting this, results highlighted the N- and O-linked glycosyltransferases as a group of genes aiding, or restricting, protein secretion in a protein specific manner. These highly product-specific genes enable bespoke cell line designs that cater to the unique secretory requirements of different r-proteins, and allows for a more rational selection of cell hosts for a given r-protein.

Materials and Methods

Experimental Design

This study was performed in three main steps. Initially, difficult to express r-proteins were produced in various expression systems (various cell lines and protocols) to evaluate performance differences in protein-specific or unspecific secreted production as determined by absorbance at 280 nm of purified r-proteins or western blotting and LC-MS/MS analysis of cell culture supernatants. In a second step, a transcriptomic evaluation of cell lines with distinct differences in performance was conducted in order to evaluate underlying secretory pathway components with impact on r-protein secretion, including differential expression analysis, gene

set enrichment analysis and protein feature analysis. Genes of interest, identified by the transcriptomic profiling were subjected to co-expression analysis together with a difficult to express protein in CHO and HEK293 to evaluate the impact on r-protein secretion. ELISA or western blotting was used to determine relative r-protein titers compared to expression of the difficult to express protein alone.

Cell lines and medium

ExpiCHO-S (Gibco™) cells were cultivated in ExpiCHO expression medium. 293-F (Gibco™) and Freestyle™ 293-F (Gibco™) cells were cultivated in FreeStyle™ 293 expression medium (Gibco™). FreeStyle™ CHO expression medium (Gibco™) supplemented with 8 mM GlutaMAX™ (Gibco™) was used for Freestyle™ CHO-S cells (Gibco™). Cells were cultivated in 125 ml Erlenmeyer shake flasks at 37°C, 8°C CO_2 and 125 rpm. CHOEBNALT85, CHOEBNALT85-1E9 and 293ALL cells were cultivated according to manufacturer's recommendations (Icosagen Cell Factory OÜ, Tartu, Estonia).

Plasmids and expression constructs

For expression validation of difficult to express proteins the pQMCF vector or the in house designed pKTH16 or pKTH16_dPur plasmid was used. Expressions in both vectors are driven by the CMV promoter. The pQMCF generic expression cassette included an N-terminal CD33 leader sequence[115] followed by a short spacer sequence (AAA) and a C-terminal TEV and human protein C tag[116]. The pQMCF-plasmid without gene insert served as empty vector control. Moreover the in house pKTH16 plasmid was used as expression vector with the transgene expressed fused to a N-terminal CD33 signal peptide (no spacer sequence between signal peptide and mature sequence) and a C-terminal human protein C-tag. Transgenes for co-expression validation in combination with a difficult to express protein were cloned into the pKTH16 vector in fusion with a C-terminal FLAG tag. An empty pKTH16 vector, not encoding a transgene, was used as negative control in co-expression experiments.

ExpiCHO transfection, cultivation and harvest

The pKTH16_dPur plasmid was used for expression of transgenes. The transfections and cultivations were performed according to the manufacturers ExpiCHO standard protocol. One day prior to the transfection, the cells were seeded at $3x10^6$ cells/ml. At the day of transfection, the cells were split to $6*10^6$ cells/ml in 25 ml ExpiCHO expression medium. The ExpiFectamine reagent and 20 µg of plasmid DNA were diluted separately in OptiPRO SFM and then mixed together and incubated at room temperature for three minutes before addition to cells. The cells were cultivated at 37°C, 8% CO_2 and 125 rpm. The day after the transfection (18-22 hours post-transfection), ExpiFectamine CHO Enhancer and ExpiCHO Feed were added to the cells. The cells were harvested at day 8 post-transfection.

Affinity protein purification

Recombinant proteins were purified using the Anti-Protein C Affinity Matrix (11815024001, Roche) on an ASPEC liquid handling instrument (Gilson). The matrix was washed three times with equilibration buffer (20 mM Tris, 0,1 M NaCl, 2 mM $CaCl_2$, adjusted to pH 7,5). The protein sample was filtered through a 0,45 µm filter and then incubated with the purification matrix overnight on a rock n roll at 4°C prior to packing of the matrix-protein mixture on columns. The column was equilibrated with 20 mM Tris, 0,1 M NaCl, 2 mM $CaCl_2$, at pH 7,5 and washing was performed with 20 mM Tris, 1 M NaCl, 2mM $CaCl_2$, at pH 7,5. HPC4-tagged proteins were eluted by EDTA (20 mM Tris, 0,1 M NaCl, 5 mM EDTA, pH 7,5. The elution fractions were loaded on a SDS-PAGE gel to examine the purity and the yield of the elution fractions. The elution fractions showing strong and pure bands were desalted and buffer exchanged to 1xPBS. Protein concentrations were determined by absorbance measurements at 280 nm.

Medium-scale episomal stable expression (pQMCF system), cultivation and harvest in CHO and HEK293

293ALL and CHOEBNALT85-1E9 cells 007 (Icosagen, Tartu, Estonia) were transfected using Reagent 007 (Icosagen, Tartu, Estonia) and cultivations were performed at 30-35 ml scale at 37°C for the first three days, followed by incubation at 30°C until end of cultivation. Supernatant

and cells were harvested at day 7 for 293ALL and day 9 for CHOEBNALT85-1E9. The supernatant was clarified by centrifugation at 1800 xg for 45 min at 20°C. Cells were stored in RNAlater™ stabilization solution (Invitrogen™) for downstream RNA extraction.

Small-scale transient transfection, cultivation and harvest in CHO and HEK293

The pQMCF plasmids encoding 22 difficult to express target proteins and the pKTH16 plasmid encoding NRTN and NRTN pp were used for transient expression in Freestyle™ CHO, 293-F™ and Freestyle™ 293-F. At 24 hours before transfection, cells were split to 0.6 (Freestyle™ CHO) or 0.7 (293-F and Freestyle™ 293-F) million cells/ml in 125-ml Erlenmeyer shake flasks (Corning). On the day of transfection, the culture medium was exchanged by centrifugation and cells were resuspended in fresh medium at a cell density of 1 million cells/ml. Cells were transfected with 25 kDa linear PEI (Polysciences Inc.) at a DNA:PEI ratio of 1:4 (Freestyle™ CHO) or 1:3 (293-F and Freestyle™ 293-F) where 1 µg DNA was added per 1 million cells. Each plasmid was transfected in duplicate wells. The pD2529-CMVM-03 plasmid (Atum) expressing DasherGFP was used to monitor the transfection. Cultivation was performed in 24 deep-well plates with 2 ml cell suspension per well at 37°C, 8% CO_2 and 250 rpm in humidified incubators. At 24 h post-transfection, the transfection efficiency was monitored by flow cytometry (Gallios Flow cytometer, Beckman Coulter) of GFP-expressing cells based on mean fluorescence intensities in the FL-1 channel. At 72 hours post transfection the cell culture was harvested. Cells and supernatants from each well were separated by centrifugation (500 x g, 3 min). Half of the cell pellet was resuspended and stored in RNAlater™ stabilization solution (Invitrogen™) according to manufacturer's recommendations for subsequent RNA extraction.

Expression level evaluation and protein characterization by western blot

For western blot analysis of samples, cell pellets were initially lysed in M-PER solution (Thermo Fisher Scientific) and samples were separated by SDS-PAGE (Criterion TGX Precast gels, 4-12%, Bio-Rad) under denaturing conditions. Proteins were transferred onto PVDF membranes (Trans-Blot Turbo Transfer Pack, Bio-Rad) using the Trans-Blot Turbo Blotting

System (Bio-Rad), followed by blocking of membranes with 5% milk in TBST (0.05% Tween-20). Washing of membranes was performed with TBST and r-proteins were stained using a primary anti-HPC4-antibody (0.2 µg/ml, Icosagen) followed by the secondary goat anti-human HRP-conjugated antibody (1:4000, A18805, Invitrogen). Stained proteins were detected using Immobilon Western Chemiluminescent HRP Substrate (Millipore) and image acquisition using a ChemiDoc Imaging system (Bio-rad). For protein abundance estimations, each supernatant sample was run in duplicate and volumetric band intensities were fitted onto a standard curve generated by a dilution series of HPC4-tagged EPO on the same membrane using the Image Lab software (Bio-Rad).

Transcriptome profiling

RNA was extracted using RNeasy plus Mini Kit (Qiagen) according to manufacturer's guidelines and the quality of the isolated RNA was evaluated by BIOanalyzer 2100 (Agilent, Santa Clara, CA) using the Agilent RNA 6000 Nano Kit. RNA sequencing was performed at GATC Biotech (Konstanz, Germany) using the Inview transcriptomics Discover platform. Sequence data for RNA-Seq were quality controlled using FastQC and summarized with multiQC[117]. Trimmomatic[118] was used to trim low-quality bases from the reads. The CHO-K1[1,119] and the human GRCh38.p12 reference genomes were extended to incorporate the transgene sequences so that the transcripts of the heterologous secretome can be quantified. Reads were then quasi-mapped to either the extended CHO-K1 or the human GRCh38.p12 genome based on the cell line of origin and quantified with Salmon[120]. To compare the transcriptome usage across CHO and HEK293 cell lines, an ortholog conversion table[38] was used to convert CHO genes to their human orthologs. The functional groups and the color scheme for the transcriptome usage were adapted from the Proteomaps tool[121].

Differential expression was performed using DESeq2[122]. To facilitate the comparison between CHO and HEK293 cells, CHO genes were converted to human orthologs using the conversion table described above. With tximport[123], the transcript-level abundances were

integrated into gene-level counts to be compatible with DESeq2. Three different types of comparisons were carried out by specifying the corresponding design matrices. For cell line comparisons, gene expression profiles from all producers were compared between HEK293 and CHO cells. To estimate the degree of gene activation upon recombinant production, the CHO and HEK293 producer cells were compared with their non-producing counterparts respectively. To better account for sample variation, the fold changes were shrunken towards a beta prior to reduce effect sizes for low confidence fold change estimates and improve gene fold changes rankings[124], eliminating the need for additional filtering. To obtain genes that show the greatest activation disparity between HEK293 and CHO cells, the absolute difference of fold changes for all genes across the two cell lines were ranked, and the top 20 (0.12%) of them were chosen for further investigation. To estimate the differential activation between cell lines across r-proteins, the DESeq2 design matrix was expanded to include an interaction term between the cell lines and the r-protein identities.

To not be overshadowed by the expression of extreme genes and to pay more equal attention to all the genes within the secretory pathway, a variance-stabilizing transformation was applied[125], similar in implementation to a log-transformation on the secretory pathway gene expression.

Co-expression validation of gene outliers between HEK293 and CHO

Each pKTH16_dPur plasmid encoding gene outliers between HEK293 and CHO, or an empty pKTH16_dPur vector control, was co-transfected with the pQMCF plasmid encoding THBS4 in duplicates into 293-F and Freestyle CHO-S cells using PEI as described above. In total 1 ug of plasmid was transfected per 1 million cells and plasmid ratios of 1:1, 1:2 and/or 1:10 (gene outlier:THBS4) was used. Cells were cultivated in deep-well plates as described above and cells and supernatants were harvested at 72 hours post transfection. Some gene outliers were also validated in combination with ARTN. The expression of THBS4 in each sample was evaluated by

sandwich ELISA using a human anti-HPC4-antibody (Icosagen) as capture antibody, rabbit anti-THBS4 (antibody HPRK2400008 kindly provided from the Human Protein Atlas) as primary antibody and a HRP-conjugated swine anti-rabbit antibody (p039901-2, Dako) combined with TMB substrate (Thermo Fisher Scientific) for detection. The relative secreted expression of ARTN was determined by western blotting of culture supernatants using the rabbit antibody HPRK3140989 from the Human Protein Atlas and an HRP-conjugated swine anti-rabbit antibody (p039901-2, Dako) combined with Immobilon Western Chemiluminescent HRP Substrate (Millipore). For western blot validation of outlier gene expression, cells were lysed using M-PER or Mem-PER mammalian protein extraction reagents (Thermo Fisher Scientific) and blotted onto PVDF membranes as described above. Proteins were detected using a monoclonal anti-FLAG M2 antibody (F3165, Sigma/Merck Millipore) and an HRP-conjugated polyclonal goat anti-mouse antibody (P0447, Dako).

Pathway and protein feature analysis

Gene set enrichment analysis was used to calculate the significantly activated pathways from gene-level differential expression profiles. Canonical pathway annotation was obtained from MSigDB[126]. Additionally, manually curated gene sets on various secretory pathway subsystems were referenced from[48]. A normalized enrichment score (NES) representing the gene-set enrichment analysis (GSEA) statistic [127] was calculated to quantify the overall direction of regulation for each gene set along with an accompanying permutation p-value[128].

PTM information for each r-protein integrated from UniProt and phosphosite. Among the various types of common PTMs that occur in the cell, we considered glycosylation and disulfide bonds due to their occurrences in our panel of r-proteins and their relevance to the secretory pathway. For each r-protein, a glycosylation- and disulfide bond-index quantifying the level of enrichment of the respective PTM was calculated by dividing the total number of occurrences of the PTM in question in the r-protein by the length of the protein. The enzymes responsible for the

synthesis of glycans and disulfide bonds were obtained from[129] and KEGG[130] respectively, and their corresponding expression changes from CHO to HEK293 were summarized using the GSEA enrichment score statistic for each r-protein. A Bayesian linear regression model (formula given below) was used to assess the relationship between titer improvement from CHO to HEK293 and the enzyme expression. To further deconvolute the effects of PTMs on this dependency, an interaction term between the PTM index and the enzyme expression was added to the linear model to capture how the correlation between the enzyme activities and the titer improvements changes across r-proteins with different PTM indices. The model coefficients were estimated with Markov chain Monte Carlo (MCMC) via the rethinking package with default parameters[131].

$$\begin{aligned}
\mathrm{LFC}_{\mathrm{Titer}}[i] &\sim \mathrm{Normal}(\mu[i], \sigma) && \text{(For each clone } i\text{, draw titer improvement LFC from } \mu[i] \text{ with standard deviation } \sigma\text{)} \\
\mu[i] &= \mathrm{a} + b_{\mathrm{PTM}} \cdot \mathrm{PTM_index}[i] + b_{\mathrm{Enzyme_expr}} \cdot \mathrm{Enzyme_expr}[i] + \\
&\quad b_{\mathrm{Interaction}} \cdot \mathrm{PTM_index}[i] \cdot \mathrm{Enzyme_expr}[i] && \text{(break } \mu[i] \text{ down based on PTM, enzyme expression and an interaction term)} \\
&\textbf{Priors:} \\
a &\sim \mathrm{Normal}(0, 1) \\
b_{\mathrm{PTM}}, b_{\mathrm{Enzyme_expr}}, b_{\mathrm{Interaction}} &\sim \mathrm{Normal}(0, 0.25) \\
\sigma &\sim \mathrm{Exponential}(1)
\end{aligned}$$

Statistical Analysis

Statistical analysis of significantly different expression levels of THBS4 and ARTN from co-expression experiments was performed in GraphPad Prism 7 using ANOVA one-way analysis followed by Durnetts's test comparing all expression levels to a control (THBS4 or ARTN co-expressed with an empty plasmid). The details of statistical analysis of the transcriptomic profiling, including differential expression analysis and protein feature analysis is described in the respective sections above.

Data and materials availability

Raw RNA sequencing data are deposited to the Sequence Read Archive (SRA) with accession number SRP281874, and the corresponding processed data are available on the Gene Expression Omnibus (GEO) via accession number GSE157729. All the other data are contained

in the article and its supplementary information or available upon request. The source code for the figures are available from https://github.com/LewisLabUCSD/CHO_HEK.

Chapter 2, in part, is currently being prepared for submission for publication of the material as it may appear in Malm M, Kuo CC, Barzadd MM, Mebrahtu A, Wistbacka N, Razavi R, Volk AL, Lundqvist M, Kotol D, Edfors F, Gräslund T, Chotteau V, Field R, Varley PG, Roth RG, Lewis NE, Hatton D, Rockberg J. “Harnessing secretory pathway differences between HEK293 and CHO to rescue production of difficult to express proteins” (Submitted). The dissertation author was the one of the two primary investigators and authors of this material.

CHAPTER 3: TRANSCRIPTOMIC ANALYSIS OF RECOMBINANT PROTEIN-PRODUCING CHO CELLS REVEAL PRODUCT-DEPENDENT HOST RESPONSE

Introduction

The human secretome project [70,132] has comprehensively characterized an important subset of the human proteome that is produced through the secretory pathway. Secreted proteins often play important roles in cell signaling [133], and those that are present in blood can extend their influence beyond the confines of the human tissues [134]. This makes secreted proteins appealing recombinant therapeutics candidates for the biopharmaceutical industry. Chinese hamster ovary (CHO) cell line has been the most popular mammalian expression system for recombinant proteins because of its scalability and compliance with human post-translational modifications (PTMs) [135,136]. To systematically measure the potential of CHO cells to produce the human secretome, an effort to express the entire human secretome recombinantly in CHO cells was initiated as a companion project to the human secretome project. So far, almost 1,300 proteins have successfully been produced and purified in CHO cells in a controlled, high throughput fashion [137]. However, despite optimization efforts, only 65% of the secretome could be successfully expressed by CHO cells above the quality threshold. Furthermore, among the proteins that passed quality checks, titers differed by several orders of magnitude depending on the protein. To better understand the determinants of protein titers, we RNA-sequenced a panel CHO cell clones each producing one of the 95 different recombinant proteins, along with one non-producing clone. We found that transgenes consistently take up roughly 3% of the entire transcriptome, making it one of the most highly expressed genes in most of the producer clones. However, less than 5% of the variance can be attributed to differences in transgene abundance. To further identify determinants of recombinant protein yields, we focused on properties of the recombinant proteins and the conditions of the host cells.

A machine learning approach was used to unravel the factors underlying productivity based on features of the recombinant proteins. More specifically, we curated a comprehensive feature set according to the (amino acid) sequences of the secreted proteins. The features range from basic properties of the protein such as molecular weight to complex structural features derived from homology modeling. We developed a robust machine learning pipeline to perform automatic feature selection and identified a set of sequence features with the strongest influence on protein titers.

To gain further insights into the In addition, expression data is presented for all the gene constructs in the CHO cell factory to allow in-depth analysis of the relationship between protein sequence and yield. For some representative CHO expression clones, including both high and low producers, omics data have been generated and data is also presented using a human cell line HEK293 for a selection of the clones with low or no production in CHO. In conclusion, we will provide an open access knowledge resource to facilitate basic and applied research covering the proteins actively secreted in human cells, tissues and organs.

Results

Variation in achieved recombinant protein yield cannot be explained by transgene mRNA abundance

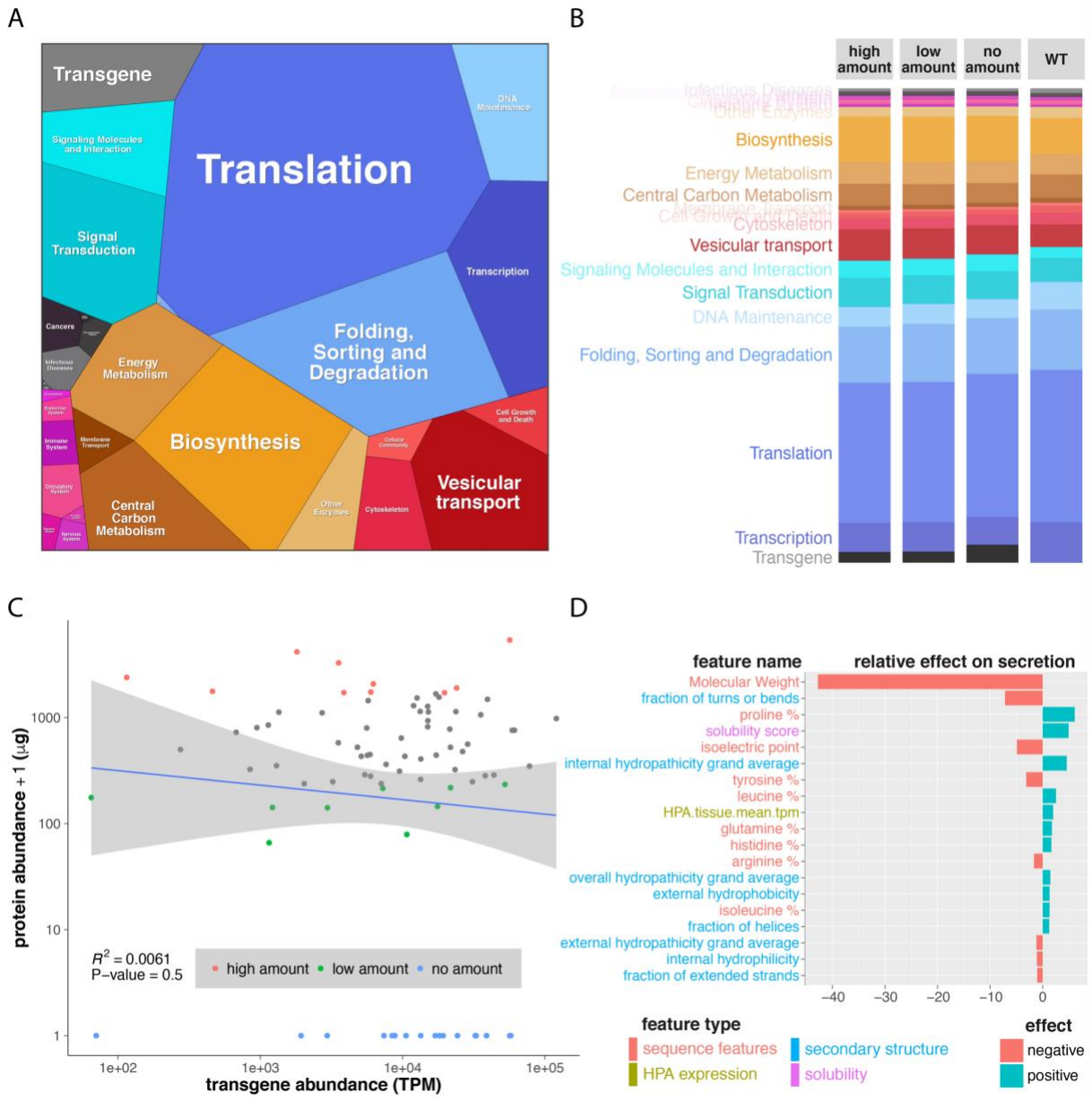


Figure 9. The contribution of transgene mRNA level, host cell transcriptome, and protein structural properties on the yield of recombinant proteins in CHO cells.
CHO cells expressing the entire human secretome were developed, and RNA-Seq was conducted on 96 representative clones, each secreting a different human protein. (A) The cells expressed a wide range of transcripts supporting a wide range of cellular processes. The cells dedicated a substantial amount of resources on mRNAs associated with translation and protein synthesis (blue). (B) Overall, the transcriptome usage is mostly consistent across cell lines showing different levels of recombinant protein production. (C) The transgene abundance (TPM) of each recombinant gene was compared to the secreted protein abundance (µg). (D) To quantify the impact of secreted protein characteristics and their effects on secretion, a machine learning pipeline was used to identify features of the secreted proteins that contribute the most (both positively and negatively) to protein secretion.

To better characterize the library of CHO cells producing the human secretome, we selected for RNA-sequencing 95 clones of CHO cells producing various recombinant proteins based on the yield, growth characteristics and overall viability. An additional non-producing clone was also RNA-sequenced (methods). The mRNA abundance for the transgenes encoding the human secreted proteins, along with the endogenous CHO genes are summarized in Suppl. Table 4.

We visualized the global transcriptome usage of the CHO cells using a Proteomap [121] (Fig. 9A), wherein mRNA abundance is summarized into blocks of cellular processes with the area representing the fraction of the transcriptome dedicated to a particular cellular process. On average, the transgenes are among the most highly expressed genes in the CHO cell transcriptome, taking up ~3% of the transcriptome and the levels of mRNA were comparable among cell lines, regardless of whether the cell lines secreted high, low, or no amounts of recombinant protein (Fig. 3B). Additionally, the major cellular processes remain relatively stable across different tiers of producers and interestingly, the expression level of genes associated with translation show slightly higher expression in cell lines with the lowest recombinant protein yield. Incidentally, cellular processes associated with vesicular transport were down-regulated in non-producers.

We then evaluated if the variation in protein production yields can be explained by transgene mRNA levels. Across the producers, the transgene mRNA levels only explained 5% (Fig. 9B) of the variance in recombinant protein productivity, compared to previous reports of ~40% for endogenous genes in mammalian cells across various conditions [121,138,139]. Incidentally, cell lines that fail to secrete detectable levels of recombinant protein tend to express higher levels of transgene mRNAs compared to clones of higher yields. This suggests that difficult-to-express proteins are less limited by transcript level, compared to other factors.

We further examined how other transcriptomic determinants in the host cells impacted the amount of protein achieved. To evaluate if any particular pathway differentiated between the cell lines that gave high vs. low amounts of recombinant protein products, we conducted Gene Set Enrichment Analysis (GSEA) [127,140] and found several dysregulated cellular responses. This includes genes targeted by KRAS activation, whose expression was significantly down-regulated in high producers. As KRAS contributes to the Warburg effect [141], the reduction in KRAS signaling in high producers may improve cell growth, contributing to higher recombinant protein yields.

Difficult-to-express recombinant proteins are characterized by higher molecular weight and more frequent turns

While mRNA levels of the transgenes are largely uninformative about recombinant protein titers, we wonder if attributes of each secreted protein could explain protein yields. We curated a compendium of various features characteristic of each SecP in our recombinant proteomics dataset. These range from simple features such as molecular weights to more complex signatures including the structural properties of the protein. We then used machine learning to model how the features of the proteins affect the protein yield (Fig. 9D). Note that the transgene mRNA level was excluded from this analysis to focus on structural features of the recombinant proteins. Among more than 150 features curated for the secreted proteins, the molecular weight of the proteins impacts achieved recombinant protein yields the most (Suppl. Table 4). Large proteins pose additional demands, as they have more post-translational modifications, provide more opportunities for misfolding and aggregation, more difficulty of trafficking, and impose increased energy consumption in their production. Recombinant protein titers are also impacted considerably by other features, such as the fraction of the protein sequence to be in structural turns or bends, the amount of proline, hydrophobicity, etc. Together, protein-specific features explain 45% of the variance in protein secretion and these attributes can help inform recombinant proteins candidate selection for future production runs.

Differential expression underlies titer differences between similarly structured recombinant proteins

The aforementioned protein-specific features highlight the properties common to difficult-to-express proteins. However, they generally do not serve as actionable engineering targets for titer enhancement as they are integral to the identity and function of the recombinant protein. To boost recombinant protein productivity, efforts have initially focused on bioprocess optimization before targeted genetic manipulation became more widely explored in the last decade as the available genomic resources continue to grow. When it comes to enhancing specific productivity, components within the secretory pathway have traditionally been popular engineering targets because they are directly involved in various stages of protein secretion. However, the same engineering target can result in improvement in productivity that varies greatly depending on the product of interest, suggesting a dependency between the secretory pathway and its products. Such product-specific relationships in the secretory pathway have been more thoroughly examined recently [48,142], and studies have demonstrated significant titer improvements when engineering targets are selected based on the properties of the secreted protein [143].

To identify engineering opportunities within CHO cells, we proposed to analyze functional differential expression while accounting for differences in intrinsic protein properties. We first selected various pairs of secreted proteins with similar structural and functional properties based on their T-SNE embeddings and prioritized clonal pairs with drastically different protein titers. The interactive web app we developed to facilitate clonal pair selection is available on https://curtis999.shinyapps.io/candidate/.

Methods

Sequence processing and RNA-seq quantification

Sequence data for RNA-Seq were quality controlled using FastQC and summarized with multiQC [117]. Trimmomatic [118] was used to trim low-quality bases from the reads. The CHO-K1 reference genome [119] was extended to incorporate the transgene sequences so that the

transcripts of the heterologous secretome can be quantified. Reads were then quasi-mapped to the extended CHO-K1 genome and quantified with Salmon [120].

Proteomaps

The Proteomaps provides a means to visualize genes and the corresponding cellular processes for mouse and other model organisms. Thus, an ortholog conversion table [38] was used to convert the CHO-K1 genes to their mouse homologs, wherein a two-way BLAST, protein sequences extracted from Refseq 68 [144] and a gene name were used in conjunction to maximize coverage of overlapping CHO and mouse genes. The resulting Proteomaps cover roughly 70% of the transcripts as Proteomaps does not include annotation for all mouse genes, nor do all CHO genes have mouse homologs. Nevertheless, the coverage is significant enough that most major cellular processes are represented.

Transcriptomic determinants of protein secretion

The differential expression between subsets of samples that are high (n = 15) and low producers (n = 11) was performed with DESeq2 [122]. GSEA [127] was then used to determine the significantly up- and down-regulated cellular processes.

Protein features importance

By regressing the actual protein yields on key properties of the respective secreted proteins, our machine learning pipeline can identify protein-specific determinants of the actual protein yield. To fully characterize the properties of the secretome, we extended upon the features from our pilot study [145] on the expression determinants of the human protein fragments used in the creation of the antibodies for the HPA project. We compiled a compendium of various features characteristic of each SecP. The features, ranging from simple features such as molecular weights to more high-level signatures such as the structural properties of the protein, fall into four broad categories (Fig. 10).

Feature type	Feature name	description	Number of features	Data source/ software packages
sequence features	*AA.comp*	Amino acid composition	20	this study
	AA.comp	Amino acid composition correlation with AA composition in native CHO cells	2	this study
	Mol.Weight	Molecular weight	1	this study
	PSIM	Number of post-translational modification sites	6	10.1371/journal.pone.0063284
	PTM.detailed	Number of unconventional post-translational modification sites	12	10.1093/nar/gkl124
	RNA_MFEs_exp_norm	RNA minimum free energy normalized wrt sequence length	1	10.1186/1748-7188-6-26
Experimental abundance	*GTEX_human_grouped*	secretome expression in various human tissues	16	10.1126/science.1262110
	HPA_protein	Protein level across different tissues	17	10.1126/science.1260419
	HPA.tissue.mean	Mean protein level across different tissues	1	
	Matt	Protein and mRNA copy numbers, half-lives, transcription rates and translation rate constants in mouse fibroblasts	6	10.1038/nature10098
	PrESTs.df	Production yield of fusion proteins with fractions of human secretome in yeast	1	10.1093/bioinformatics/btx207
	RNA96_WT	Expression of the CHO ortholog of the product protein	1	this study
physical_features	*Protparam_gravy*	Grand average of hydropathicity (GRAVY)	1	ProtParam
	Protparam_instability	Instability index, an estimate of the stability of the protein in a test tube.	1	
	Protparam_isoelectric_point	protein isoelectric point	1	
Structural predictions	*acc_frac*	Solvent-accessible fraction	1	10.1093/nar/gki396
	acc_hydrophilic_in	% Hydrophilic solvent-inaccessible residues	1	
	acc_hydrophilic_out	% Hydrophilic solvent-accessible residues	1	
	acc_hydrophobic_in	% Hydrophobic solvent-inaccessible residues	1	
	acc_hydrophobic_out	% Hydrophobic solvent-accessible residues	1	
	acc20_mean	Mean accessibility score	1	
	out_gravy	GRAVY of outer and inner residues	1	
	ss_helix, ss_ext, ss_c	3-category Secondary Structures	3	
	ss8_helix, ss8_ext, ss8_turn, ss8_helix3, ss8_pi_helix, ss8_bridge, ss8_bend, ss8_coil	8-category Secondary Structures	8	
	solubility	Predicted protein solubility	2	10.1093/bioinformatics/btx345

Figure 10. List of features used in the expression data analysis and their sources.

The features were then filtered univariately in which features with low maximal information coefficients [146] are removed. The remaining features are then passed to 8 different machine learning algorithms: random forest, Gradient Boosting Machine, partial least square, glmnet.fit, cubist, glmboost.fit, glmnet_significancePenalty.fit and SVM. For models that require hyper-parameter fine tuning, a grid-search was used to obtain the optimal parameters over the 15 bootstrap resamples. The performance of each model is assessed through 0.632bootstrap [147] to better capture the prediction errors typically underestimated by vanilla bootstrapping. Models that achieved an R-squared better than 0.65 on the validation set yet are diverse in terms of prediction

errors are then selected to create an ensemble model, where the models are weighted by simple linear regression. As the linear ensemble model outperformed the single models, it was used to determine the factors contributing to poor secretion by averaging the vectors of feature importance generated by each model according to their corresponding weights in the ensemble model.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Kuo CC, Masson HO, Lewis NE. “Transcriptomic analysis of recombinant protein-producing CHO cells reveal product-dependent host response” The dissertation author was the co-first investigator and author of this material.

CHAPTER 4: IN SITU DETECTION OF PROTEIN INTERACTIONS

Abstract

Despite their therapeutic potential, many protein drugs remain inaccessible to patients since they are difficult to secrete. Each recombinant protein has unique physicochemical properties and requires different machinery for proper folding, assembly, and post-translational modifications (PTMs). Here we aimed to identify the machinery supporting recombinant protein secretion by measuring the protein-protein interaction (PPI) networks of four different recombinant proteins (SERPINA1, SERPINC1, SERPING1 and SeAP) with various PTMs and structural motifs using the proximity-dependent biotin identification (BioID) method. We identified PPIs associated with specific features of the secreted proteins using a Bayesian statistical model, and found proteins involved in protein folding, disulfide bond formation and N-glycosylation were positively correlated with the corresponding features of the four model proteins. Among others, oxidative folding enzymes showed the strongest association with disulfide bond formation, supporting their critical roles in proper folding and maintaining the ER stability. Knockdown of disulfide-isomerase PDIA4, a measured interactor with significance for SERPINC1 but not SERPINA1, led to the decreased secretion of SERPINC1, which relies on its extensive disulfide bonds, compared to SERPINA1, which has comparatively less disulfide bonds. Proximity-dependent labeling successfully identified the transient interactions supporting synthesis of secreted recombinant proteins and refined our understanding of key molecular mechanisms of the secretory pathway during recombinant protein production.

Introduction

Therapeutic proteins are increasingly important for treating diverse diseases, including cancers, autoimmunity/inflammation, infectious diseases, and genetic disorders. For example, the plasma protein therapeutics market is expected to grow by $36 billion (USD) by 2024. Mammalian

cells are the dominant production system due to their ability to perform PTMs that are required for drug safety and function [148,149]. However, the complexities associated with the mammalian secretory machinery remains a bottleneck in recombinant protein production [97]. The secretory pathway machinery includes >575 gene products tasked with the synthesis, folding, PTMs, quality control, and trafficking of secreted proteins (SecPs) [47,48,150,151]. Numerous components of the secretory pathway (SecMs) have been engineered to increase the capacity of the secretion. However, the precision and efficiency of the mammalian secretory pathway results from the coordinated effort of these secretory machinery components including chaperones, modifying enzymes (e.g., protein disulfide isomerases and glycosyltransferases), and transporters within the secretory pathway. Overexpression of heterologous proteins in this tightly regulated and complex system could impact its functionality and homeostasis, resulting in adaptive responses that can impair both protein quantity and quality [98,152]. More importantly, variability in the structures and modifications of recombinant proteins could necessitate a customized secretion machinery to handle this diversity, but the secretory machinery of recombinant protein producing cells has not been adapted to facilitate the high titer secretion desired for most recombinant proteins. A previous study also showed human protein secretory pathway genes are expressed in a tissue-specific pattern to support the diversity of secreted proteins and their modifications [48], suggesting that expression of several SecMs is regulated to support client SecPs in the secretory pathway. Unfortunately, the SecMs needed to support any given secreted protein remain unknown. Thus, there is a need to elucidate the SecMs that support the expression of different recombinant proteins with specific features. This can guide mammalian synthetic biology efforts to engineer enhanced cells capable of expressing proteins of different kinds in a client-specific manner.

PPI networks are invaluable tools for deciphering the molecular basis of biological processes. New proximity dependent labeling methods such as BioID [153,154] and APEX [155] can identify weak and transient interactions in living cells, along with stable interactions. Furthermore, BioID offers a high-throughput approach for systematic detection of intracellular PPIs occurring in

various cellular compartments and has been used to characterize PPI networks and subcellular organization [156]. BioID relies on expressing a protein of interest fused to a promiscuous biotin ligase (BirA) that can biotinylate the proximal interactors in nanometer-scale labeling radius [154]. For example, this approach has mapped protein interactions at human centrosomes and cilia [157,158], focal adhesions [159], nuclear pore [154] and ER membrane-bound ribosomes [160] . Here we used BioID2, an improved smaller biotin ligase for BioID [156,161], to explore how the SecMs involved vary for different secreted therapeutic proteins (Fig. 11). Specifically, BioID2 was employed to identify SecMs that interact with three SERPIN-family proteins (SERPINA1: treatment for Alpha-1-antitrypsin deficiency, SERPINC1: treatment for Hereditary antithrombin deficiency, and SERPING1: treatment for acute attacks of hereditary angioedema) and secreted embryonic alkaline phosphatase (SeAP), which is a truncated form of Alkaline Phosphatase, Placental Type (ALPP). These proteins vary in their PTMs (e.g., glycosylation, disulfide bond and residue modifications) and have different amino acid sequences that consequently form different local motifs. Using a Bayesian statistical model, we identified the critical PPIs that are positively correlated with each protein feature. Identification of these PPIs will refine our understanding of how the secretory pathway functions during the expression of the recombinant proteins and introduce novel targets for secretory pathway engineering in a client specific manner.

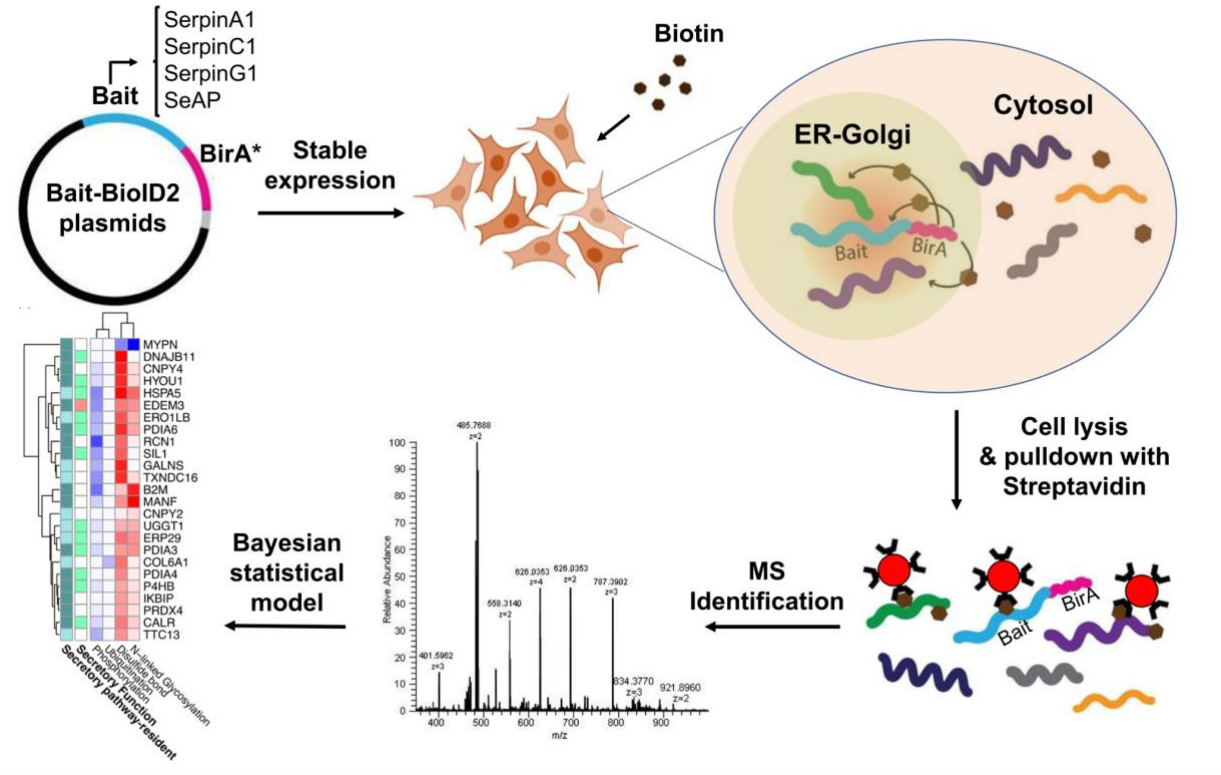


Figure 11. Flowchart of the BioID2 application to detect in situ interactions supporting therapeutic proteins secretion.

Materials and Methods

Molecular cloning and generation of stable cell lines

All plasmids used in this study were constructed by PCR and Gibson isothermal assembly. The expression ORFs, hereafter named bait-BirA, were constructed by fusing BioID2 to the C-terminal of each model protein (with a glycine-serine linker added between) and a 3XFLAG tag at C-terminal to simplify the immuno detection. ORFs were inserted into pcDNA5/FRT (Invitrogen), which allows targeted integration of the transgenes into the host genome. Gibson assembly primers were designed by SnapGene software and used to amplify the corresponding fragments and vectors with long overlapping overhangs, which were then assembled using Gibson Assembly Master Mix (NEB). To obtain secretable BioID2 (without any bait protein), Gibson

assembly was employed to fuse the signal peptide of SERPINC1 gene to the N-terminal of BirA (hereafter referred to as Signal-BirA). Assembly products were transformed to the chemically competent E. coli, and recombinant plasmids were verified by restriction digestion and sequencing. For all experiments, Flp-In 293 cells (Invitrogen) were cultured in DMEM media supplemented with fetal bovine serum (10 %) and antibiotics (penicillin, 100 U mL−1 and streptomycin, 100 µg mL−1) and maintained at 37 °C under 5 % CO2. For generating stable cell lines, Flp-In 293 cells were seeded in 6 well plates at a density of 0.5×10^6 cells per well the day before transfection. Cells were then co-transfected with each pcDNA5/FRT vector containing expression cassette and pOG44 plasmid using Lipofectamine® 2000 according to the manufacturer's directions. After recovery from transfection, cells were grown in DMEM containing 10% FBS, 1% PenStrep, and 150 µg/mL Hygromycin B to select hygromycin-resistant cells. Individual resistant colonies were isolated, pooled, and seeded in 24-well plates for further scaling up and screened for expression of the fusion proteins by Western Blotting.

Immunofluorescence

Recombinant HEK293 cells expressing BioID2 fusions were grown in complete medium supplemented with 50 uM biotin on coverslips until 70% confluent. Cells were then fixed in PBS containing 4% PFA for 10 min at room temperature. Blocking was performed by incubating fixed cells with 1% BSA and 5% normal goat serum in PBST. Anti-flag mouse monoclonal antibody-Dylight 650 conjugate (Thermofisher), targeting the bait-BirA, and streptavidin-DyLight 594 conjugate (Thermofisher), targeting the biotinylated proteins, were diluted at 1:300 and 1:1000 in blocking buffer, respectively and incubated with fixed cells for 30 minutes at room temperature. Cells were then washed, counterstained with DAPI, mounted on the slide using antifade vectashield mountant, and imaged using Leica SP8 Confocal with Lightning Deconvolution. Colocalization quantification was performed for the deconvolved images using Fiji's (ImageJ 1.52p) Coloc_2 analysis tool between the 650 (anti-flag) and 594 (Streptavidin) channels [162]. This

tool generates a comprehensive report for evaluating pixel intensity colocalization of two channels by various methods such as Pearson's Coefficient (range: -1.0 to 1.0), Manders' Colocalization Coefficients (MCC, range: 0 to 1.0), and Li's Intensity Correlation Quotient (IQC, range: -0.5 to 0.5) [163,164]. Background pixel intensity was subtracted using Fiji's rolling ball algorithm and a region of interest (ROI). Thresholds were determined using Coloc_2's bisection method, which is further used to adjust for background. Above threshold metrics were reported.

RNAi knockdown experiment

esiRNA targeting PDIA4, PDIA6, and ERp44 were ordered from Sigma. HEK293 cell expressing SERPINC1-BirA and SERPINA1-BirA were seeded at 1X105 cells/well in 12-well plates with complete medium and reverse transfected with 144 ng of the appropriate PDI specific esiRNA, Luciferase esiRNA as a negative control, or KIF11 esiRNA as positive control using Lipofectamine RNAiMAX (Invitrogen). All transfections were performed according to the manufacturer's guidelines. Each experiment was done in triplicates and targeted gene knockdown by esiRNA was allowed to occur for 96 hrs. Culture supernatants and cell pellets were then harvested, clarified by low-speed centrifugation, and then aliquoted and stored at −80°C for further experiments.

Western blotting

To validate the secretion of bait-BirA proteins, supernatants of cultures expressing fusion proteins were collected, centrifuged to remove cell debris, and 30 ul were loaded on SDS-PAGE gel for electrophoresis. The resolved proteins were then transblotted to nitrocellulose membranes using the Trans-Blot Turbo Transfer System from Bio-RAD. The membrane was blocked with 5% skim milk in TBST and probed with HRP-conjugated anti-flag mouse monoclonal antibody (Thermofisher) diluted at 1:10000 in the blocking buffer. The membrane was washed, and Clarity Western ECL Substrate was added. Proteins' bonds were visualized using G:Box Gel Image Analysis Systems (SYNGENE). For staining of intracellular biotinylated proteins, cells were grown

in complete medium supplemented with 50 µM biotin, lysed by RIPA buffer, and protein content was quantified using Bradford assay. 20 ug of total protein was loaded, resolved and transblotted as described earlier. The membrane was blocked by 3% BSA in TBST and probed with HRP-conjugated streptavidin diluted in blocking buffer at 1:2000 ratio. For visualizing the proteins' bands, the same Clarity Western ECL Substrate was used.

Quantitative western-blots were used to determine knockdown (KD) efficiency as well as impact on SERPIN secretion. For KD efficiency, cell pellets from each KD experiment were lysed with RIPA buffer, and approximately 25µg of protein lysate was loaded onto the SDS-PAGE gel for PDIA4, PDIA6, and ERp44 experiments. Resolved proteins were transblotted onto separate nitrocellulose membranes as described earlier. Each membrane was blocked with Intercept (TBS) Blocking Buffer from LI-COR and probed with Rabbit anti PDIA4 (1:2000), PDIA6 (1:2000), or ERp44 (1:2500) monoclonal antibodies, respectively. A housekeeping protein in each lysate was targeted for normalization with either a Mouse anti Alpha-tubulin (1:20,000) or anti Beta-actin (1:10,000) monoclonal antibodies. For visualization, IR spectra was utilized by staining with LI-COR Goat anti Mouse conjugated to 680 (1:15,000) and anti Rabbit conjugated to 800 (1:15,000). Images were taken and analyzed with Image Studio Lite Version 5.2 for relative quantification. Impact on SERPIN secretion was determined from the aliquots of clarified supernatants from esiRNA transfected cultures using the same blotting method as above. 30 ul of supernatants were loaded on SDS-PAGE gel, and transblotted onto nitrocellulose membrane. Transblotted membrane was probed using Rabbit anti Flag (Proteintech, 1:800) followed by Goat anti Rabbit 800 (1:15,000). KD efficiency and the effect of PDI knockdown on secretion of the model proteins was measured in comparison with the negative control of each cell line transfected with Luciferase esiRNA as negative control (see above).

Mass Spectrometry

Cells were grown in 245 mm plates (one plate per biological replicate in triplicate) to approximately 70% confluence in complete media and then incubated for 24 h with 50 µM biotin.

Cells were harvested and washed twice in cold PBS, lysed with vigorous shaking (20 Hz) in 8M urea, 50mM ammonium bicarbonate lysis buffer, extracted proteins were centrifuged at 14,000 x g to remove cellular debris and quantified by BCA assay (Thermo Scientific) as per manufacturer recommendations. Affinity purification of biotinylated proteins was carried out in a Bravo AssayMap platform (Agilent) using AssayMap streptavidin cartridges (Agilent), and the bound proteins were subjected to on-cartridge digestion with mass spec grade Trypsin/Lys-C Rapid digestion enzyme (Promega, Madison, WI) at 70ºC for 2h. Digested peptides were then desalted in the Bravo platform using AssayMap C18 cartridges and the organic solvent was removed in a SpeedVac concentrator prior to LC-MS/MS analysis. Dried peptides were reconstituted with 2% acetonitrile, 0.1% formic acid, and analyzed by LC-MS/MS using a Proxeon EASY nanoLC system (Thermo Fisher Scientific) coupled to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific). Peptides were separated using an analytical C18 Acclaim PepMap column 0.075 x 500 mm, 2µm particles (Thermo Scientific) in a 93-min linear gradient of 2-28% solvent B (80% acetonitrile, 0.1% formic acid) at a flow rate of 300nL/min. The mass spectrometer was operated in positive data-dependent acquisition mode. MS1 spectra were measured with a resolution of 70,000, an AGC target of 1e6 and a mass range from 350 to 1700 m/z. Up to 12 MS2 spectra per duty cycle were triggered, fragmented by HCD, and acquired with a resolution of 17,500 and an AGC target of 5e4, an isolation window of 1.6 m/z and a normalized collision energy of 25. Dynamic exclusion was enabled with a duration of 20 sec.

MS data Analysis

All mass spectra were analyzed with MaxQuant software [165] version 1.5.5.1. MS/MS spectra were searched against the Homo sapiens Uniprot protein sequence database (version January 2018) and GPM cRAP sequences (commonly known protein contaminants). Precursor mass tolerance was set to 20ppm and 4.5ppm for the first search where initial mass recalibration was completed and for the main search, respectively. Product ions were searched with a mass tolerance 0.5 Da. The maximum precursor ion charge state used for searching was 7.

Carbamidomethylation of cysteines was searched as a fixed modification, while oxidation of methionines and acetylation of protein N-terminal were searched as variable modifications. Enzyme was set to trypsin in a specific mode and a maximum of two missed cleavages was allowed for searching. The target-decoy-based false discovery rate (FDR) filter for spectrum and protein identification was set to 1%. Enrichment of proteins in streptavidin affinity purifications were calculated as the ratio of intensity. To remove the systematic biases introduced during various steps of sample processing and data generation, datasets were normalized using the LOESS method [166] integrated into Normalyzer [167]. Perseus software [168] was employed for data preparation, filtering, and computation of differential protein abundance. The DEP package [169] was used to explore whether missing values in the dataset are biased to lower intense proteins. Left-censored imputation was performed using random draws from shifted distribution. A Student's t-test with a multi-sample permutation-based correction for an FDR of 0.05 was employed to identify differentially expressed proteins using log2 transformed data.

Detection of significant interactions

The threshold for significant interactions was determined using the known secretory pathway components as a gold standard. We set the cutoffs for FDR at 0.1 and removed all interactors with negative fold changes, as this optimizes the enrichment of known secretory pathway components among the significant interactors. The enrichment for two independent secretory pathway-related gene sets also peaked around the cutoffs set through the gene set of known secretory pathway components, suggesting the optimal cutoffs are robust to the gold standards chosen.

Estimation of preferential interaction between protein features and interactors with a Bayesian modeling framework

To identify patterns of interactions between individual bait-BirA proteins and their interactors, we first obtained and summarized several protein features across model proteins. The protein properties considered include shared structural motifs [170], known sites of PTM from

Uniprot and phosphosite [170,171]. The complete protein feature composition for each of the model proteins is illustrated in. Given the interactions between the SecPs and their interactors, we can predict the important structural features implicated in the interactions between the SecPs and a given SecM. The interactions between the BirA-fused samples and the secretory pathway interactors can be pooled according to shared properties of the SecPs to reveal interdependencies between components of the secretory pathway and their products.

To test if some secretory machinery components preferentially interact with certain protein features, we first calculated the effective total frequency ($\delta_{f,g}$) of interactions between each feature-gene pair (f,g) by going through every SecP in our data and counting the number of times this feature occurs in a bait-BirA protein p (f_p).

$$\delta_{f,g} = \sum_{p \in \text{secPs}} \mathbf{f}_p \cdot \mathbb{I}_{\mathbf{interact}}(p,g) \qquad \text{(bait-BirA protein: } p\text{; interactor: } g\text{)}$$

$$\mathbb{I}_{\mathbf{interact}}(p,g) := \begin{cases} 1 & \text{if p and g interact} \\ 0 & \text{otherwise.} \end{cases}$$

The number f_p is added to the total frequency, $\delta_{f,g}$, only when p and g interact (when $\mathbb{I}_{interact}(p,g) = 1$). This effective interaction frequency, $\delta_{f,g}$ essentially linked the features pooled across the model proteins to a given interactor g, taking us closer to estimating interaction affinity between a feature and an interactor. To further account for SecP and feature promiscuity, we implement an estimate of the tendency for f and g to interact. The interaction synergy $s_{f,g}$, the tendency for an interactor g to interact with a feature more than expected by chance, was estimated by a Bayesian modeling approach. More specifically, we seek to decouple the interaction synergy from the observed interaction frequency $\delta_{f,g}$ by modeling the interaction frequency $\delta_{f,g}$ with a Poisson regression.

$$\delta_{f,g} \sim \text{Poisson}(\lambda_{f,g}) \quad \text{(Interaction frequency between } f, g \text{ follows a Poisson distribution with mean } \lambda_{f,g})$$
$$\log \lambda_{f,g} = \alpha + \mu_f + \mu_g + s_{fg}$$

Interaction effects:

$$\alpha \sim \text{Normal}(0, 1) \quad \text{(Average interaction frequency across all } f, g \text{ pairs)}$$
$$\mu_f \sim \text{Normal}(0, \sigma_f) \quad \text{(Feature promiscuity for } f\text{)}$$
$$\mu_g \sim \text{Normal}(0, \sigma_g) \quad \text{(Interactor promiscuity for } g\text{)}$$
$$s_{fg} \sim \text{Normal}(0, \sigma_s) \quad \text{(Interaction synergy between } f, g\text{)}$$

Hyper priors:

$$\sigma_f \sim \text{Exponential}(1)$$
$$\sigma_g \sim \text{Exponential}(1)$$
$$\sigma_s \sim \text{Exponential}(1)$$

The mean for the Poisson distribution is parameterized by the sum of feature promiscuity (number of SecMs g connected to a feature) μ_f , interactor promiscuity (number of SecPs with a feature f interacting with a SecM) μ_g , the interaction synergy $s_{f,g}$ and an intercept variable. The feature promiscuity μ_f quantifies the probability of a feature f to partake in an interaction with any secMs. Likewise, the interactor promiscuity μ_g measures the tendency for a secM g to interact with any features. After subtracting out the μ_f and μ_g from the log-transformed mean $\lambda_{f,g}$ for the Poisson distribution, we arrive at the coefficient of interest-- the interaction synergy $s_{f,g}$ between f and g. It quantifies the degree to which f and g interact more than by random chance. In previous work, this approach has correctly estimated epistasis intensity[66]. To better regularize the promiscuities, their Bayesian priors are all normally distributed around 0, with their variances parameterized by the hyper priors σ_f , σ_g and σ_s which follow an exponential distribution. The intercept α is parameterized by a standard normal distribution. We used the rethinking R package [172] to construct the model and sample the coefficients.

Results

BioID can successfully tag proteins colocalized with secreted proteins

We first investigated if intracellular PPIs between each SecP and their supporting SecMs can be measured using the BioID method. To do this, each bait-BirA was expressed in HEK293 cells using the Flp-In™ system (see materials and methods) for targeted integration of the transgenes into the same genomic locus to ensure comparable transcription rates of each

transgene. Variations in mRNA level caused by random integration can trigger adaptive response such as the unfolded protein response in some cell lines which reciprocally alters the active PPIs network involved in the secretion. We observed successful secretion of bait-BirA proteins into culture supernatant, evaluated by Western blot (Fig. 12A). Thus, the BirA fusion did not prevent secretion of the model proteins, and it is expected that they enter the secretory pathway where they are processed and packaged for secretion. We also verified the biotinylation profile by western blot for each cell line in the presence and absence of biotin. The biotinylation profile of the bait-BirA cells is different when biotin is added to the culture with substantial increased biotinylation of specific proteins, while no obvious change is observed for WT (Fig. 12B), suggesting that BioID2 successfully tagged specific proteins within the cells. Colocalization of the bait-BirA proteins and the biotinylated proteins was then studied by multicolor co-immunofluorescence microscopy to test whether biotinylated proteins are actual partners of the model proteins. The results demonstrated successful labeling of the interactors by BirA through colocalization of the biotin-labeled proteins and bait-BirA, while WT did not show increased biotinylation under the same experimental conditions (Fig. 13). To quantify the colocalizations, we calculated different colocalization metrics (see methods) from the images and compared to the WT, and the results confirmed the specificity of the BioID labeling system to tag the proximal proteins (Li's ICQ value closer to 0.5 and Pearson's R value and Manders' Colocalization Coefficients closer to one demonstrate a dependent protein staining pattern between the red and green channels).

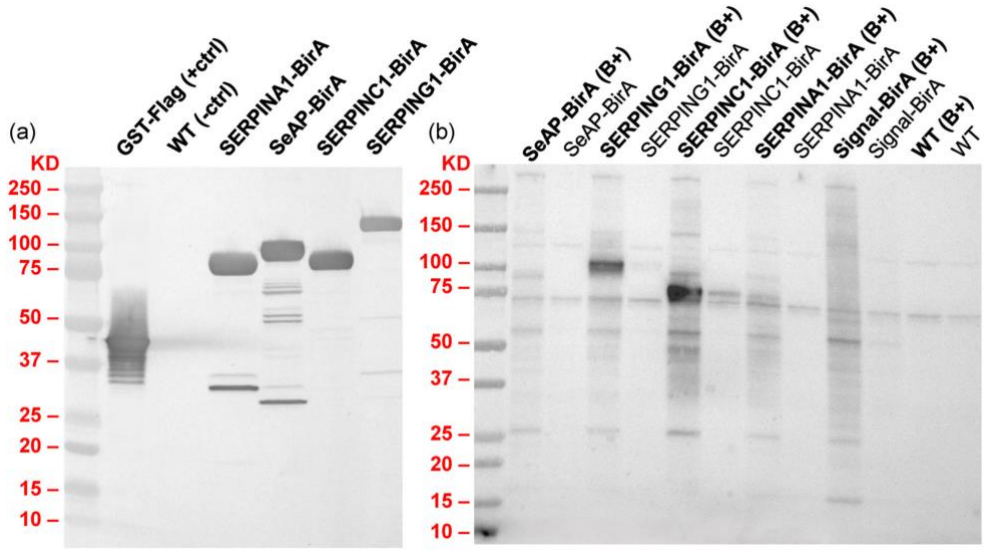


Figure 12. Expression of bait-BirA proteins results in a substantial increase in biotinylated proteins.

(a) Successful secretion of the bait-BirA proteins into the culture supernatant was evaluated by Western blot using HRP-anti-flag antibody.

(b) The immunoblotting biotinylation profiling of the model proteins and WT control in HEK293 cells with HRP-streptavidin. When the BirA domain was fused to the model proteins, biotin addition led to the biotinylation of a subset of proteins (B+) which are not seen in WT or absence of biotin. This demonstrates that the BioID labeling system tags interactions as secreted proteins are synthesized and trafficked through the secretory pathway. A few endogenously biotinylated proteins appear in the absence of biotin and in the WT.

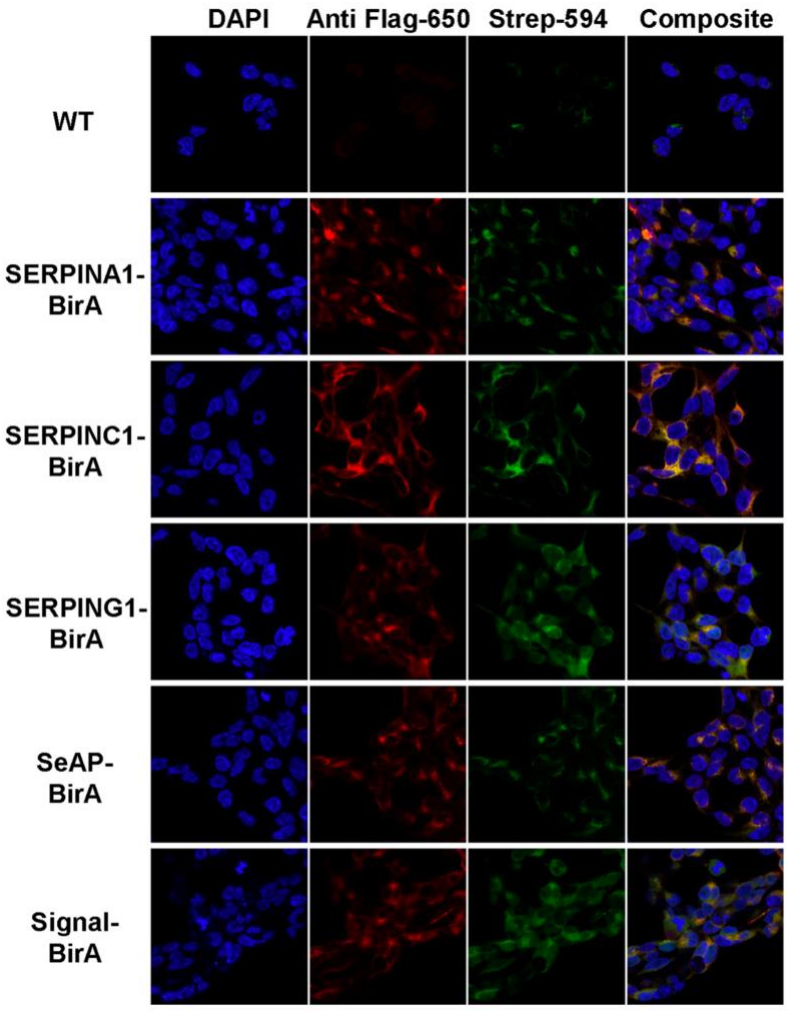


Figure 13. Bait-BirA fusion proteins are colocalized with biotin-staining.
Co-Immunofluorescence demonstrated the intracellular colocalization of the biotin-labeled proteins (stained with Streptavidin-Dylight 594 and illustrated in green color) and bait-BirA (stained with anti-flag monoclonal antibody-Dylight 650 and illustrated in red color), while WT did not show increased biotinylation under the same experimental conditions.

WT cells revealed endogenous biotinylation landscape

After successful tagging of the proximal proteins we aimed to identify the interactions with each bait protein. For this, cells were lysed, and biotinylated proteins were purified using streptavidin. Purified proteins were digested by trypsin, and peptides were subjected to LC-MS/MS, and the biotinylated proteins in samples were mapped using MaxQuant [165]. Differentially biotinylated proteins were then identified in each sample compared to the WT using Perseus [168]. When implemented in a WT control cell line, we identified proteins that are biotinylated

endogenously along with bona fide interactors. These include proteins that bind biotin as a cofactor such as carboxylases which is problematic with streptavidin-based protein detection [173]. Although the extent to which the general endogenous biotinylation has not been systematically quantified, the biotinylated proteins isolated from the WT sample showed considerable overlap with interacting proteins detected in other model protein samples, suggesting endogenous biotinylation may be more pervasive than previously believed. Thus, using the interaction partners detected from the WT sample as background, we filtered out interactions detected in each model bait-BirA sample that were likely a result of endogenous biotinylation. Among the top differentially biotinylated proteins in bait-BirA samples, the bait proteins showed the highest log-fold change (LFC) (Fig. 14). This observation is expected because the bait protein is a potential substrate for BirA located in the closest vicinity of the enzyme and is considered as evidence to show the biotinylation system is working properly within the cell.

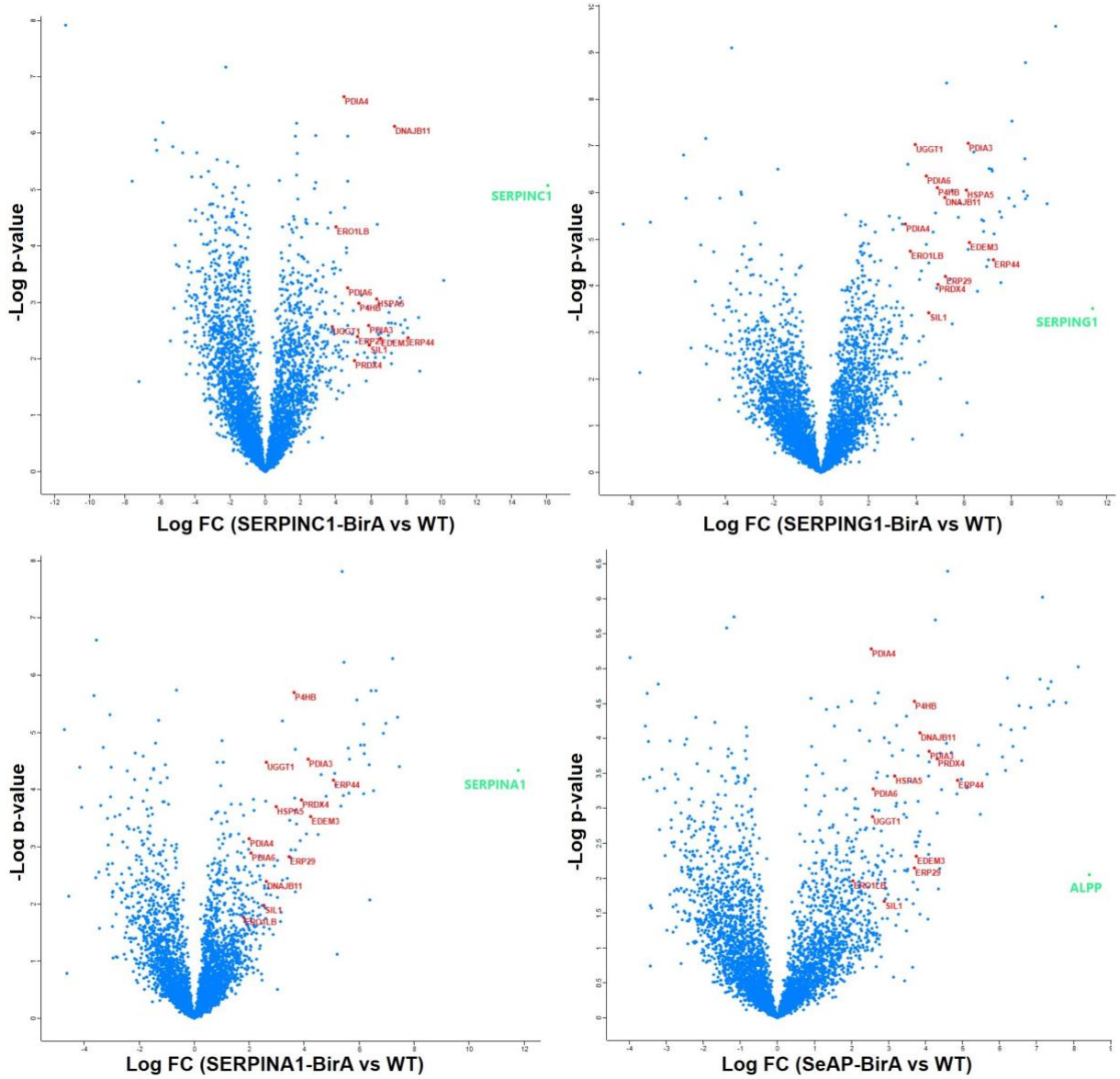


Figure 14. Dozens of proteins show significantly increased biotinylation after expression of bait-BirA proteins.

Volcano plot showing the distribution of the quantified biotinylated proteins by MS according to p-value and fold change. As depicted the bait-protein significantly showed the highest fold change compared to WT almost in all cases, indicating the capability of the BioID labeling system to tag the in vivo interactions within the live cells. The key interactors involved in disulfide bond and N-glycosylation formation (see text) are highlighted in red. SeAP is a truncated form of Alkaline phosphatase, placental type (ALPP).

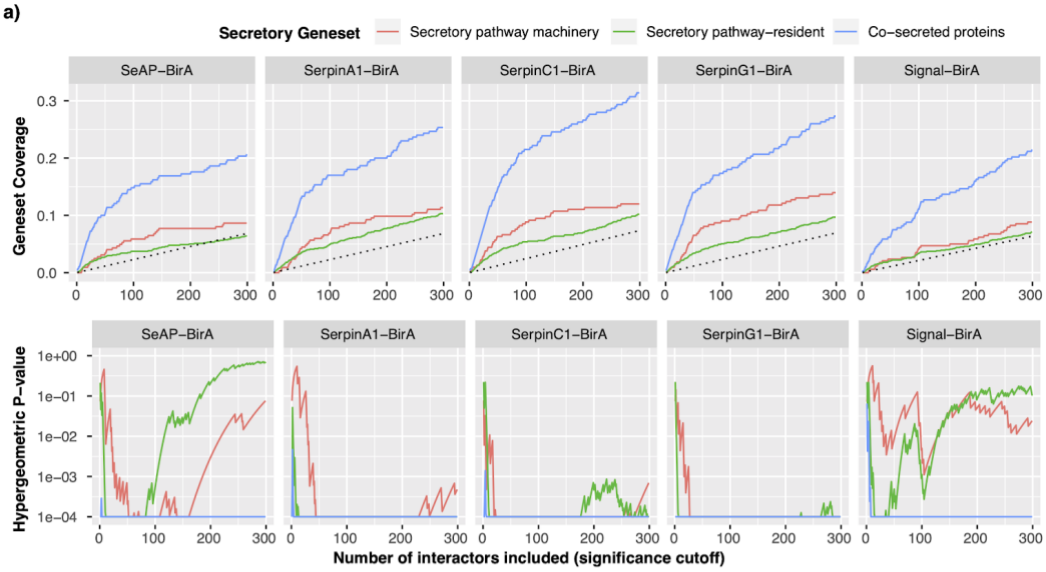
Interactors are enriched for secretory pathway components and co-secreted proteins

In theory, the model proteins would come in frequent contact with members of the secretory pathway, and other co-secreted proteins. To determine if our setup captures the

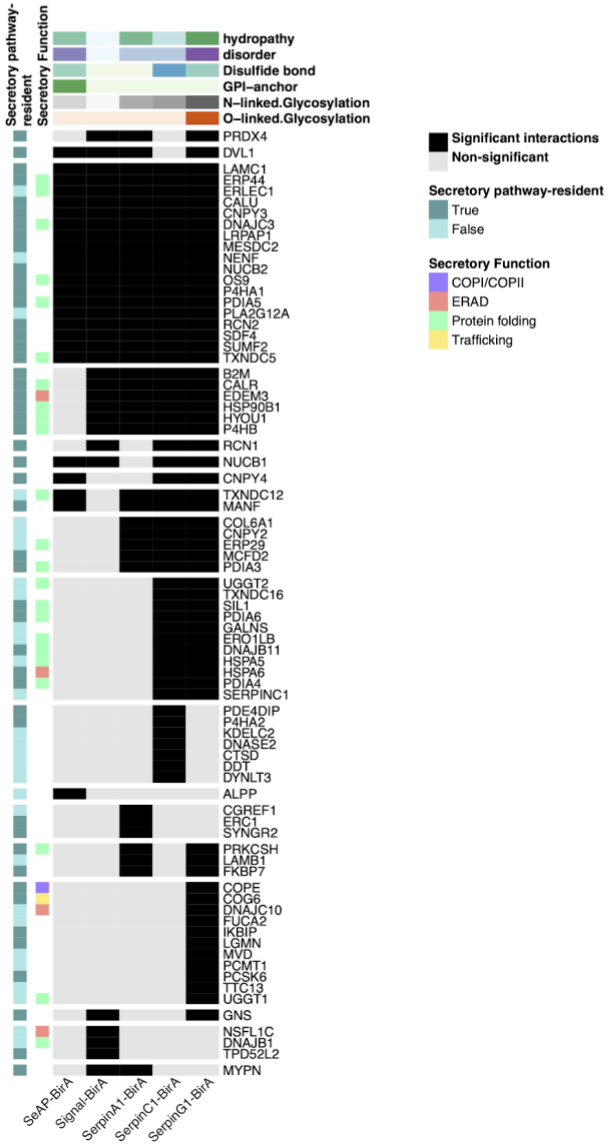
secretory pathway-related proteins more than by random chance, we analyzed the enrichment for 3 independent secretory pathway-relevant gene sets at various fold change and p-value thresholds. We saw a significant enrichment for the secretory pathway machinery, secretory-resident and co-secreted proteins among probable PPIs across all model protein samples (Fig. 15), with peak enrichment occurring at the top 100-300 most significant interactors by positive fold change. This corresponds to a significance cutoff of a fold change of 3 or greater enrichment and an adjusted p-value < 0.1, in model proteins compared to WT control (Fig. 15B for all significant interactions) for each secreted protein (see methods). The secretory machinery components are more enriched among the top 300 hits for all model proteins than other co-secreted proteins, suggesting more frequent interactions between the secretory pathway machinery and their products than the crosstalk between co-secreted proteins. Probable PPIs detected in all model proteins (n=19) and hits shared among all SERPIN gene products are significantly enriched for proteins involved in protein folding (Fig. 15B). Indeed, molecular chaperones are highly promiscuous when assisting protein folding due to their inherent flexibility [174]. Apart from the shared interactions, PPIs for each model protein differ substantially. Thus, the question remained if these private interactions correspond to unique properties of each model protein.

Figure 15. Interacting proteins are enriched for secretory pathway machinery.
(a) To determine if significant interactions enrich for secretory pathway-related genes, we performed an iterative enrichment analysis in which we included the interactors with the greatest fold changes first and iteratively added interactors with lower fold changes. The y-axis indicates the overall coverage of 3 secretory pathway-related gene sets and the x-axis the significance cutoffs (rank ordered by fold change). The coverage of the gene set (top) along with their corresponding hypergeometric enrichment p-value (bottom) are shown, denoting the probability of obtaining an overlap larger than the one observed if no enrichment existed. The top 300 hits for each secretable BirA sample (Fig. S3 for all hits) showed significant enrichment of the secretory pathway components and co-secreted proteins among the most significant hits for all samples except Signal-BirA (which is a lone secreted BirA and not a mammalian secreted protein). (b) Quantified interactions between interactors (y-axis) and the model proteins (x-axis), where the shadowed entries indicate significant interactions. The features of the model proteins, detailed in Fig. S4, are summarized for each model protein on the bottom panel and the secretory pathway attributes for the interactors are labeled on the left.

a)


Secretory Geneset
Secretory pathway machinery
Secretory pathway−resident
Co−secreted proteins
SeAP−BirA
SerpinA1−BirA
SerpinC1−BirA
SerpinG1−BirA
Signal−BirA
Geneset Coverage
Hypergeometric P−value
Number of interactors included (significance cutoff)

b)


Secretory pathway-resident
Secretory Function
hydropathy
disorder
Disulfide bond
GPI−anchor
N−linked.Glycosylation
O−linked.Glycosylation
PRDX4
DVL1
LAMC1
ERP44
ERLEC1
CALU
CNPY3
DNAJC3
LRPAP1
MESDC2
NENF
NUCB2
OS9
P4HA1
PDIA5
PLA2G12A
RCN2
SDF4
SUMF2
TXNDC5
B2M
CALR
EDEM3
HSP90B1
HYOU1
P4HB
RCN1
NUCB1
CNPY4
TXNDC12
MANF
COL6A1
CNPY2
ERP29
MCFD2
PDIA3
UGGT2
TXNDC16
SIL1
PDIA6
GALNS
ERO1LB
DNAJB11
HSPA5
HSPA6
PDIA4
SERPINC1
PDE4DIP
P4HA2
KDELC2
DNASE2
CTSD
DDT
DYNLT3
ALPP
CGREF1
ERC1
SYNGR2
PRKCSH
LAMB1
FKBP7
COPE
COG6
DNAJC10
FUCA2
IKBIP
LGMN
MVD
PCMT1
PCSK6
TTC13
UGGT1
GNS
NSFL1C
DNAJB1
TPD52L2
MYPN
SeAP-BirA
Signal-BirA
SerpinA1-BirA
SerpinC1-BirA
SerpinG1-BirA
Significant interactions
Non-significant
Secretory pathway-resident
True
False
Secretory Function
COPI/COPII
ERAD
Protein folding
Trafficking

Private interactors reflect post-translational and structural features of model proteins

PPIs in the secretory pathway mediate the folding, modification and transportation of secreted proteins [100,175–178]. Incidentally, co-expression analysis has linked certain PTMs across the secretome to the expression of their responsible enzymes. For example, PDIs are consistently upregulated in tissues secreting disulfide-rich proteins [48]. As the bait-BirA proteins differ in structural composition and PTMs, we wondered if bait-BirA proteins with shared features have higher affinity for specific interactors. More specifically, we hypothesize that proteins requiring a specific PTM would preferentially interact with the secretory machinery components responsible for the PTM synthesis. To test if such preferential interaction exists, we summarized various PTM and structural properties across model proteins and analyzed their associations with the corresponding secretory machinery using a Bayesian modeling framework (see methods). Among the studied PTMs, bait-BirA proteins with disulfide bonds and N-linked glycans demonstrated higher affinity towards specific interactions (Fig. 16A) that are known to help secretion of proteins with the corresponding PTMs. Thus, we analyzed the detected interactions associated with glycosylation, disulfide bond addition, and protein folding.

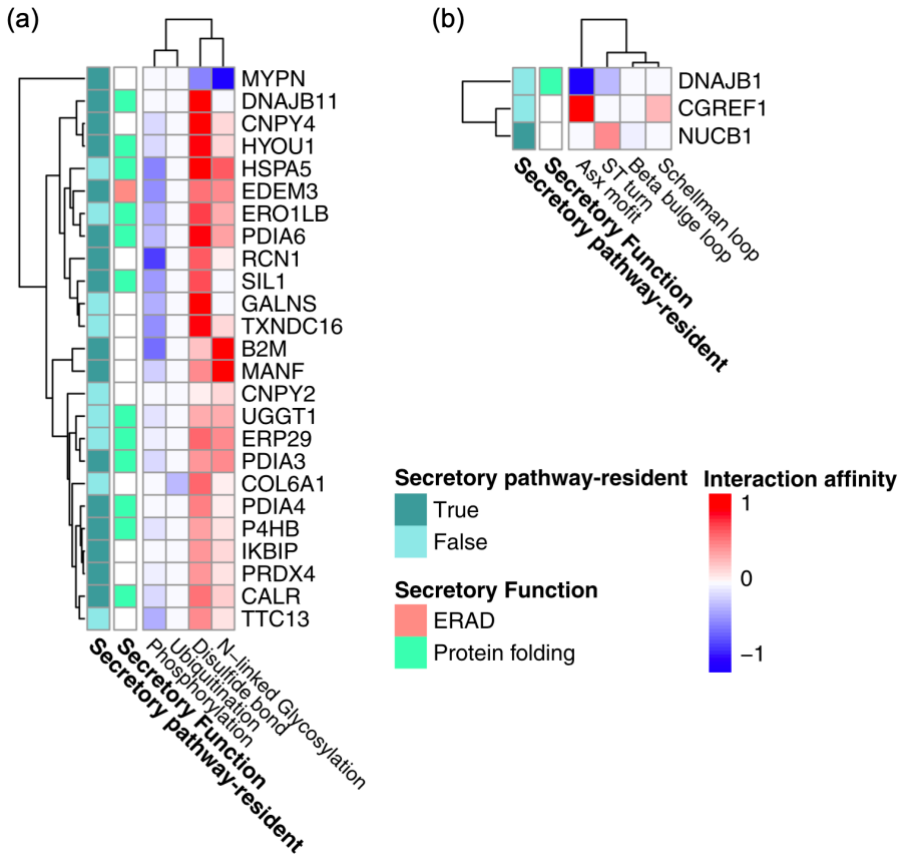


Figure 16. Detected interactors correlate with protein features.
Interactors were associated with specific (a) PTMs and (b) structural features of model proteins. The heatmap shows the standardized interaction affinities estimated between certain interactors and PTMs or structural features across all model proteins (see methods). Only interactors having significant associations with model protein features are shown.

Proteins with increased glycosylation are associated with quality control pathways

We detected significant interactions in the Calnexin/Calreticulin cycle and related processes for more heavily glycosylated proteins (Fig. 16A). For example, the glycosylated baits interacted with calreticulin (CALR), a calcium-binding chaperone that promotes folding, oligomeric assembly, and quality control of glycoproteins in the ER[179]. They also interacted with UGGT1, which recognizes glycoproteins with minor folding defects and reglucosylates single N-glycans near the misfolded part of the protein. Glycosylated proteins are then recognized by CALR for recycling to the ER and refolding or degradation [180]. Two members of the PDI family, PDIA3 and ERp29, which form a complex with calreticulin/calnexin, also showed association with N-glycosylated baites (Fig. 16A) suggesting their role in glycoprotein folding and quality control.

Calnexin/Calreticulin-PDIA3 complexes promote the oxidative folding of nascent polypeptides [181] and ERp29 promotes isomerization of peptidyl-prolyl bonds to attain the native polypeptide structure [181,182]. Proteins with chaperone activity, such as HSPA5 (Fig. 16A), were also found to interact with N-linked glycan-containing bait-BirA. HSPA5 is a component of the glycoprotein quality-control (GQC) system which recognizes glycoproteins with amino acid substitutions, and targets them for ER-associated degradation (ERAD) [183]. EDEM3, another interactor associated with the N-glycan containing proteins (Fig. 16A), is a glycosyltransferase involved in ERAD mediated degradation of glycoproteins by catalyzing mannose trimming from Man8GlcNAc2 to Man7GlcNAc2 in N-glycans [184]. Given that most of these molecular chaperones and enzymes are involved in ERAD mediated degradation of the misfolded glycoproteins these findings suggest the quality control pathways are critical for synthesizing and secreting proteins with N-linked glycans.

Disulfide bond formation is rate-limiting in protein secretion

Several members of the PDI family including P4HB, PDIA3, PDIA4 and PDIA6 significantly interacted with model-BirAs containing more disulfide bonds (Fig. 16A). These enzymes catalyze the formation, breakage and rearrangement of disulfide bonds through the thioredoxin-like domains [185]. The identification of various PDIs highlights the importance of the oxidative folding enzymes in protein folding and maintaining stability that can limit the efficiency of protein secretion. The proteins with more disulfide bonds also interact with major ER chaperones HSPA5 and DNAJB11, a co-chaperone of HSPA5, that play a key role in protein folding and quality control in the ER lumen [186,187], highlighting their important role in secretion of the disulfide bond enriched proteins. The PDI, ERp44 showed the strongest association (LFC > 8) with disulfide bond enriched proteins i.e. SERPINC1 and SERPING1. ERp44 mediates the ER retention of the oxidoreductase Ero1α (an oxidoreductin that reoxidizes P4HB to enable additional rounds of disulfide formation) through the formation of reversible mixed disulfides [188]. Hence, the strong association of ERp44 highlights the importance of the thiol-mediated ER protein retention in

disulfide bond formation, particularly when secretory is loaded with the proteins with more disulfide bonds. In addition, ERO1LB, PRDX4 and SIL1 were ER-localized enzymes that were associated with disulfide bond formation. ERO1LB efficiently reoxidizes P4HB [189], PRDX4 couples hydrogen peroxide catabolism with oxidative protein folding by reducing hydrogen peroxide [190], and SIL1 can reverse HSPA5 cysteine oxidation which alters its chaperone activity to cope with suboptimal folding conditions [191]. The identification of these oxidoreductase enzymes highlights the importance of ER redox homeostasis in disulfide bond formation and protecting cells from the consequences of misfolded proteins.

To validate the importance of specific PDI interactions that showed highest fold change and specificity of effect on productivity of proteins with more disulfide bonds, we knocked down PDIA4, PDIA6, or ERp44 in cells expressing either SERPINC1-BirA or SERPINA1-BirA, using an orthogonal RNAi approach, i.e. esiRNAs [192]. HEK293 cells expressing SERPINC1-BirA or SERPINA1-BirA were transfected with esiRNA against PDIA4, PDIA6, ERp44, or LUC and KIF11 as negative/positive controls, respectively. Knockdown experiments successfully resulted in >70% reduction in cellular expression of the targeted PDIs (Fig. 17A). We observed a 42% reduction in SERPINC1-BirA secretion following knockdown of PDIA4 (p-value = 0.032), meanwhile a significantly lesser reduction of SERPINA1-BirA secretion was seen (Fig. 17B). If differences in KD efficacy between experiments are accounted for, normalization of SERPINs secretion by KD efficiencies resulted in 54% reduction of SERPINC1-BirA secretion (p-value = 0.026) while no significant reduction was seen in SERPINA1-BirA secretion (Fig. 17C). Neither knockdown of PDIA6 nor ERp44 produced a significant reduction in either SERPINC1-BirA nor SERPINA1-BirA secretion. BioID analysis indicated that PDIA4 interacts significantly with SERPINC1-BirA but not with SERPINA1-BirA, while PDIA6 did not show considerable interaction with SERPINA1 nor SERPINC1. Therefore, PDIA4 may work as a private interactor for SERPINC1 secretion as recognized by knockdown results. PDIA6 could potentially work as a private interactor for SERPING1 secretion as predicted by BioID. ERp44 had a higher quantified

fold change interaction with SERPINC1-BirA, but it also showed significant interaction with SERPINA1-BirA, SERPING1-BirA and less strongly with SeAP-BirA, suggesting ERp44 may function in a more public manner.

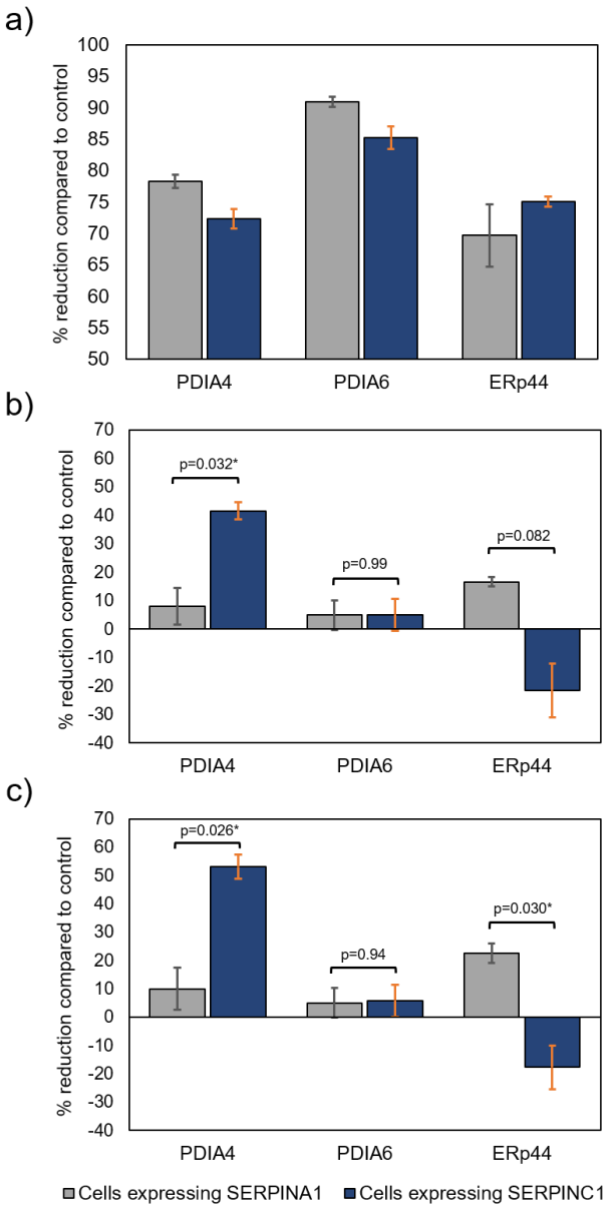


Figure 17. Effects of esiRNA mediated knockdown of isomerases PDIA4, PDIA6, and ERp44 on SERPIN secretion.

Quantitative western blots were used to find the relative abundance of each PDI and SERPIN after esiRNA transfection in both cells expressing SERPINA1 and SERPINC1. We have reported these results as (a) KD efficiency of each PDI measured by percent reduction of PDI abundance compared to the negative control, (b) the effect of each PDI KD on SERPINA1 and SERPINC1 secretion, and (c) the normalized effects of PDI KD on SERPIN secretion using the KD efficiency, assuming greater KD results would result in a greater loss of SERPIN secretion. SERPINC1, which has a multitude of disulfide PTMs, shows a greater reduction in SERPIN production in PDIA4 experiments compared to SERPINA1, which does not have as many disulfide bridges in its structure. This is consistent with expected interactions found with BioID analysis.

Identified PPIs are associated with structural motifs on bait proteins

In addition to PTMs, the Bayesian modeling framework found associations between SecP structural features and the SecMs (Fig. 16B). For example, model proteins depleted in the asx motif [193] showed higher tendency to interact with DNAJB1, a molecular chaperone of the HSP40 family. The asx motif impacts N-glycan occupancy of Asn-X-Thr/Ser sites, depending on the ability of the peptide to adopt an Asx-turn motif [194,195]. As another example, NUCB1, a chaperone-like amyloid binding protein that prevents protein aggregation [196], interacted more strongly with our proteins with more ST-turns (Fig. 16B). ST turns occur frequently at the N-termini of α-helices [197] and are regarded as helix capping features which stabilize α-helices in proteins [198]. Thus, the enriched interaction of the NUCB1 with St-turn suggests that it can help stabilize folding of protein with a predominant α-helical secondary structure.

Discussion

BioID has been used to profile the proteome of different cellular compartments and molecular complex systems [156]. However, this is the first time that BioID has been used to identify the proteome of the secretory pathway during the recombinant protein expression. Numerous PPIs guiding the folding, modifications, and trafficking of the secreted and membrane proteins through the secretory pathway are transient [199,200] and therefore cannot be captured by conventional methods such as co-immunoprecipitation. Consistent with previous studies [201], these results showed the BioID can detect weak and transient interactions in situ, and therefore it is a powerful approach to study luminal processes involved in protein secretion. We found that disulfide bridge formation enzymes showed the strongest association with bait proteins enriched in disulfide bonds, supporting their critical roles in protein secretion and maintaining ER stability. A previous study on difficult to express (DTE) monoclonal antibodies showed less recognition by PDI impairs disulfide bridge formation within the antibody light chain (LC) which can initiate the intracellular degradation by the ubiquitin proteasome system via ERAD [202]. Thus, insufficient

interaction between the secreted proteins with enriched disulfide bonds and PDIs can limit secretion efficiency and serve as a rate-limiting step for protein production. In another study, the tissue specific analysis of SecMs expression showed a positive correlation between the expression of P4HB and PDIA4 and liver tissue where numerous disulfide bond enriched proteins are secreted by hepatocytes [48]. These observations are clear evidence that suggests the tissue-specific fine-tuning of the PDI family expression level in response to the enrichment of the disulfide sites. Together, these results showed PDIs are actively involved in adaptive responses and secretion of proteins with dominant disulfide bonds which are crucial for restoring ER stability, and therefore yielding the recombinant proteins. Given the associations between the SecMs and the features of the model proteins, we also hypothesized that SecMs preferentially interacting with bait-BirA proteins that carry certain structural features may be essential for the secretion of those proteins. While evidence linking SecMs to the structural motifs is lacking, many molecular chaperones selectively interact with certain sequence and structural elements to favor the particular folding pathways [203]. For example, chaperones of the HSP70 family evolved to bind extended β strand peptides; interestingly, the associations identified between chaperones and asx motif and ST turn represent a novel association for further study.

While we show BioID works well for studying the synthesis of secreted proteins, we acknowledge that biotin-based methods have some limitations as well. Biotin is actively imported into the cytoplasm of cells and can freely diffuse to the nucleus, but it may not be as accessible in the secretory pathway, thus reducing labeling efficacy in that compartment [204]. Here we showed this challenge is not an insurmountable issue, in that the BioID2 construct successfully identified many expected luminal interactions. BioID2 requires less biotin supplementation, and exhibits enhanced labeling of proximate protein [161] allowing for BioID to be introduced to new systems where biotinylation supplementation may not be easily accomplished [201]. More recently, two promiscuous mutants of biotin ligase, TurboID and miniTurbo, have been developed to catalyze proximity labeling even with much greater efficiency [205] and therefore can be considered as an

effective method when proximal labeling of the endomembrane organelles is desired. It is possible that the coverage of biotin labeling may also be limited in the early processes in the secretory pathway due to the maturation process of the biotin ligase, i.e. biotinylation only can occur once the biotin ligase has been fully transcribed and functional. Therefore, the C-terminal location of birA may limit the labeling of some early components until it is fully folded. Our results did reveal that there was coverage of the early secretory pathway interactions, but the possibility of missing interactions cannot be ruled out. So, BioID should be seen as a complementary method to other PPIs and systems biology approaches for complete characterization of cellular interactions.

In summary, we demonstrate here an approach to identify the protein interactions that synthesize and support secreted proteins, and thus define the product-specific secretory pathway. The identification of such machinery opens avenues for mammalian synthetic biology, wherein biotherapeutic production hosts can be rationally engineered to improve the titer and quality of diverse proteins in a client specific manner.

Chapter 4, in full, is a reprint of the material as it appears in Samoudi M, Kuo CC, Robinson CM, Shams-Ud-Doha K, Schinn SM, Kol S, Weiss L, Bjorn SP, Voldborg BG, Campos AR, Lewis NE. “In situ detection of protein interactions for recombinant therapeutic enzymes”. *Biotechnology and Bioengineering*, 2020. The dissertation author was the co-first investigator and author of this material.

CHAPTER 5: ESSENTIAL COMPONENTS OF THE SECRETORY PATHWAY CORRELATING WITH HIGH PRODUCTIVITY IN ANTIBODY-PRODUCING CHO CELLS

Abstract

CHO cells are the industrial mammalian cell line of choice for producing recombinant therapeutic proteins, e.g., monoclonal antibodies, cytokines and vaccines. However, while some proteins are efficiently expressed, many fail to secrete well, thus making many proteins inaccessible as therapeutics. Despite adequate mRNA levels, proteins may fail to express well due to bottlenecks downstream of transcription. Thus, to enable efficient manufacturing of high-value proteins, we must decipher molecular mechanisms controlling the secretory pathway during protein expression. Such knowledge would inform rational strategies for secretory pathway engineering and recombinant protein production. Here, we established methods to unravel the host cell secretory pathway machinery that directly regulates protein secretion, specifically secretion of monoclonal antibodies (mAbs). We used proximity-biotinylation [206,207] with an Fc-mediated biotinylation by antibody recognition (FcBAR) method, to quantify protein-protein interaction (PPI) networks regulating protein secretion and identify PPIs correlating with secretion rate. To systematically analyze how changes in PPI strength impact secretion, we developed a stochastic queuing model that incorporates PPIs and enzyme expression to estimate and locate potential bottlenecks within the secretory pathway. Our tools guide host cell engineering for biomanufacturing of diverse proteins, and provide deep insights into the functions of the mammalian secretory pathway.

Introduction

Eukaryotic cells, including Chinese hamster ovary (CHO) cells are essential cell factories used for producing recombinant therapeutic and industrial proteins, e.g., monoclonal antibodies,

enzymes, cytokines and vaccines. However, many important proteins are difficult to produce even when the cells have adequate mRNA levels. Since they fail to secrete well from the cells, they are thus inaccessible as products. Many of these proteins may fail to express well due to bottlenecks in the cellular machinery used by the secretory pathway, e.g. involving protein translation and/or posttranslational events during protein production. Thus, to enable efficient manufacturing of high-value proteins, we must decipher molecular mechanisms controlling the secretory pathway during protein expression for each protein of interest. Such knowledge would inform rational strategies for engineering the secretory pathway of cell factories to improve recombinant protein production. Through this work we are developing methods to define the molecular machinery a cell factory needs to produce high-value proteins, and advanced machine learning techniques to guide cell engineering.

Here we analyzed in situ protein-protein interactions generated by proximity biotinylation with Fc-mediated biotinylation by antibody recognition (FcBAR) between the mAbs/related proteins and the cellular secretory pathway machinery. We applied FcBAR to Rituximab-producing CHO cells and identified essential PPIs for a secreted protein and those correlating with secretion rate. To systematically analyze how changes in PPI strength impact secretion, we developed a stochastic queuing model that estimates and locates potential bottlenecks within the secretory pathway.

This computational framework will help define roles of interacting proteins and prioritize PPIs for further study. Our tools will guide host cell engineering for biomanufacturing of diverse proteins, and provide deep insights into the functions of the mammalian secretory pathway.

Methods

To estimate the degrees to which Rituximab production is throttled for each clone, and more specifically, to determine the subcellular localization of production bottlenecks, we constructed a queuing model that traces the reactions Rituximab undergoes within the secretory

pathway. We built product-specific reaction networks [208] for both the light and heavy chains of Rituximab that take into account the synthesis, post translational modification and transportation as prescribed by their PTM and sequence features. We parsed the gene-protein-reactions (GPRs) at the enzyme level and arranged reactions into groups of functional modules (Fig. 18). Boolean logic was applied to summarize the gene expression for the module, where the maximum expression from the parallel alternative reactions was used as the proxy for module expression whereas the minimum gene expression was used for chains. For PPI intensities, the average LFQs were used as the representative for each module (Fig. 18B).

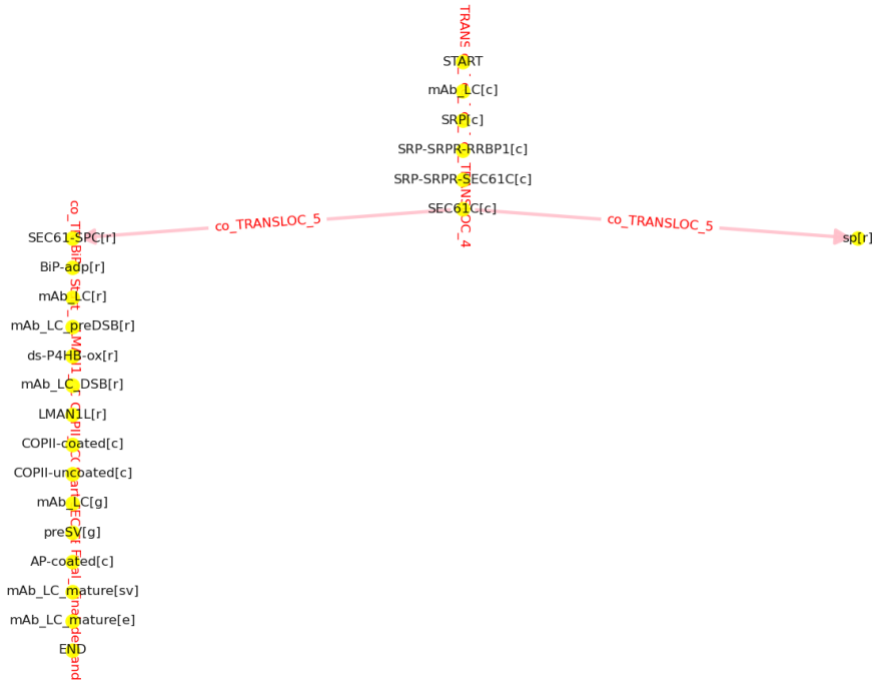
To simulate congestion at each module, we adopted principles in queuing theory and extended our PPI support framework to account for the stochastic transitions across modules (Eq. 1). Using the original reaction network as the backbone, we turned each module into a queuing mode and approximated the transitions and levels of congestion by optimizing the model parameters such that measured PPIs and yield best match model predictions (Fig. 19A). During transition and congestion optimization, the module gene expression is taken into account in which more highly expressed modules are less prone to congestion (Fig. 19B). We approximated the posterior distribution of model parameters given the measured PPIs and yield by MCMC sampling (Fig. 19CD).

$$P(\mathbf{v}_{PPI}|\mathbf{L},\mathbf{T},\boldsymbol{\theta},\mathbf{C},\mathbf{S},\mathbf{e}) = P(\mathbf{v}_{PPI}|\mathbf{L})\prod_{i=1}^{t}(P(\mathbf{l}_i|\mathbf{t}_i,\mathbf{l}_{i-1})P(\mathbf{t}_i|\boldsymbol{\theta},\mathbf{c}_i)P(\mathbf{c}_i|\boldsymbol{\theta},\mathbf{l}_{i-1},\mathbf{e})P(\mathbf{s}_i|\mathbf{m}))P(\mathbf{t}_0)P(\mathbf{c}_0)P(\mathbf{l}_0)$$

$$\mathbf{l}_i \sim \text{Multinomial}(\mathbf{l}_{i-1}+\mathbf{s}_i,\mathbf{t}_i)$$
$$\mathbf{t}_i \sim \text{Dirichlet}(\boldsymbol{\theta}+\text{diag}(\mathbf{c}_i))$$
$$\mathbf{c}_i \leftarrow \lambda_a\left(\textstyle\sum_n \boldsymbol{\theta}_{m,n}\right)\cdot\max\left\{0,\log\frac{\lambda_b+\mathbf{l}_i}{\lambda_c+\mathbf{e}}\right\}$$
$$\mathbf{s}_i \sim \text{Poisson}(\lambda_s\mathbf{m})$$

Equation 1

A



B

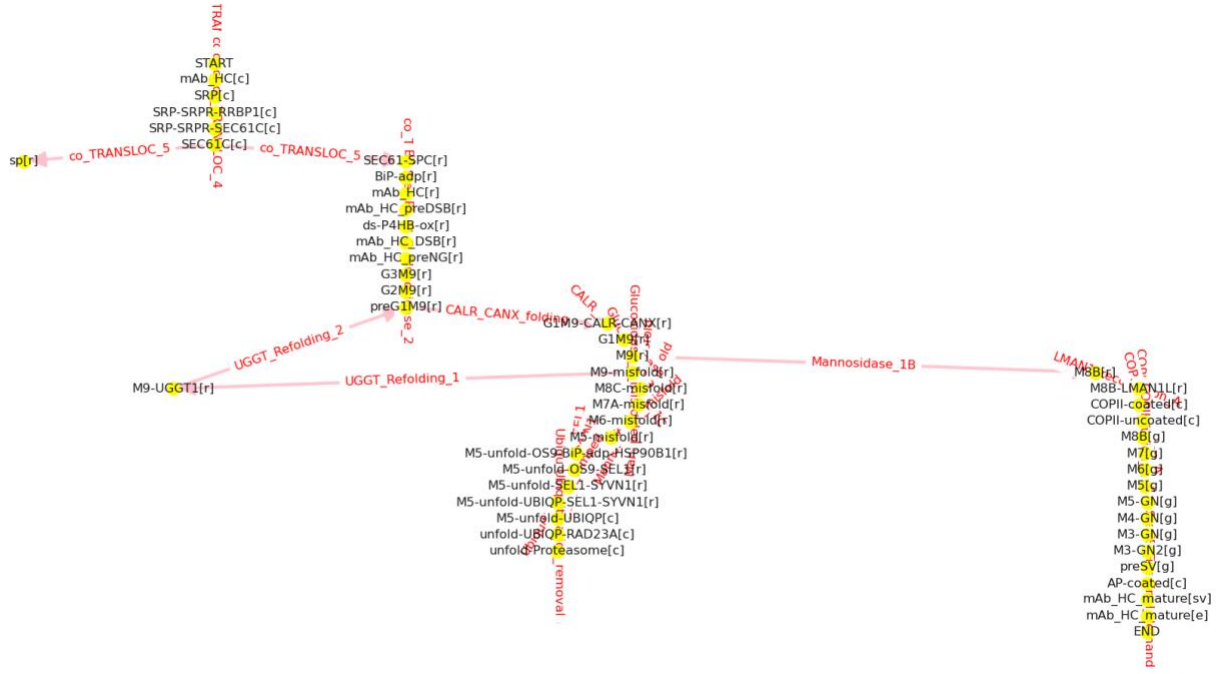
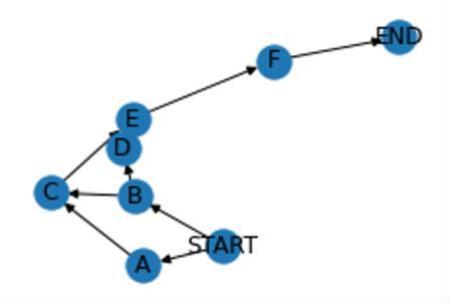


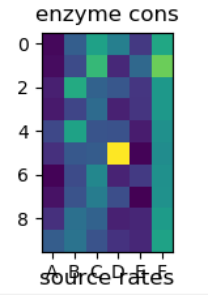
Figure 18. Modular reaction networks for Rituximab light and heavy chains.
(A) Rituximab light chain reaction network.
(B) Rituximab heavy chain reaction network with module expression overlaid. Edge color and edge label denote module gene expression and averaged PPI intensities respectively.

Figure 19. Toy model and stochastic transition at various time points.
(A) A toy reaction network consisting of several parallel and chain reactions.
(B) Enzyme concentration m_n at each module n.
(C) Location of molecules across time points, as sampled by the stochastic queuing model.
(D) The congestion rate at each module across the 10 clones.

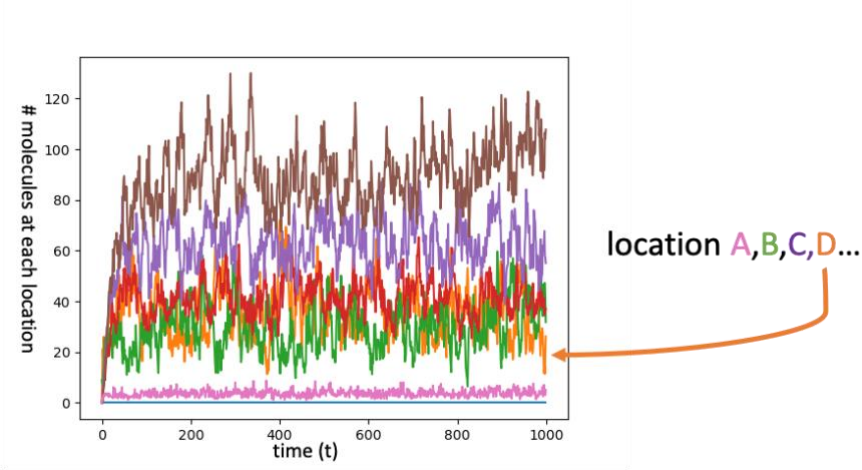
A


END
F
E
D
C
B
START
A

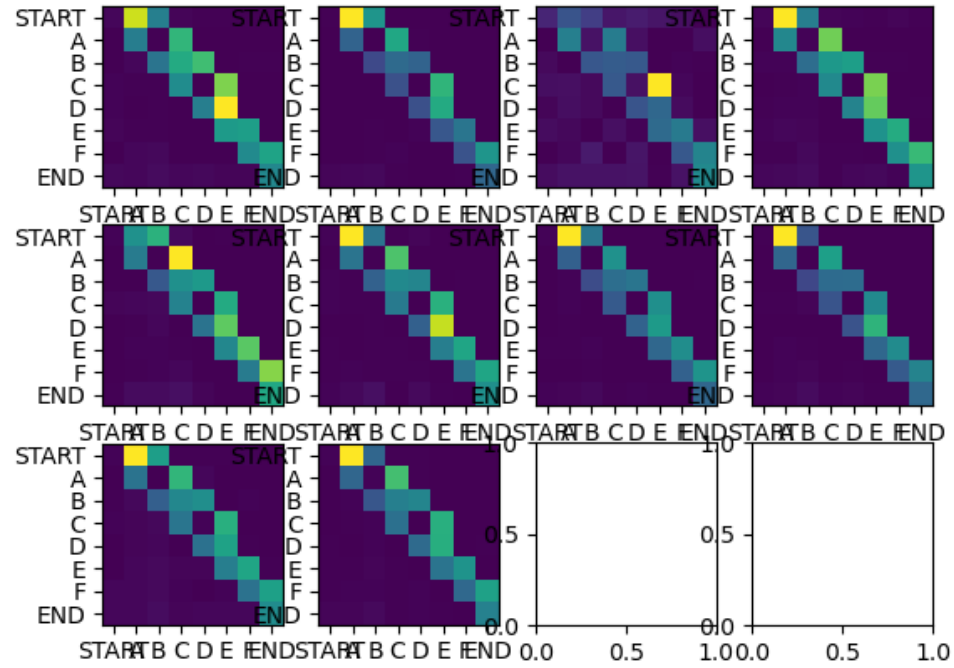
B


enzyme cons
0
2
4
6
8
A B C D E F
source rates

C


molecules at each location
0
20
40
60
80
100
120
0
200
400
600
800
1000
time (t)
location A,B,C,D...

D


START
A
B
C
D
E
F
END
0.0
0.5
1.0

Results

FcBAR captured interactions between Rituximab and secretory pathway machinery

In theory, secreted proteins would come in frequent contact with members of the secretory pathway, and other co-secreted proteins. To determine if our setup captures the secretory pathway-related proteins more than by random chance, we analyzed the enrichment for 3 independent secretory pathway-relevant gene sets at various fold change and p-value thresholds. We saw a significant enrichment for the secretory pathway machinery, secretory-resident and co-secreted proteins among probable PPIs across all Rituximab-producing clones (Fig. 20).

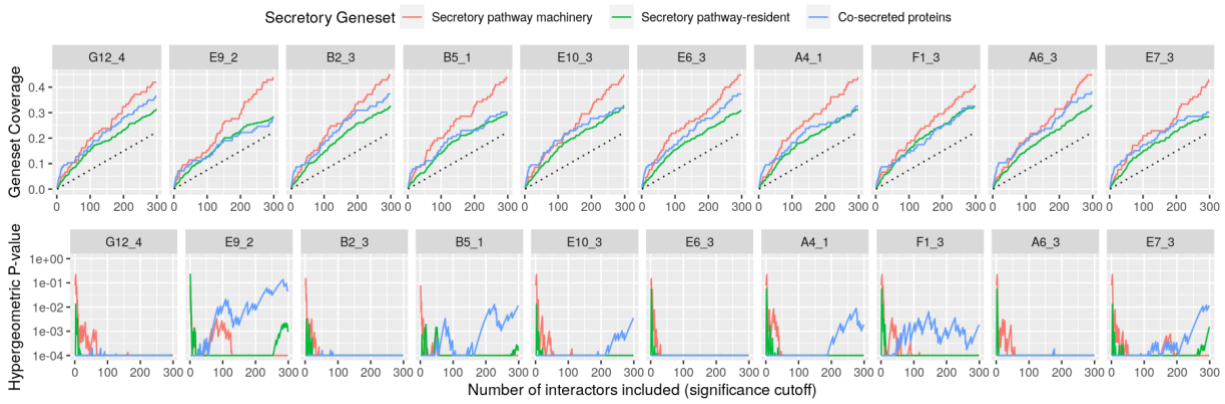


Figure 20. Interacting proteins are enriched for secretory pathway machinery.
To determine if significant interactions enrich for secretory pathway-related genes, we performed an iterative enrichment analysis in which we included the interactors with the greatest fold changes first and iteratively added interactors with lower fold changes. The y-axis indicates the overall coverage of 3 secretory pathway-related gene sets and the x-axis the significance cutoffs (rank ordered by fold change). The coverage of the gene set (top) along with their corresponding hypergeometric enrichment p-value (bottom) are shown, denoting the probability of obtaining an overlap larger than the one observed if no enrichment existed. The top 300 hits for each secretable BirA sample (Fig. S3 for all hits) showed significant enrichment of the secretory pathway components and co-secreted proteins among the most significant hits for all samples except Signal-BirA (which is a lone secreted BirA and not a mammalian secreted protein).

Rituximab mRNA levels correlated with titers

Although some studies find that transgene mRNA levels do not correlate with the amount of secreted protein[209,210], we performed a simple linear regression to estimate and decouple the effect of mRNA levels on protein titers. A positive correlation between protein titers and Rituximab light chain mRNA levels was found (bootstrapping p=0.03). A combination of high light and heavy chain (bootstrapping p=0.1) mRNA levels results in even more elevated titers (Fig. 21).

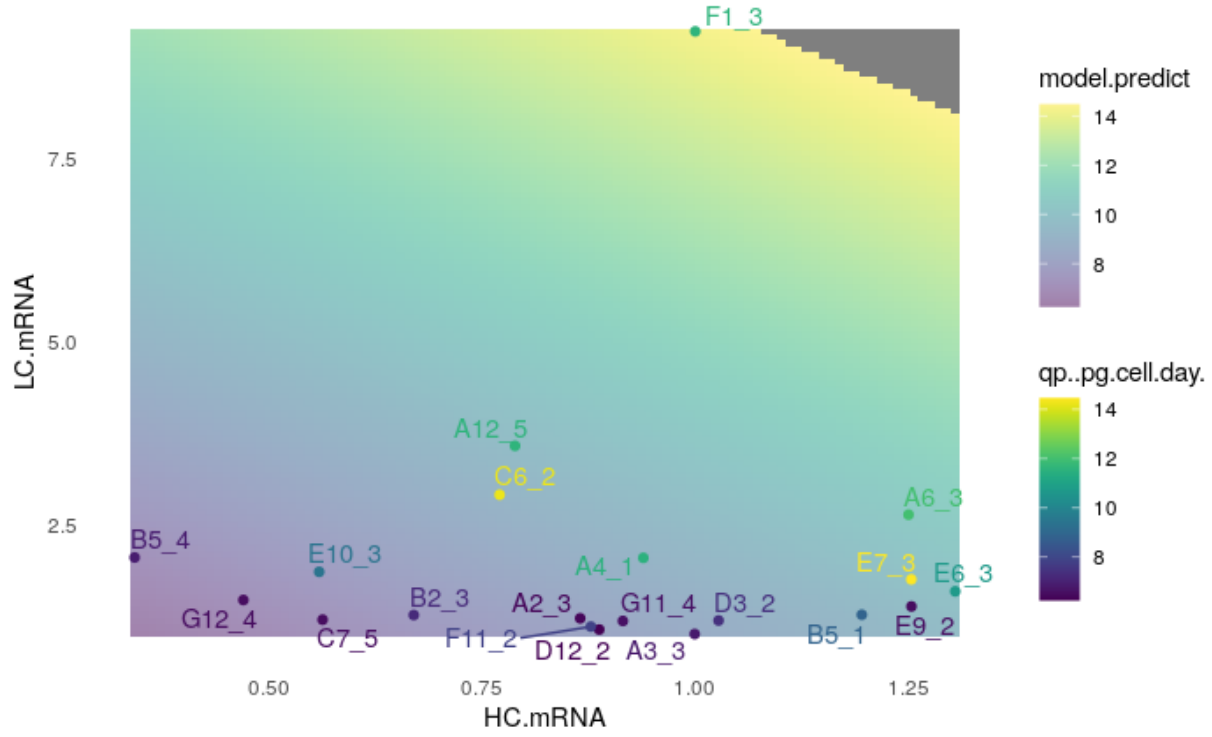


Figure 21. Rituximab specific productivity and model predictions.
Rituximab light and heavy chain and corresponding specific productivity
the true qp (color of the data points) and the model prediction (background). The model errors can be calculated as the difference between the two. The major model outliers with high errors are clones E7_3, F1_3, B9_2 and C6_2.

Rituximab in high producers tend to interact less frequently with secMs

To evaluate if certain subsystems within the secretory pathway show titer-dependent interaction, we summarized interactions across proteins in a given secretory pathway subsystem across different productivity metrics, and noticed a strong negative correlation between high producers and secretory pathway interactions (Fig. 22).

This suggests Rituximab in high producers either clears the secretory pathway much more efficiently than it does in low producers, or alternative secretion pathways may be involved.

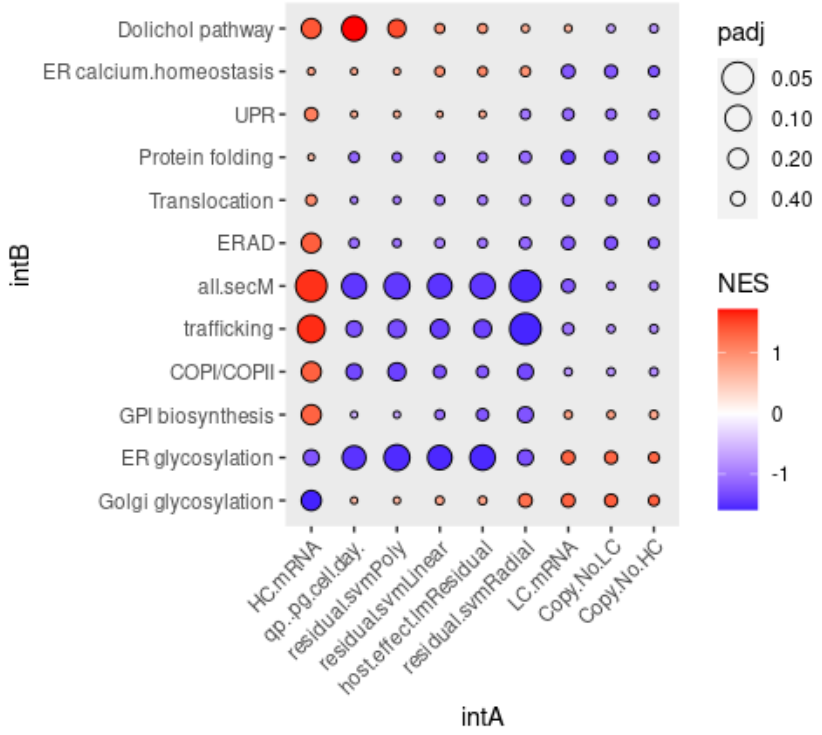
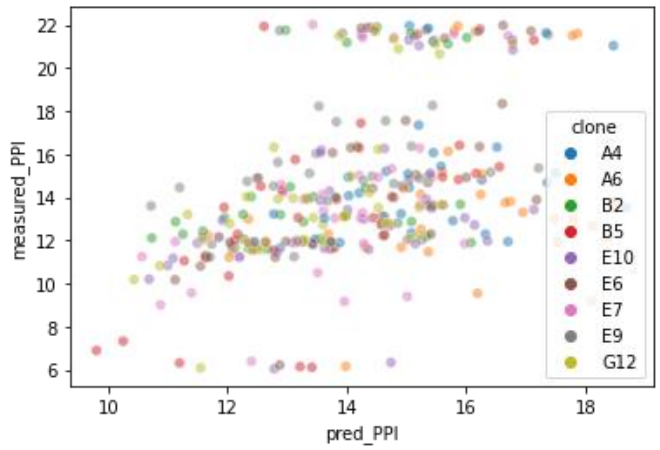


Figure 22. Pathway-level correlation analysis.
To see if interactions with genes from certain functional groups favor/ hinder r-protein production, pathway-level enrichment of the correlation coefficients was performed. Key pathways whose constituent genes show consistent positive/ negative correlations with titer variables are indicated by red/ blue dots respectively.

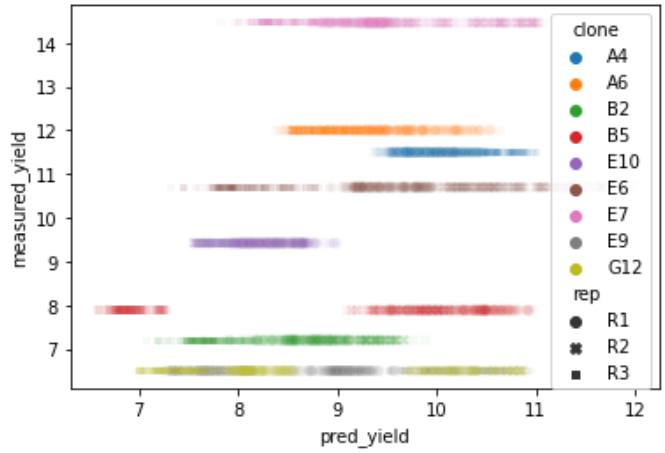
Stochastic model approximated PPIs and specific productivities with high accuracy

The model works by jointly optimizing parameters for congestion and transition such that predicted protein accumulation at each module and the protein yield match the actual protein measurement and yield as closely as possible. The parameters of interest are the congestion rate at each site and key transition probabilities for Rituximab heavy chain. Note that transition probability for the Rituximab light chain synthesis network is trivial as it adopts a chain structure with no branching points (Fig. 18A). To ensure accurate and robust parameter estimation, we first examined the degree to which posterior predictive PPI and yield match their respective measurements. For the rituximab light chain, both the predicted PPIs and the yield are distributed closely around their measurements (Fig. 23AB).

A



B



C

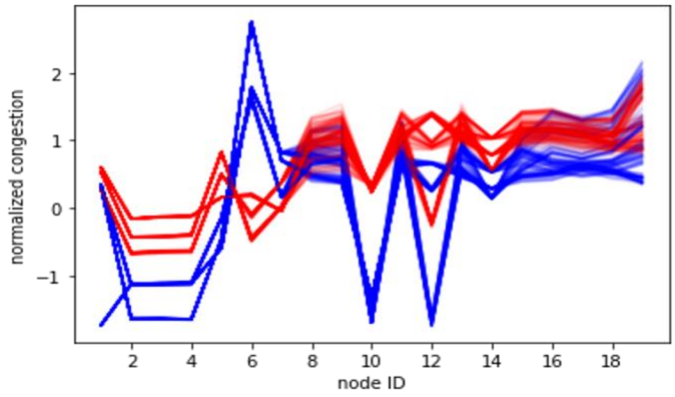


Figure 23. PPIs and specific productivities across clones approximated with posterior model distributions.

Predicted congestion

While we do not see a general association between yield and congestion, pairwise comparison of the titer outliers (highest yield: E7, rows 18-20; lowest yield: E9, rows 21-23) reveals that modules in lower yield clones tend to be more congested (Fig. 23C).

Chapter 5, in part, is currently being prepared for submission for publication of the material. Samoudi M, Kuo CC, Robinson CM, Lewis NE. “Essential components of the secretory pathway correlating with high productivity in antibody-producing CHO cells” The dissertation author was the co-first investigator and author of this material.

CHAPTER 6: DYSREGULATION OF THE SECRETORY PATHWAY CONNECTS ALZHEIMER'S DISEASE GENETICS TO AGGREGATE FORMATION

Summary

Amyloid disorders, such as Alzheimer's disease (AD), involve aggregation of secreted proteins. However, it is largely unclear how hundreds of secretory pathway proteins contribute to amyloid formation. We developed a systems biology framework that integrates expression data with protein-protein interaction networks to estimate a tissue's fitness for producing specific secreted proteins. Using this, we analyzed the fitness of the secretory pathway of various brain regions and cell types for synthesizing the Alzheimer's disease-associated amyloid-precursor protein (APP). While key amyloidogenic pathway components were not differentially expressed in AD brain, we found Aβ deposition correlates with repressed expression of the secretory pathway components proximal to APP and systemic up-regulation of the secretory pathway components proximal to β- and γ-secretases in AD. Our analyses suggest that perturbations from 3 AD risk loci cascade through the APP secretory machinery support network and into the endocytosis pathway. This suggests amyloidogenesis is associated with dysregulation of secretory pathway components supporting APP, thus suggesting novel therapeutic targets for AD treatment.

Introduction

No mammalian cell exists alone. Indeed, each cell dedicates >1/3 of its protein-coding genes to interact directly with other cells and its environment [70,211], using hormones and receptors for communication, enzymes and other proteins to modify their extracellular matrix, transporters for exchanging metabolites, etc. The mammalian secretory pathway is tasked with the synthesis, post-translational modification (PTM), quality control, and trafficking of these secreted proteins (secPs) [150,151]. SecPs account for >25% of the total proteome mass[212,213], and are among the most

tissue-specific genes in the human genome [211]. The precision and efficiency of the mammalian secretory pathway result from the concerted effort of hundreds of secretory machinery components (secMs) including chaperones, enzymes, transporters, glycosyltransferases, metabolites and lipids within the secretory pathway [47,48,97,214]. Since many secPs relay signals between cells or modify a cell's microenvironment, each cell must carefully regulate the synthesis and localization of each secP.

Perturbations to the secretory pathway result in misfolded proteins, which induce ER stress and apoptosis. In amyloid diseases, the misfolded proteins can aggregate into toxic amyloid fibrils, which ultimately lead to cell death. Aβ deposition, a major pathological hallmark of Alzheimer's disease (AD), stems from the perturbed processing of the transmembrane amyloid precursor protein (APP). In Aβ aggregation, alternative proteolytic cleavage of amyloid precursor peptide by β- (rather than α-secretase) releases the secreted form of APP, the aggregation-prone Aβ1-42.[215–217] Furthermore, additional PTMs in the secretory pathway may affect APP cleavage [218,219], phosphorylation [220], glycosylation [221–224] and trafficking [219,225]. However, protein aggregation could stem from the perturbation of diverse processes, but no systematic exploration of all processes supporting proper APP processing has been done [226]. Several large-scale GWAS also identified more than 45 AD risk loci [227–229], although for many loci, it remains unclear how they induce AD pathology. Furthermore, a large part of AD heritability remains unknown [227,230]. The genetic landscape of late-onset Alzheimer's (LOAD) is highly heterogeneous, with multiple complex molecular interactions contributing to the disease phenotype. Therefore, the discovery of concerted expression changes in LOAD, such as the remodeling of immune-specific modules requires systems approaches on large datasets [231].

To unravel the molecular changes leading to Aβ deposition, we focused on the roles of the secretory pathway in amyloidogenesis. The secretory pathway is responsible for the processing, quality control and trafficking of key components of the amyloidogenic pathway [232,233], such as APP and the secretases, so we investigated if there is a systemic dysregulation of the

secMs supporting their production and processing. To do this, we first developed a network-based approach that leverages protein-protein interaction (PPI) and mRNA and protein abundance data to quantify a cell or tissues' "secretory machinery support". This measures the fitness of a tissue or cell for properly secreting a specific secP based on the expression of its supporting secMs. Next, we investigated if there are disruptions in the secretory machinery support for key players of the amyloidogenic pathway (i.e., APP and the secretases), leading to increased Aβ deposition in LOAD, based on data from several large-scale clinical bulk- and single-cell RNA-Seq datasets [234,235]. We found significant dysregulation of the secretory pathway proximal to APP and the secretases, and this dysregulation is a major determinant of Aβ deposition. We further demonstrated that the concerted expression changes in the secretory support modules for the APP, BACE1, and PSEN1 can be linked to known AD risk genes and their regulation targets. In terms of subcellular localization, the core perturbed network enriches for known hotspots for Aβ production such as ER, cytosol and endosomes. Moreover, we found that the AD risk loci activate endocytosis via the core support network, and we identified a candidate TF binding motif that is conserved in the promoter regions of the interaction network genes. Together, our analyses suggest mechanisms underlying impaired protein secretion, which could propose novel therapeutic targets for the treatment of AD. It also proposes mechanisms by which AD genetics imbalance the secretory pathway, thus resulting in Aβ deposition, cell death, and cognitive impairment.

Results

Secreted proteins and secretory machinery show similar tissue-specific expression

The secretory pathway synthesizes and transports a variety of secreted proteins, each with different requirements for their synthesis and secretion (e.g., different physicochemical properties and post-translational modifications). With the human secretome being one of the most

tissue-specific subsets of the human proteome [211,236], we hypothesized each tissue expresses just the secMs needed to synthesize and process secPs from the tissue. Supporting this, we observed that clustering tissues by secP gene expression grouped tissues similarly as when clustering by secM gene expression (Fig. 24, p-value = 0.0145). Thus, the secMs are not merely housekeeping proteins always expressed to support any proteins being secreted; rather, they express in a tissue-specific fashion to meet the demands of different tissues [48]. However, the question remains if the pairings between secMs expressed in each tissue represent those needed to specifically support the secPs they secrete.

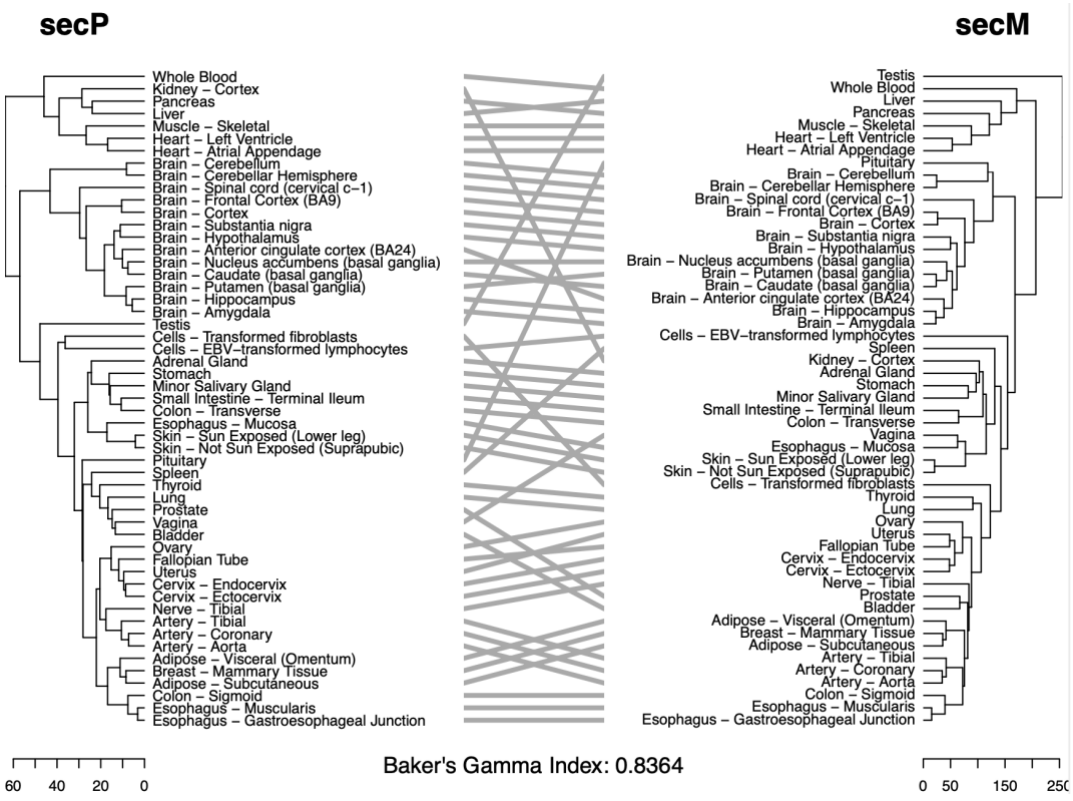


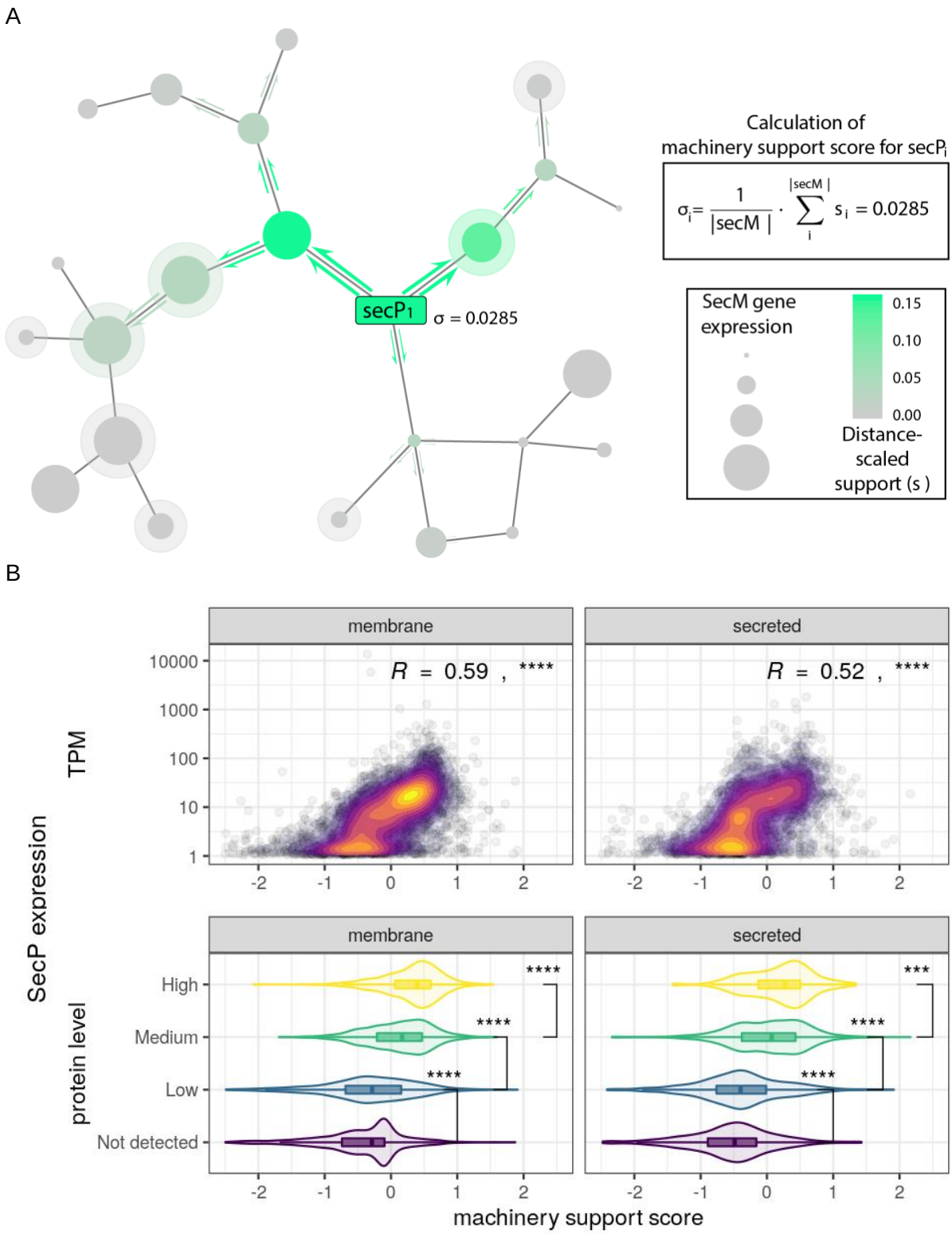
Figure 24. The secMs and the secPs show coordinated expression profiles across different human tissues.
The tissue similarity structures from the perspective of the secretome and the secretory pathway expression were represented by two hierarchical clustering dendrograms, which were then compared with a tanglegram [237]. Gene expression of the secretory pathway and its clients show a high level of coordination across human tissues, with precision significantly better than expected by sampling genes from each tissue (bootstrapping p-value 0.013 and 0.045 for data from GTEx [238] and Human Protein Atlas [211] respectively).

Tissue specific expression of secMs predict expression of their client-secreted proteins

To further dissect the pattern of secP-secM co-regulation seen across tissues, we incorporated two sources of information: protein-protein interactions (PPIs) and secM gene expression. We harness PPIs to identify the secMs relevant to each secP, since PPIs are one of the major modalities through which machinery proteins in the secretory pathway assist protein secretion [100,176,239,240]. Further, secMs responsible for secP post-translational modifications are well-captured by PPIs between the secPs and the secMs. To focus on spatially proximal interactions, we filtered the PPIs for interactions between secMs and other secretory pathway-resident proteins for each secP (see Methods), resulting in a "secM support network" consisting of 3658 genes. By overlaying secM gene expression on this network, one can quantify the secM support for secretion of each secP. To systematically quantify the fitness of the secM support network for producing each secP in a tissue, a machinery support score is calculated for each secP by a random walk algorithm that integrates secM gene expression levels proximal to each secP in its PPI network. More specifically, for each secP, we added the protein to the secM support network, and centered the network on the secP. We then performed a random walk on the secM support network starting from the secP. We adapted the transition probabilities of the random walk to incorporate gene expression of the secMs so that propagation is constrained by not only PPI network topologies but also the expression of the secM components, allowing one to contextualize cell- and disease-specific interactomes (methods and supplementary note; Fig. 25A). The algorithm assigns a component score to each protein in the network, representing its availability to the secP of interest. The "secretory machinery support score" (i.e., the average component scores the secretory pathway components receive from the random walk) then quantifies the overall secretory pathway support for the given secP. Using this approach, we found the secretory machinery support score for each secP increases in tissues wherein the secP is more highly expressed (Fig. 25B). Further, the machinery support score considerably improves

the prediction of secP protein abundance from mRNA expression across tissues. Thus, by accounting for the mRNA/protein expression of PPIs surrounding each secP, the machinery support score quantifies a tissue's relative fitness for synthesizing and secreting the secP of interest.

Figure 25. Expression data can quantify a tissue or cell's fitness for synthesizing a secreted or membrane protein.
(A) In our systems biology approach, the mRNA or protein abundance is overlaid on a PPI network surrounding a secreted protein (secP). The secP synthesis fitness is quantified by summing the secM expression, scaled by distance from the secP (computed by a random walk), yielding a quantitative "machinery support score". The calculation of the support score also provides a sub-network of proteins contributing to secP synthesis. (B) We quantified the machinery support score for every secreted protein in all tissues in the Human Protein Atlas, and found a clear correlation between the secP expression and the relative machinery support score. This correlation was seen for both mRNA (top, spearman correlation coefficient, see methods) and protein (bottom, t-test) abundance. Thus, our machinery support score allows one to quantify how fit a cell or tissue is for properly expressing and processing a secreted or membrane protein.


A
secP1
σ = 0.0285
Calculation of
machinery support score for secPi
σi= 1/|secM| · Σ si = 0.0285
SecM gene
expression
0.15
0.10
0.05
0.00
Distance-
scaled
support (s)
B
membrane
secreted
R = 0.59 , ****
R = 0.52 , ****
TPM
SecP expression
protein level
High
Medium
Low
Not detected
****
***
machinery support score

Aβ deposition in Alzheimer's disease is characterized by perturbed secretory support of amyloid precursor protein

The co-regulation of the secP and their cognate secMs results from millions of years of evolution. Thus, the question arises whether perturbations to such co-regulation could underlie the molecular pathology in AD. Specifically, Aβ deposition is a major hallmark of AD-pathology. The precursor to Aβ, APP, is moderately to highly expressed (high transcript levels and moderate protein levels) in the cerebral cortex. While APP overexpression from APP duplication can cause early-onset (familial) AD [241,242], sporadic (non-familial) AD does not show differential transcript abundance for APP between AD and non-AD individuals despite the increase in Aβ plaques [243]. However, APP does undergo post-transcriptional processing, with pathogenic Aβ being released from APP following sequential cleavage by β- and γ-secretases, while the α-secretase promotes the correct processing of APP.

To test the relevance of secretory pathway expression to AD, we analyzed RNA-Seq data from 4 brain regions in 298 AD and age-matched control subjects (from the Mount Sinai Brain Bank [234]) and single-cell RNA-Seq from the prefrontal cortex (Brodmann area 10) of 48 individuals [235] (Fig. 26). APP was not differentially expressed in AD brains at both the single-cell [235] and the tissue level [234]. This is not surprising since the amyloidogenic pathway giving rise to neurotoxic Aβ takes place post-translationally [244]. Additionally, expression of neither BACE1 nor PSEN1 correlated with plaque abundance in affected brain regions (Fig. 26). However, each gene had significant changes in machinery support scores correlating with severity (Fig. 26B). Specifically, the supporting machinery score for APP decreased in affected brain regions, showing suppressed scores in cells with amyloid deposition ($p<0.0066$). The largest effect was in cell types that are major producers of Aβ, including neurons [232,233,245], reactive astrocytes [246–249] (Fig. 26A,), and in brain regions affected early in the onset of AD (Fig. 26B). However, BACE1 and PSEN1, which aid in amyloidogenesis, showed an opposite trend, with affected cells increasing machinery support for these secretases (Fig. 26B; p-values for Brodmann area 36 (parahippocampal gyrus,

BM36) and Brodmann area 44 (inferior frontal gyrus, BM44): p <0.0038 and p<0.014 for β-secretase; p<0.13 and p<0.083 for γ-secretase).

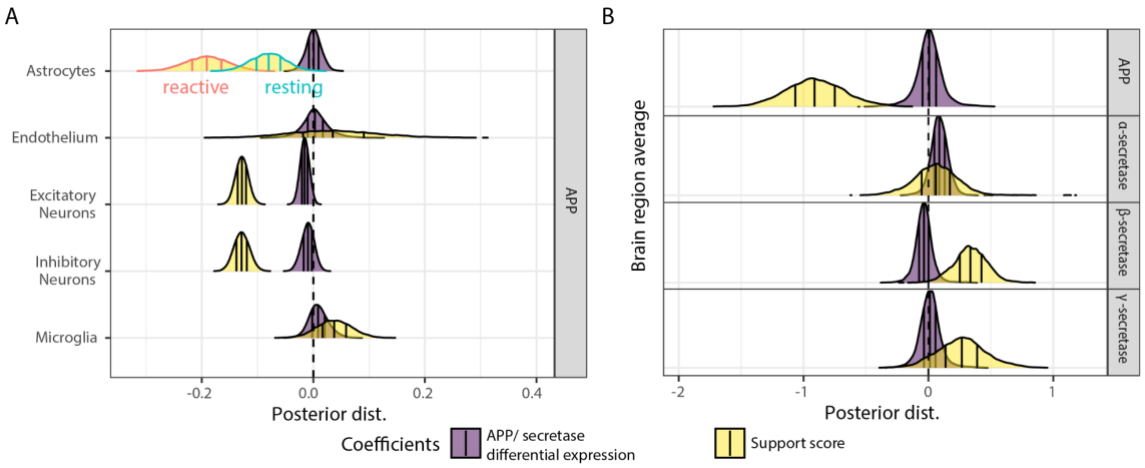


Figure 26. AD-relevant genes show perturbed secretory machinery support scores. (A) Overall, APP expression does not correlate with plaque densities across individuals in single-cell RNA-Seq. However, support scores show a negative correlation with plaque density, suggesting the secMs supporting proteostasis of APP are suppressed in AD. (B) Similar trends were seen in all brain regions surveyed (BM10, BM22, BM36, BM44, Figure S5). On average, no correlation is found between plaque abundance and gene expression of APP, BACE1 or PSEN1. However, secM support score for APP shows a negative correlation while the support scores for BACE1 and PSEN1 positively correlate with amyloid formation, suggesting AD pathogenesis involves dysregulation of the secretory pathway.

Secretory pathway support of APP is most strongly suppressed proximal to APP

The secretory machinery support score for APP is significantly decreased in AD. However, it is unclear if the decline in APP support score is due to a general suppression of many secMs throughout the secretory pathway or a local repression in which only the secMs most proximal to APP are down-regulated. To test this, we defined a core support network for APP based on network proximity and gene expression. As the abundance and proximity to APP vary across the secMs in the network, we break down the APP support score into the individual component scores for each protein in the secM support network. This quantifies each secM's corresponding contribution to the secretory support of APP. When the secMs are rank-ordered by their individual

component scores, their contribution to the APP support score follows a pattern of exponential decay, suggesting that the support score is mostly determined by a smaller number of proteins with high proximity to APP (Fig. 27A). While the entire APP support network is not differentially expressed between AD and healthy brains, we wonder if this is the case with secMs that are major contributors to the support score. When we overlaid the differential expression across brain regions and cell types and considered progressively smaller subsets of the APP support network consisting of proteins with the highest component scores, we saw the strongest repression at around 20-30 secMs, suggesting that the proteins nearest to APP are the most suppressed (Fig. 27A).

Changes in APP-supporting PPIs are regulated by AD risk loci

Large GWAS screens found AD risk genes impacting many pathways [228,250]. Although the secretory pathway is tasked with the synthesis and processing of APP, it is not generally implicated in LOAD pathogenesis, since secretory pathway genes are not enriched among LOAD risk genes from large-scale GWAS studies [228]. However, our results show transcriptional perturbations of machinery support for key amyloidogenic genes; thus, there may be a concerted regulatory change for modules supporting APP production and proteostasis in the secretory pathway. Thus, we tested if regulatory AD risk loci (i.e., transcription factors) regulate the secMs interacting with APP. While the entire APP support network does not enrich for the AD risk genes nor their regulatory targets, we wondered if AD risk loci selectively target the secMs more proximal to APP in the PPI network. As we assessed enrichment for AD risk gene targets in subnetworks that were progressively closer to APP (i.e., more focused on the support network interacting most directly with APP), we found the secMs supporting APP were increasingly enriched for targets of AD risk regulatory genes (Fig. 27B). The results are reproducible across multiple GWAS significance thresholds. Since the enrichment of AD risk gene targets steadily increased in statistical significance with increasing proximity to APP in the PPI network, we focused on the top 20 proteins interacting the most closely with APP from the support network. This resulted in an APP support subnetwork where the enrichment significance for AD risk gene targets peaks. This cutoff also coincides with the subnetwork with the strongest suppression of secM expression (Fig. 27A).

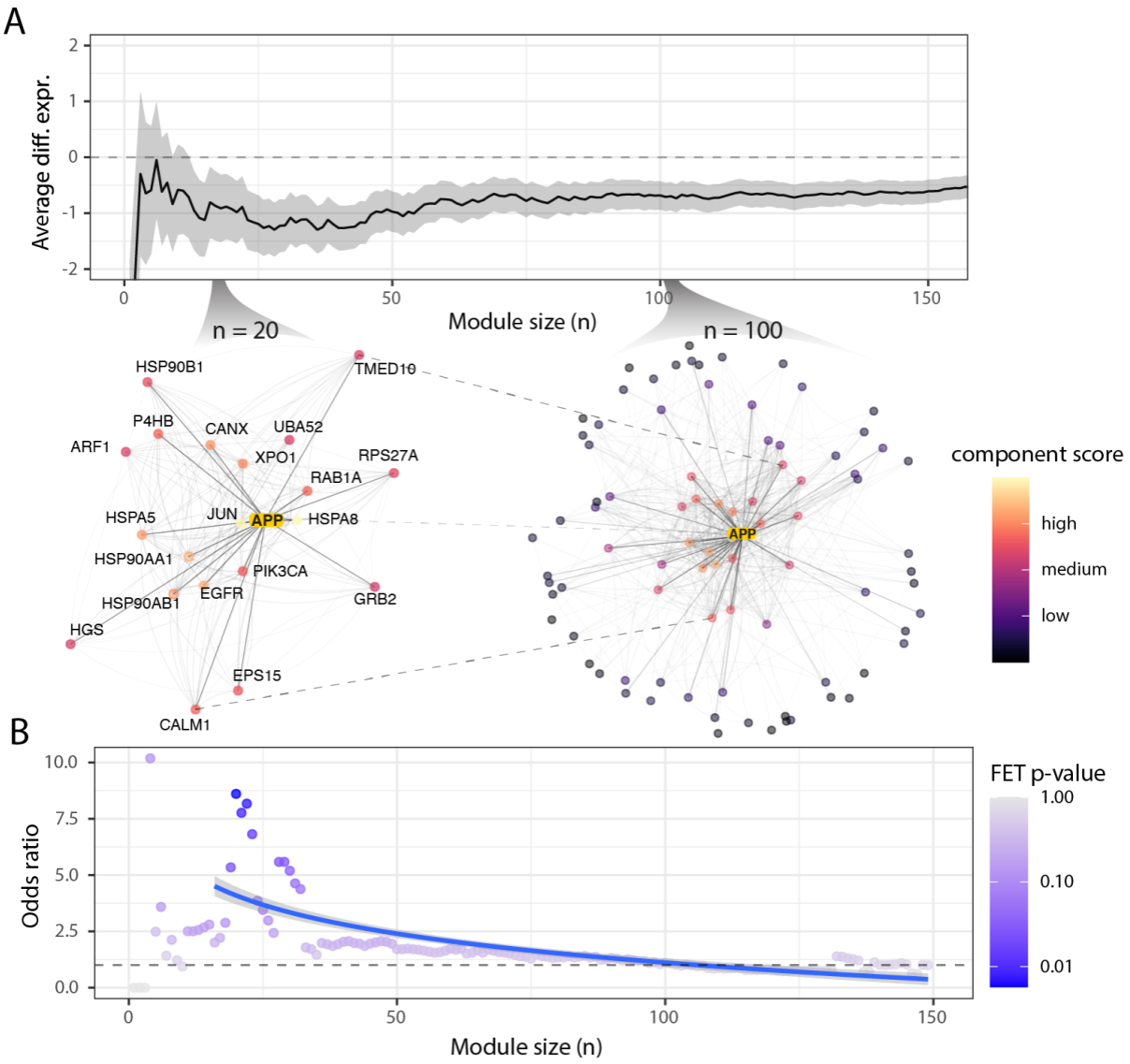


Figure 27. The APP secM proteostasis network is not enriched for AD risk genes, but is enriched for AD risk gene regulatory targets.

The networks of secM proteins supporting APP production at two support component score cutoffs, representing the top n = 20 / n = 100 proteins contributing the most to the APP's support score. The top 20 proteins with the highest component scores are labeled. (A) Starting from n=1 where the secM with the highest component score was considered, we incrementally included secMs less proximal to APP. The sizes of the subnetworks are indicated by the x-axis. At each iteration, we calculated the average differential expression (y-axis) between AD and healthy subjects for each subnetwork. The strongest repression of the AD support network occurs at around n=15~30. (B) We also calculated the degree to which the subnetwork enriched for regulatory targets of known AD risk genes above the genome-wide significance threshold (y-axis). The regulatory targets of AD risk genes are generally depleted among the general non-APP-specific secMs (Figure S9 for full trend across all 3685 subnetworks), but targets of AD risk genes enrich strongly among the core secMs closest to APP.

Core support network overlaps significantly with genomic loci with differential histone acetylation in AD brain

Epigenetic alterations have been linked to neurodegeneration in human AD brains and AD mouse models [251,252]. Thus, we analyzed data from three epigenome-wide association studies to investigate if the core support network is overrepresented for hotspots of aberrant epigenomic reprogramming in AD. We found that proteins proximal to APP on the support network show significant enrichment for Aβ related epigenetic changes measured in H3K9ac profiles from 669 aged human prefrontal cortices [253] (Fig. 28, top row). Additionally, the enrichment around the core network is higher for H3K9ac peaks annotated as being in enhancer domains than those in promoter domains. In another epigenome-wide association study comparing aging- and AD-related histone acetylation changes [254], we found the H3K122ac, H3K27ac and H3K9ac peaks that differ significantly were disproportionately located near the core support network. Interestingly, while AD and aged brains often share similar epigenetic signatures [255], enrichment of AD-related peaks is stronger in the core support network than that of aging-related peaks. Thus, we observe considerable epigenetic changes in human AD brain focused around the APP supporting network.

The enriched epigenetic changes were further captured in a mouse model of AD. Specifically, histone methylation and acetylation marks were profiled in CK-p25 mice with increased Aβ levels and controls [256]. While the core network is depleted for significantly altered H3K4me3 peaks in CK-p25 mice, AD-associated H3K27ac alterations are significantly enriched among proteins proximal to APP (Fig. 28, bottom row). This is in line with our previous observation in which the core support network is a hotspot for AD-related acetylation marks, especially around the enhancer domains.

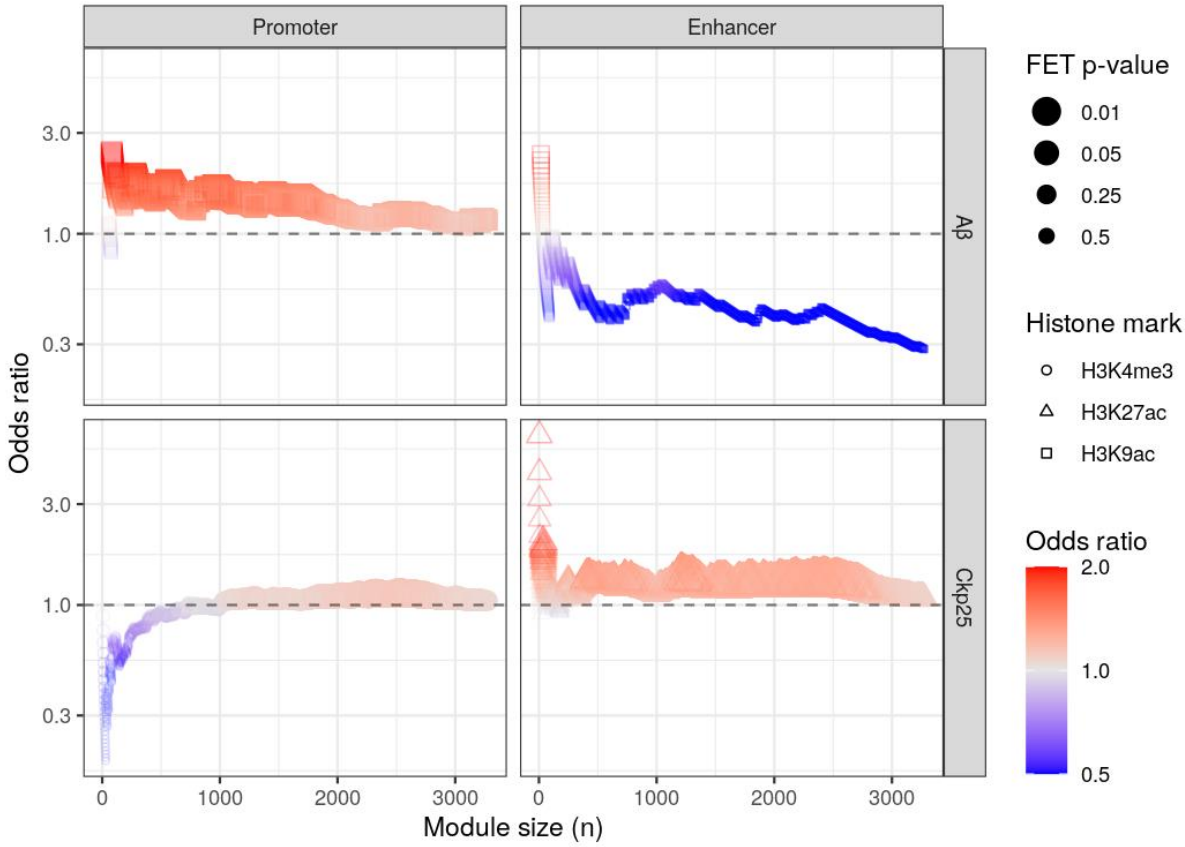


Figure 28. The APP core support network is enriched for genes whose enhancer regions contain AD-specific histone marks.
Following the notations from Figure 4, in each subplot the y-axis (odds ratio) shows the degree to which the core support network overlaps with genomic loci with differential histone modifications is indicated evaluated at modules of various sizes (x-axis), with a smaller n indicating a smaller subnetwork with only the proteins most proximal to APP. The subplots are arranged based on the region in which the differentially-enriched peaks are detected (columns), and the conditions compared (top row, Aβ-associated changes; bottom row, CK-p25 mice with increased Aβ levels vs. controls).

AD risk loci activate endocytosis via the core support network

We analyzed the content of the core support network, and found it is enriched for genes in the amyloidogenic pathway. Specifically, we saw interactions were concentrated in the ER, endosomes, and the cytosol (FDR p<1.1e-3), consistent with the localization of amyloidogenesis. For example, the endosome hosts intracellular Aβ production with its β-secretase, and is enlarged in autopsies from AD [257] and stem cell models [258]. To further unravel the link between endocytosis

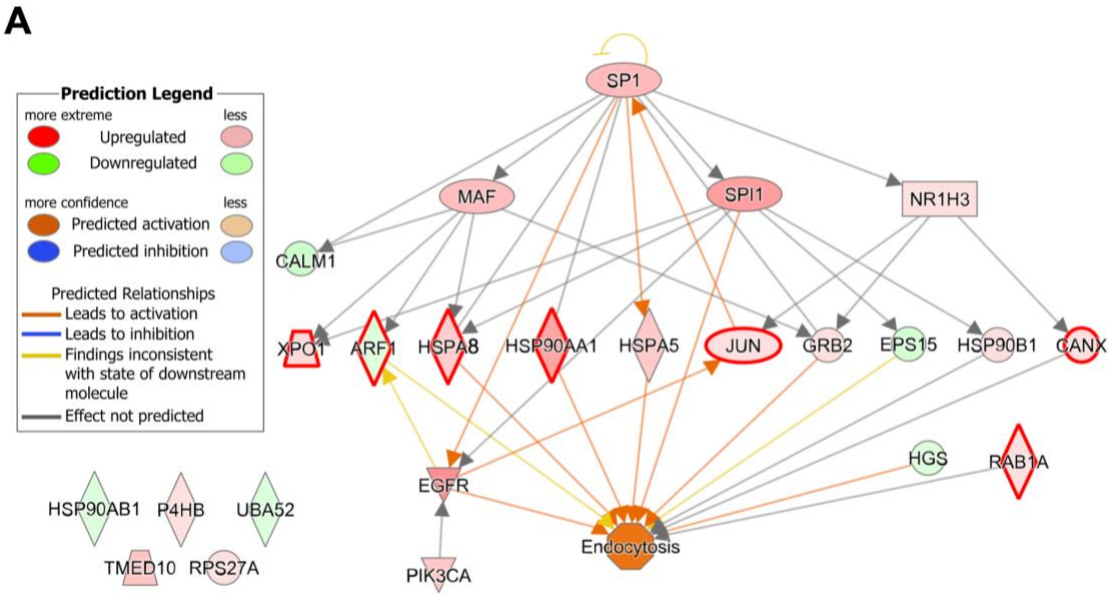
and the core support network, we analyzed the patterns of the core network differential expression between AD and controls across multiple cohort studies using gene regulatory networks obtained from ENCODE [259] and Ingenuity Pathway Analysis (IPA) [260]. Complementing our previous observation that a significant portion of the core support network is endosome-resident, we saw significant up-regulation of genes associated with endocytosis (p-value = 8.95e-14), mediated by the core supporting machinery (Fig. 29) across various brain regions.

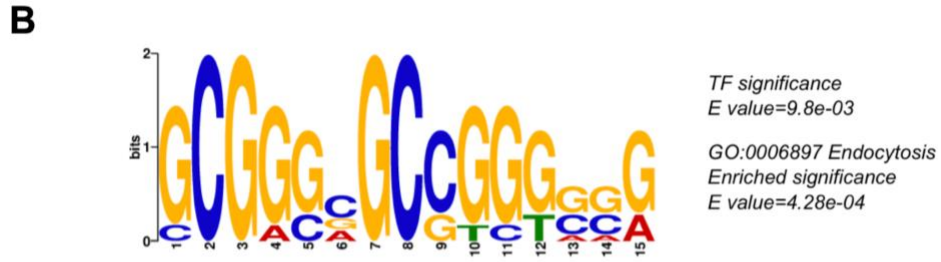
We further analyzed the APP core support network and identified transcription factors (TFs) that are most strongly associated with the perturbed secM module. The APP core support network coincides significantly with the regulatory targets of 3 genes from AD risk loci: NR1H3, MAF and SPI1 [227–229,250]. To investigate the extent to which these AD risk genes perturb transcription of the supporting machinery in AD, we analyzed the differentially expressed genes between AD and controls across multiple cohort studies using gene regulatory networks obtained from ENCODE [259] and Ingenuity pathway analysis (IPA) [260].

In addition to predicting upstream transcriptional regulators associated with amyloidogenesis from gene signatures or differentially expressed genes, we searched for conserved TF binding sites among the top-ranked dysregulated genes of the core-network in Alzheimer's Disease. Specifically, we conducted de novo TF binding site motif discovery for the 20 genes in the core support network (Fig. 29B; see Methods). We identified one significant TF motif (E value= 9.8e-3) that is present in 7 of the genes in the core support network, which is associated with the SP1, SP2, and SP3 transcription factors. SP1 is important in AD [261,262] and is predicted to bind the 3 genes from AD risk loci (NR1H3, MAF, and SPI1) with high confidence via three elite enhancers [263] GH11J047390 (GH score=2.2), GH16J079764 (GH score=2.4), and GH11J047247 (GH score=2.1) respectively. Furthermore, several motif binding sites across the core support network significantly overlap with loci with major epigenetic alterations in AD (Fig. 29B) [253–255]. For example, the motif binding site at the promoter region of ARF1 completely overlaps with a histone acetylation mark H3K122ac that is significantly altered in AD but not aged

subjects (chr1:228259654-228280877, Wilcoxon Rank Sum P-value 0.004). Another locus of significantly altered H3K122ac peaks in AD individuals (chr2:65353363-65359754, Wilcoxon Rank Sum P-value 0.01) overlaps with the motif binding site at RAB1A. The motif binding site at HSP90AA1 overlaps with significantly altered peaks for H3K122ac and H3K9ac, which are repressed and upregulated respectively in AD (chr14:102543639-102575586, Wilcoxon Rank Sum P-value 0.05 and chr14:102543639-102575586, Wilcoxon Rank Sum P-value 0.05). Thus, a perturbation to multiple TFs could disrupt the APP core support network, wherein both genetic risk genes (NR1H3, MAF, and SPI1) and global regulators can contribute to the dysregulation of the core network.

Further analysis suggests the AD risk genes and the core support network genes are further co-regulated with an activation of endocytosis in the AD pathogenesis. We further characterized the functional association of the conserved TF motif by scanning all promoters of genes in the genome for the motif (see Methods for details). The conserved TF motif was significantly enriched in the promoter regions of known endocytic pathway genes (p-value=4.28e-4) and several other pathways relevant to AD pathogenesis. Together, these results suggest a concerted change in endosomal activities and dysregulated pathways between normal and AD brains that arises due to the differential expression of the core supporting machinery surrounding APP.





Name	Align. P-value	Overlap w/ AD Epigenetic Var.	Sites
XPO1	1.58e-7		ccgcaggttg GCGGGAGCCGGTGGG ggcgggccgg
ARF1	2.82e-7	+	gcgcgtcact GCGGCCGCCGGGGCG ccgaggccag
HSPA8	3.66e-7		GGGTTTGGGC GCGGGAGCCGGGCAG GAGAAGACGG
JUN	1.31e-6		cggtg GCGGCCGCCGGTGGA tgacttcggg
P4HB	1.41e-6		aacttccggt CCGGGGGCCGGGGGG cgtgccagtt
RAB1A	1.86e-6	+	TGCCGGGGGT GCGGGGGCGGGGAGG GGGCCGGGGT
CANX	3.68e-6		AGAACCCAGC GCGAGCGCCTGGCCG GATTGGCGTG
HSP90AA1	5.44e-6	+	ttctttggca GCGGGCGCGGCGCCA gtcaggacga

Figure 29. The regulatory relationships surrounding the core support network.
(A) Regulatory structures of the 20 proteins from the core support network were constructed using IPA. Gene expression profiles from the 4 brain regions (BM10, BM22, BM36, BM44) [234], the Mayo Study [264] and the ROSMAP study (Religious Order Study and Memory and Aging Project) [265] were averaged and overlaid on the core support network. The endocytosis pathway is strongly activated in AD brains via the core support network, which in turn is regulated by 3 genes from AD risk loci-- NR1H3, MAF and SPI1. The proteins harboring binding sites for the TF motif (shown in panel B) were outlined in red. (B) The TF motif (top panel) aligns significantly to 8 genes from the entire support network (bottom panel), 7 of which belong to the core support network. Align. P-value, statistical significance of the motif alignment to the promoter region of each gene; overlap w/AD epigenetic Var., whether the motif binding site completely overlaps with significant epigenetic alterations in AD.

Discussion

In the past four decades, the Alzheimer's research community has made huge strides towards elucidating molecular processes contributing to amyloid deposition. However, despite involving aggregation of secreted proteins, it remained unclear to what extent the core processes of protein secretion and proteostasis are involved. Here we developed a systems biology approach that analyzed the interactions between key amyloidogenic components and the secretory pathway. This approach predicted the propensity for amyloid deposition at the single-cell level. To gain systems-level insights into LOAD, we used the framework to identify a subset of the secretory pathway components on which concerted suppression and several regulatory elements of AD converged. We further demonstrated that an increase in endocytic activities in LOAD can be attributed to key AD risk genes via the core support network.

LOAD is a complex disease. The identification of three rare mutations in APP, PSEN1 and PSEN2 that occur in early-onset familial AD and the discovery of APOE constitute our primary knowledge of the genetic landscape of LOAD. More recently, high-throughput technologies such as GWAS and whole exome sequencing have identified more than 45 genetic risk loci of LOAD. However, the additional risk loci exert only very small risk effects [266], and the link between genetic risk variants and amyloid deposition remains incompletely understood. Despite the highly heterogeneous expression of APP and other key amyloidogenic components across AD and healthy subjects, we demonstrated concerted down-regulation of secretory machinery proximal to APP in AD patients. This highlights the secretory pathway as a determinant of amyloid deposition, which had not been a major focus of AD research. Incidentally, the proteostasis network, with which the secretory pathway shares a significant overlap, has been an increasingly popular target of protein aggregation and aging studies. The human chaperone network, a major component of the proteostasis network, undergoes continual remodeling during an organism's lifespan [267–269]. However, in the aging AD brain, the directions of regulation of these chaperones

are rather uncoordinated across different chaperone families and even within the same family [268]. Our observations of concerted repression of key proximal secretory pathway components show that improper expression of the secretory pathway, of which the chaperone network is a subset, is associated with the deposition of amyloid.

Even though the secretory pathway is in charge of post-translational processing and targeting of APP up until its cleavage by the secretases, its implications in AD have been insufficiently researched primarily due to the lack of AD risk genes in the pathway [270]. Our results showed that genes contributing the most to the APP support network in the secretory pathway are significantly enriched for targets of AD risk genes and AD related epigenetic changes, suggesting a mechanistic link between genetic and epigenetic variants of AD and secretory pathway dysregulation, complementing previous systems-level approaches to understanding LOAD with a focus on immune- and microglia-specific modules [231]. To further unravel the link, we examined the core support network consisting of secretory pathway components most proximal to APP. We noticed 3 AD risk genes that are also transcription factors showing regulatory evidence over the core support network according to both curated and de novo pathway analyses. More importantly, we demonstrated a regulatory cascade originating from the 3 AD risk genes, mediated by the core support network and into the endocytosis pathway. The endosome, where the β-secretase is localized to and where its acidic pH is optimal for enzymatic cleavage, is a major site of intracellular Aβ production. Our findings thus offer a mechanistic view of amyloidogenesis involving the secretory pathway and the endosomes. This is in line with observations in embryonic cortical neurons that showed increased Aβ levels as a result of increased endocytic pathway activities and reuptake in APP in aged cells [271]. We also observed significant enrichment of endosomal-localized proteins in the core support network for APP, further lending credence to the involvement of the secretory pathway in activating the endocytic pathway.

The dominant model of AD pathogenesis, the amyloid hypothesis [272–274], posits that AD pathogenesis and the rest of the disease process such as tau tangle formation [275,276] result from the accumulation of Aβ via the imbalance between Aβ production and clearance. We examined the capacity of the secretory pathway in the context of Aβ production and processing, where the secretory support of APP and β- and γ-secretases were analyzed. While the concerted dysregulation of the secretory support for these key amyloidogenic components in AD brains theoretically leads to increased Aβ production, the impact on Aβ clearance warrants further investigation. It is worth noting that detectable Aβ deposition can precede the onset of AD by more than 15 years [277,278], which likely coincides with the onset of the decline of the proteostasis network. Our findings highlight the roles of the secretory pathway in amyloidogenesis, which open new possibilities for early diagnosis and treatment research on LOAD. Furthermore, our systems approach can be further applied to other diseases in which the secretory pathway is perturbed, such as perturbed hormone secretion in endocrine disorders, changes in hepatokine secretion nonalcoholic fatty liver disease [279,280], and the secretion of diverse proteins in cancer [281].

Methods

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Semi-quantitative proteomic and transcriptomic data across human tissues	The Human Protein Atlas	https://www.proteinatlas.org/humanproteome/tissue
Quantitative proteomic data across human tissues	[282]	ArrayExpress:E-MTAB-2836
Transcriptomic profiles across human tissues	[238]	https://www.gtexportal.org/home/index.html
ROSMAP scRNA-seq	[235]	Synapse:syn18485175
ROSMAP bulk RNA-seq	[234]	Synapse:syn3157743
ROSMAP clinical metadata	[265]	Synapse:syn3157322
PCNet v1.3	[283]	NDEX UUID: 4de852d9-9908-11e9-bcaf-0ac135e8bacf
Secretory pathway components	[97]	https://github.com/LewisLabUCSD/MammalianSecretoryRecon
Subcellular localization	The Cell Atlas	https://www.proteinatlas.org/humanproteome/cell
Aβ-related H3K9ac profiles	[253]	Synapse:syn4896408
Aging- and AD-related H3K122ac, H3K27ac and H3K9ac profiles	[254]	GSE153875
H3K27ac and H3K4me3 profiles from CK-p25 and control mice	[256]	GSE65159
LOAD risk loci, IGAP 2019 rare variant meta-analysis	[228]	NIAGADS:NG00075
LOAD risk loci, IGAP 2013 meta-analysis	[250]	NIAGADS:NG00036
Software and Algorithms		
Source code and interactive notebooks for expression-guided random walk	This study	https://github.com/LewisLabUCSD/AD_secretory_pathway
Code for data processing, modeling, and figure generation	This study	https://github.com/LewisLabUCSD/AD_secretory_pathway_figs
Rethinking R package	[172]	https://github.com/rmcelreath/rethinking
Ingenuity pathway analysis	[260]	http://www.ingenuity.com
MEME suite	[284]	https://meme-suite.org/meme/

Resource Availability

Materials Availability

This study did not generate new materials.

Data and Code Availability

This paper analyses existing publicly available data. These datasets' accession numbers are provided in the key resource table.

Source code and interactive notebooks for performing expression-guided random walk are publicly available at: https://github.com/LewisLabUCSD/AD_secretory_pathway

Scripts used to generate the figures presented in this paper are publicly available at https://github.com/LewisLabUCSD/AD_secretory_pathway_figs

Method Details

Calculation of secretory pathway support scores

Random walk on interactome

To contextualize the secretion of a given secP, we used network propagation to quantify the influence of gene expression across neighboring genes. Let $G(V, E)$ denote an undirected interactome with vertex set V containing n proteins and an edge set E the m interactions between them. Let w_{ij}be the edge weight ($w_{ij} = 1$ if G is undirected) between edges i and j and A be the adjacency matrix of G where $A_{ij} = \{w_{ij}\ if\ \{v_i, v_j\} \in E;\ 0\ otherwise\}$. Let $x(t)$ be the location of the walk at time t. Note that given the previous walk location i at time $t-1$, we can represent the probability of the walker moving from location i to j in a single step as (Eq. 2):

$$W_{ij} = \Pr\left(x(t) = i | x(t-1) = j\right) = \frac{A_{ij}}{d_i}, \text{ where } \quad d_i = \sum_{j}^{n} A_{ij} \quad \text{(Equation 2).}$$

Summing probabilities from all inbound locations we have (Eq. 3):

$$\Pr\left(x(t)=j|x(t-1)\right)=\sum_{i}^{n}W_{ij}\Pr\left(x(t-1)=i\right)$$ (Equation 3).

In matrix notation, this is $\mathbf{p}(t)=\mathbf{W}\mathbf{p}(t-1)$, where W is the transition matrix and each entry W_{ij} denotes the aforementioned transition probability from i to j. In random walk with restarts [285], at each step the walk resets to the origin with probability α, and the last equation becomes $\mathbf{p_{RWR}}(t)=(1-\alpha)\mathbf{W}\mathbf{p_{RWR}}(t-1)+\alpha\mathbf{p}(0)$, where $\mathbf{p}(0)=e_k$ denotes the initial distribution if the walker starts at v_k. The restart parameter α was set to 0.1, as advised by the linear optimal model given the size of the network [283].

Expression-guided random walk with restarts

The transition matrix can be modified to incorporate gene expression into each step of the propagation. If we let $\mathbf{t}=[t_0,\ldots,t_n]^\intercal$, where t_i is the scaled expression corresponding to node v_i, $0\leq t_i\leq 1$, the expression-adjusted transition matrix can thus be given by $\widehat{\mathbf{W_{adj}}}=\mathrm{diag}(\mathbf{t})\mathbf{W}$. The choice of $\mathbf{t}$ can either be protein levels where available or mRNA abundance. We show the validity of using mRNA abundance as input in the supplementary notes. We can normalize the adjusted transition matrix by adding in self-loop: $\mathbf{W_{adj}}=\widehat{\mathbf{W_{adj}}}+\mathbf{I}_{n\times n}-\mathrm{diag}(\mathbf{1}_{n\times 1}^\intercal\widehat{\mathbf{W_{adj}}})$, and the update rule, which we termed expression-guided random walk with restarts (eRWR), now becomes: $\mathbf{p_{RWR}}(t)=(1-\alpha)\mathbf{W}\mathbf{p_{RWR}}(t-1)+\alpha\mathbf{p}(0)$.

Calculation of support scores and component scores

We performed eRWR on each secP of interest for 20 iterations (Supplementary Notes), and the final vector of probabilities $\mathbf{p_{RWR}}(t=20)$ represent the support component score for each gene on the network $G(V,E)$. The support score (σ) is the average of the support component scores of the secretory pathway proteins, $\sigma=\frac{1}{|\mathrm{secM}|}\sum_{i\in\mathrm{secM}}\mathbf{p_{RWR}}_i(t=20)$.

Context filtering of the secretory pathway support network

We used the composite consensus human interactome PCNet v1.3 [283] (NDEX UUID: 4de852d9-9908-11e9-bcaf-0ac135e8bacf) to be the static, context-agnostic network ($G_0(V_0, E_0)$). For each secP, we create a subnetwork containing the secP and other essential secretory pathway genes by filtering the network for human secretory pathway components [48,97] and secretory pathway-resident proteins [286] to constrain the network spatially, resulting in a vertex-induced subgraph $G(V = \{V_0 \cap (secP \cup secM \cup secResident)\}, \{uv | uv \in E_0 \; and \; u, v \in V\})$.

Transcriptomic and proteomic data processing and support scores calculation for normal human tissues

To calculate support scores for the normal human secretome, we used two datasets, the Human Protein Atlas (HPA) [211], and the deep proteome and transcriptome abundance atlas (deep proteome) [282] for tissue-specific transcriptomes from healthy human donors in which matching proteomic data were also available. For data from the Human Protein Atlas, we downloaded the transcriptomic data-- “RNA HPA tissue gene data” and performed log- and sigmoid-transformation on the transcript abundance (TPM) data, resulting in transformed gene expression profiles in the $(0,1)$ range. For the HPA dataset, the support score was calculated based on the tissue-median of the transformed gene expression profiles for each secP. We retained the semi-quantitative nature of the immunohistochemistry protein abundance reporting, and we calculated the support scores summary statistics for proteins belonging to each of the staining levels--"High", "Medium", "Low" & "Not detected" separately.

With the fully quantitative proteomic data from the deep proteome [282], We calculated the support scores based on the protein iBAQ values. The iBAQ abundance values were transformed in a similar fashion as the transcript abundance from the HPA dataset. Namely, they were log- and then sigmoid-transformed into the $(0,1)$ range, before being median-summarized by tissue and subsequently used in the calculation of the support scores.

Transcriptomic and proteomic data processing and support scores calculation for AD and healthy brains

To calculate support scores for key amyloidogenic pathway components in AD and healthy brains, we used two major transcriptomic datasets from the ROSMAP project (Religious Orders Study and Memory Aging Project)-- single-cell [235,287] (Synapse ID syn18485175) and bulk RNA-seq [234] (Synapse ID syn3159438) data from individuals respectively with varying degrees of Alzheimer's disease pathology. The single-cell transcriptomic dataset covers 80660 cells from the prefrontal cortex of 48 individuals. While annotations for major cell types were given, we further classified astrocytes into reactive and non-reactive astrocytes based on GFAP expression [249]. The bulk RNA-seq covers 4 brain regions (Brodmann areas 10, 22, 36, 44) of 364 individuals.

To transform the count data into appropriate expression inputs to the eRWR algorithm, count data from healthy tissue-specific transcriptomes and the AD single-cell RNA-seq data were first log-transformed to compress the extreme values. The values were then z-score standardized and passed through a logistic function, where the final transformed values have a range $(0,1)$. For the AD bulk-RNA seq data, since the counts were already normalized and transformed, they were z-score standardized and transformed by a logistic function without first being log-transformed.

Statistical analysis

Relationship between support scores of secreted proteins and protein expression

To examine the dependencies between support scores and the transcript and protein abundances of the human secretome, we calculated the support score for each secreted protein in the human secretome [70] across various human tissues. We first calculated the spearman correlation coefficients between the tissue-median support scores and the transcript and protein abundances across all secreted proteins. To assess the statistical significance of the spearman correlation coefficients, a t-statistic $t = r\sqrt{\frac{n-2}{1-r^2}}$ where n and r indicate the number of paired

observations and the pearson correlation coefficient respectively was computed. P-values were then calculated by comparing the t-statistic with its null distribution (the t-statistics approximate a t-distribution with n-2 degrees of freedom under the null hypothesis) [288].

To further quantify the statistical significance of transcript abundance and protein level in determining overall protein abundance, a Bayesian hierarchical model was created (Eq. 4) where the abundance of each protein across the 29 tissues is drawn from a linear combination of mRNA levels and the support scores weighted by their respective regression coefficients. We used the rethinking R package[172] to construct the model and sample the coefficients.

$$
\begin{aligned}
\mathrm{iBAQ}_{\mathrm{Protein}}[i,j] &\sim \mathrm{Normal}(\mu[i,j],\sigma) && \text{(For each tissue-protein combo, draw protein abundance)}\\
\mu[i,j] &= \mathrm{a[i]} + b_{\mathrm{genExp}}[i]\cdot \log \mathrm{FPKM}[i,j] + b_{\mathrm{secSupport}}[i]\cdot \mathrm{support.score}[i,j] && (i:\text{Proteins } i=1\ldots13002 \text{ across tissues } j=1\ldots29)\\
\mathrm{a}[i] &\sim \mathrm{Normal}(\overline{\mathrm{p}_\mu},\overline{\mathrm{p}_\sigma}) && \text{(Protein } i \text{ across tissues drawn from population abundance)}\\
\begin{bmatrix} b_{\mathrm{genExp}}[i] \\ b_{\mathrm{secSupport}}[i] \end{bmatrix} &\sim \mathrm{Normal}(\begin{bmatrix} \overline{e_\mu} \\ \overline{s_\mu} \end{bmatrix}, \mathbf{S}) && \text{(each coef. for genExp and secSupport for Protein}_i \text{ in tissue } j \text{ drawn from 2D Normal)}\\
\mathbf{S} &= \begin{pmatrix} \overline{e_\sigma} & 0 \\ 0 & \overline{s_\sigma} \end{pmatrix} \mathbf{R} \begin{pmatrix} \overline{e_\sigma} & 0 \\ 0 & \overline{s_\sigma} \end{pmatrix}\\
\mathbf{R} &\sim \mathrm{LKJcorr}(2)\\
\overline{\mathrm{p}_\mu}, \overline{e_\mu}, \overline{s_\mu} &\sim \mathrm{Normal}(0,1)\\
\sigma, \overline{\mathrm{p}_\sigma}, \overline{e_\sigma}, \overline{s_\sigma} &\sim \mathrm{Exponential}(1)
\end{aligned}
$$

Equation 4

Relationship between support scores of key amyloidogenic proteins and amyloid plaque densities

We built a Bayesian hierarchical model (Eq. 5) to determine the extent to which the support scores for key amyloidogenic pathway components including APP and the secretases for each cell/ sample in the single-cell (see the supplementary notes for adaptations to the model formula to account for sample covariates) and bulk RNA-seq dataset affects the amount of amyloid plaque measured. We regressed the scaled amyloid plaque densities corresponding to the individual from which the single-cell/ bulk RNA-seq sample was collected against the gene expression and secretory pathway support scores of key amyloidogenic pathway components. To regularize the

coefficients of interest, their Bayesian priors are all normally distributed around 0. The coefficients were sampled using the rethinking R package[172].

$$
\begin{aligned}
&\text{amyloid}[i] \sim \text{Normal}(\mu[i], \sigma) \\
&\quad \mu[i] = \text{a} + \\
&\qquad \sum_{c \in \textbf{covar.}} b_{\text{covar}_c} \cdot c[i] + && (\textbf{covar.} : \text{known covariates}) \\
&\qquad \sum_{k \in \textbf{Amyl}} \left(b_{\text{genExp}_k}[j] \cdot \text{genExp}_k[i] + b_{\text{secSupport}\,k}[j] \cdot \text{support.score}_k[i] \right) && (j : \text{Brain region/ cell type sample } i \text{ belongs to; } \textbf{Amyl} \subseteq \{\text{APP, secretases}\}) \\
&\quad \sigma \sim \text{Exponential}(1) \\
&\quad \text{a} \sim \text{Normal}(0, 1)
\end{aligned}
$$

$$
\begin{aligned}
&\textbf{Coefficients of interest:} \\
&\qquad b_{\text{genExp}_k}[j] && \sim \text{Normal}(\overline{\beta_{\text{genExp}_k}}, \overline{\sigma_{\text{genExp}_k}}) \\
&\qquad b_{\text{secSupport}\,k}[j] && \sim \text{Normal}(\overline{\beta_{\text{secSupport}_k}}, \overline{\sigma_{\text{secSupport}_k}}) \\
&\quad \textbf{Hyper priors:} \\
&\qquad \overline{\beta_{\text{genExp}_k}}, \overline{\beta_{\text{secSupport}_k}} && \sim \text{Normal}(0, 1); \\
&\qquad \overline{\sigma_{\text{genExp}_k}}, \overline{\sigma_{\text{secSupport}_k}} && \sim \text{Exponential}(1) \\
&\quad \textbf{Covariates:} \\
&\qquad b_{\text{covar}_c} && \sim \text{Normal}(0, 1)
\end{aligned}
$$

Equation 5

Characterizing the core support network

AD risk genes and enrichment analysis of regulatory components

We obtained 45 genome-wide significant risk loci identified by several AD GWAS studies as summarized previously [227], resulting in 176 high-confidence AD risk genes. We compiled a separate set of AD risk genes from GWAS summary statistics [228,229] for loci above the genome-wide suggestive threshold, where MAGMA [289] was used to aggregate p-values for SNPs to the gene-level independently for each GWAS dataset. P-values from the two datasets for each gene were then combined using Fisher's method, resulting in 673 AD suggestive risk genes.

The transcription factors and their targets were obtained from ENCODE [290] and ChEA [291] via the Enrichr portal [292]. To determine whether the core support network enriches for the regulatory targets of AD risk genes, we first calculated the level of overlap between the core

support network and the targets of each transcription factor using Fisher's exact test, where significantly overlapping transcription factors were defined as those with p-values of less than 0.05. A secondary enrichment was performed to quantify the level to which the significant transcription factors overlap with known AD risk genes. As mentioned earlier, two lists of AD risk genes were used. For the 673 AD suggestive risk genes, a traditional Fisher's exact test was performed. For the risk genes originating from the 45 risk genome-wide significant risk loci, instead of calculating the direct overlap between the significant transcription factors and the 176 high-confidence risk genes, we mapped the significant transcription factors back to the 45 risk loci on which Fisher's exact test was performed. This is motivated by the fact that many risk loci contain multiple risk genes that cannot be further pinpointed due to complex linkage disequilibrium patterns, a risk locus is considered hit if at least one of its mapped risk genes appears significantly enriched as a transcription factor. We performed this two-stage enrichment analysis starting from the full static support network towards the core support network by pruning back proteins furthest from APP in each iteration.

Enrichment analysis of genomic loci with AD-related epigenetic changes

Genomic coordinates for AD-related histone acetylation and methylation peaks [253–256] were mapped to the promoter and enhancer regions of genes from the entire support network according to the GRCh37 assembly. Significant epigenetic alterations were defined as those with adjusted association p-values of less than 0.05, or those labeled as epigenome-wide significant in the original study when no accompanying association statistics were available. The background was defined as the collection of all peaks detected for histone marks from each study. Following a similar enrichment analysis strategy to the one detailed in the previous section, we calculated the level of overlap between the support network and the significant epigenetic alterations via Fisher's exact test across subnetworks of various sizes.

Enrichment analysis of subcellular compartments

We compiled lists of proteins for all subcellular structures consisting of proteins known to localize to the compartment of interest within the cell [286]. We ordered the proteins in the full support network by the extent to which they deviate from their stationary support component score to control for network topology while accounting for secretory-resident proteins. To determine the degree to which the proteins from certain subcellular compartments are overrepresented in the core subnetwork, we applied Gene Set Enrichment Analysis (GSEA) [127,293] with the subcellular localization gene-sets and the ranked core support network components as input, eliminating the need for a hard significance cut-off. Subcellular compartments significantly enriched in the core subnetwork are defined as those with an FDR p-value of 0.05 or less.

Causal gene network analysis

To robustly define the core supporting subnetwork, we iteratively constructed subnetworks from proteins most proximal to APP and progressively include more distal proteins corresponding to different significance cutoffs. To robustly select the cutoff for the core supporting subnetwork, we performed the two-stage enrichment analysis on all subnetworks as detailed above (see “AD risk genes and enrichment analysis of regulatory components”). Additionally, we calculated the average differential expression between AD and healthy individuals for each subnetwork using fold changes from bulk and single-cell RNA Seq data depending on the source expression from which the subnetwork is calculated. We selected 20 proteins most proximal to APP to include in the final core subnetwork, where the cutoff coincides with the strongest enrichment of regulatory AD risk loci and the suppression of the core subnetwork.

To determine the regulator effects, we performed two network-based analyses. We first ran the upstream regulator analysis using the curated regulator networks from IPA [260]. The algorithm took as inputs the core subnetwork and the differential expression fold changes and p-values. Batch-corrected differential gene expression profiles between AD and healthy brains from

the Mount Sinai study [234], the Mayo Study [264] and the ROSMAP study (Religious Order Study and Memory and Aging Project) [265] were obtained from the AMP-AD Knowledge Portal (Synapse ID syn14237651). “Disease & functions” having considerable overlap with the core subnetwork were added, of which endocytosis is the most significant (p-value =2.34E-14).

De novo TF binding site motifs discovery and known TF binding site identification

We downloaded promoter sequences (version: GRCH38) from UCSC Genome Browser[294] for the core subnetwork. The promoter sequences are defined as sequences 1,000 bases upstream of annotated transcription start sites of RefSeq genes with annotated 5' UTRs. To conduct de novo TF binding site motifs discovery, we first ran motif discovery using the MEME suite[284] with default parameters to identify candidate TF binding site motifs within the promoter sequences by using the entire APP support network serving as background control. Then, the MEME discovered TF binding site motifs were analyzed further for matches to known TF binding sites for mammalian transcription factors in the motif databases, JASPAR Vertebrates [295], via motif comparison tool, TOMTOM[296]. We summarized all the enriched GO terms using ‘Revigo’[297] (Figure S13) on the 81 GoMo identified specific enriched GO terms in the Biological Process.

Chapter 6, in full, is a reprint of the material as it appears in Kuo CC, Chiang AW, Baghdassarian HM, Lewis NE. “Dysregulation of the secretory pathway connects Alzheimer's disease genetics to aggregate formation”. *Cell Systems*, 2021. The dissertation author was the co-first investigator and author of this material.

EPILOGUE

Recapitulation

In this dissertation, I investigated the molecular machinery of the secretory pathway in various contexts, including recombinant protein production and neurodegenerative diseases. To improve titers of therapeutic recombinant proteins in mammalian host cells, I analyzed multi-omic data and identified engineering targets within the secretory pathway that were validated to significantly improve productivity.

To better understand the product-specific process of recombinant protein synthesis, I also co-developed a transient protein interaction assay based on proximity-dependent biotin identification (BioID), and connected structural properties of the recombinant proteins with their protein-protein interaction partners. To systematically analyze how changes in PPI strength impact secretion, I developed a stochastic queuing model that incorporates PPIs and enzyme expression to estimate and locate potential bottlenecks within the secretory pathway.

Finally, I turned my attention to the human secretory pathway and its co-regulation with its products. I developed a graph-based algorithm that predicts brain β-amyloid levels accurately by incorporating the topology of protein interaction networks and single-cell RNA-seq data. I further constructed Bayesian machine learning models to identify cell-types and brain regions with increased sensitivity to secretory pathway dysregulation and identi-fied cell-type specific markers in the secretory pathway responsible for elevated β-amyloid levels. Our analyses connect amyloidogenesis to dysregulation of secretory pathway components supporting APP and suggest novel therapeutic targets for AD.

Limitations and Future Directions

Advances in omics technologies have greatly accelerated the study and development of cell lines for recombinant protein production. Mechanistic studies of protein secretion have also

benefited from the progress in multi-omic profiling and editing. However, the impact of omics technologies can be greatly increased when accompanied by systems approaches to elucidate the molecular mechanisms underlying biotherapeutic production. While systems approaches often shed light on engineering targets within the cell, the resulting titer improvements reported [298,299] often remained moderate in contrast with traditional approaches such as media and bioprocess optimization, clonal selection, and bioreactor designs [300]. Advances in omic data generation and sharing would help augment current modeling techniques, allowing for more context-specific predictions and engineering target discovery. In addition, systems approaches would also benefit from more comprehensive characterization of the secretory pathway.

As omics data are integrated with genome-wide screens, phenotypic data (e.g., specific productivity and/or cell culture longevity) and systems approaches, a clearer image of the protein secretion and associated pathways will be achieved, thus allowing the generation of custom producers and an expanded list of engineering targets. Additionally, identification of the protein interaction networks supporting biosynthesis of recombinant helps define the product-specific secretion machinery, opening avenues for mammalian cell engineering efforts, wherein biotherapeutic production hosts can be rationally engineered to improve the titer of diverse proteins in a client-specific manner.

Although the results uncovered by the secretory pathway support scoring algorithm are compelling, the system has substantial room for improvement. Currently, it is assumed all PPIs between secreted protein and secretory pathway machinery components are equally positive for secretion. As characterization of the secretory pathway continues to improve, the scoring framework can be further expanded to include the type and degree of the contribution from each secretory machinery component to protein synthesis and secretion. Additionally, the same scoring algorithm can also be adapted to other pathways of interest, allowing one to explore product-specific co-regulation beyond the protein secretory pathway.

REFERENCES

1. Lewis, N. E., Liu, X., Li, Y., Nagarajan, H., Yerganian, G., O'Brien, E., Bordbar, A., Roth, A. M., Rosenbloom, J., Bian, C., Xie, M., Chen, W., Li, N., Baycin-Hizal, D., Latif, H., Forster, J., Betenbaugh, M. J., Famili, I., Xu, X., Wang, J. & Palsson, B. O. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. *Nat. Biotechnol.* **31,** 759–765 (2013).

2. Wurm, F. M. & Hacker, D. First CHO genome. *Nat. Biotechnol.* **29,** 718–720 (2011).

3. Kantardjieff, A., Jacob, N. M., Yee, J. C., Epstein, E., Kok, Y.-J., Philp, R., Betenbaugh, M. & Hu, W.-S. Transcriptome and proteome analysis of Chinese hamster ovary cells under low temperature and butyrate treatment. *J. Biotechnol.* **145,** 143–159 (2010).

4. Rita Costa, A., Elisa Rodrigues, M., Henriques, M., Azeredo, J. & Oliveira, R. Guidelines to cell engineering for monoclonal antibody production. *Eur. J. Pharm. Biopharm.* **74,** 127–138 (2010).

5. Nishimiya, D. Proteins improving recombinant antibody production in mammalian cells. *Appl. Microbiol. Biotechnol.* **98,** 1031–1042 (2014).

6. Wurm, F. M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* **22,** 1393–1398 (2004).

7. Huang, Y.-M., Hu, W., Rustandi, E., Chang, K., Yusuf-Makagiansar, H. & Ryll, T. Maximizing productivity of CHO cell-based fed-batch culture using chemically defined media conditions and typical manufacturing equipment. *Biotechnol. Prog.* **26,** 1400–1410 (2010).

8. Golabgir, A., Gutierrez, J. M., Hefzi, H., Li, S., Palsson, B. O., Herwig, C. & Lewis, N. E. Quantitative feature extraction from the Chinese hamster ovary bioprocess bibliome using a novel meta-analysis workflow. *Biotechnol. Adv.* **34,** 621–633 (2016).

9. Hacker, D. L., De Jesus, M. & Wurm, F. M. 25 years of recombinant proteins from reactor-grown cells - where do we go from here? *Biotechnol. Adv.* **27,** 1023–1027 (2009).

10. Fischer, S., Handrick, R. & Otte, K. The art of CHO cell engineering: A comprehensive retrospect and future perspectives. *Biotechnol. Adv.* **33,** 1878–1896 (2015).

11. Seth, G., Hossler, P., Yee, J. C. & Hu, W.-S. in *Advances in Biochemical Engineering/Biotechnology* 119–164 (2006).

12. Galleguillos, S. N., Ruckerbauer, D., Gerstl, M. P., Borth, N., Hanscho, M. & Zanghellini, J. What can mathematical modelling say about CHO metabolism and protein glycosylation? *Comput. Struct. Biotechnol. J.* **15,** 212–221 (2017).

13. Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V. S., Kyriakopoulos, S., Jiménez, N. E., Zielinski, D. C., Quek, L.-E., Wulff, T., Arnsdorf, J., Li, S., Lee, J. S., Paglia, G., Loira, N., Spahn, P. N., Pedersen, L. E., Gutierrez, J. M., King, Z. A., Lund, A. M., Nagarajan, H., Thomas, A., Abdel-Haleem, A. M., Zanghellini, J., Kildegaard, H. F., Voldborg, B. G., Gerdtzen, Z. P., Betenbaugh, M. J., Palsson, B. O., Andersen, M. R., Nielsen, L. K., Borth, N., Lee, D.-Y. & Lewis, N. E. A Consensus Genome-scale

Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Syst* **3,** 434–443.e8 (2016).

14. Lewis, A. M., Abu-Absi, N. R., Borys, M. C. & Li, Z. J. The use of 'Omics technology to rationally improve industrial mammalian cell line performance. *Biotechnol. Bioeng.* **113,** 26–38 (2016).

15. Kildegaard, H. F., Baycin-Hizal, D., Lewis, N. E. & Betenbaugh, M. J. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Curr. Opin. Biotechnol.* **24,** 1102–1107 (2013).

16. Lee, J. S., Grav, L. M., Lewis, N. E. & Faustrup Kildegaard, H. CRISPR/Cas9-mediated genome engineering of CHO cell factories: Application and perspectives. *Biotechnol. J.* **10,** 979–994 (2015).

17. Grav, L. M., Lee, J. S., Gerling, S., Kallehauge, T. B., Hansen, A. H., Kol, S., Lee, G. M., Pedersen, L. E. & Kildegaard, H. F. One-step generation of triple knockout CHO cell lines using CRISPR/Cas9 and fluorescent enrichment. *Biotechnol. J.* **10,** 1446–1456 (2015).

18. Reinhart, D., Damjanovic, L., Kaisermayer, C. & Kunert, R. Benchmarking of commercially available CHO cell culture media for antibody production. *Appl. Microbiol. Biotechnol.* **99,** 4645–4657 (2015).

19. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat. Biotechnol.* **33,** 1061–1072 (2015).

20. Li, J., Menzel, C., Meier, D., Zhang, C., Dübel, S. & Jostock, T. A comparative study of different vector designs for the mammalian expression of recombinant IgG antibodies. *J. Immunol. Methods* **318,** 113–124 (2007).

21. Kober, L., Zehe, C. & Bode, J. Optimized signal peptides for the development of high expressing CHO cell lines. *Biotechnol. Bioeng.* **110,** 1164–1173 (2013).

22. Hacker, D. L., De Jesus, M. & Wurm, F. M. 25 years of recombinant proteins from reactor-grown cells - where do we go from here? *Biotechnol. Adv.* **27,** 1023–1027 (2009).

23. Porter, A. J., Dickson, A. J. & Racher, A. J. Strategies for selecting recombinant CHO cell lines for cGMP manufacturing: realizing the potential in bioreactors. *Biotechnol. Prog.* **26,** 1446–1454 (2010).

24. Priola, J. J., Calzadilla, N., Baumann, M., Borth, N., Tate, C. G. & Betenbaugh, M. J. High-throughput screening and selection of mammalian cells for enhanced protein production. *Biotechnol. J.* **11,** 853–865 (2016).

25. Droz, X., Harraghy, N., Lançon, E., Le Fourn, V., Calabrese, D., Colombet, T., Liechti, P., Rida, A., Girod, P.-A. & Mermod, N. Automated microfluidic sorting of mammalian cells labeled with magnetic microparticles for those that efficiently express and secrete a protein of interest. *Biotechnol. Bioeng.* **114,** 1791–1802 (2017).

26. Carinhas, N., Oliveira, R., Alves, P. M., Carrondo, M. J. T. & Teixeira, A. P. Systems biotechnology of animal cells: the road to prediction. *Trends Biotechnol.* **30,** 377–385 (2012).

27. Richelle, A. & Lewis, N. E. Improvements in protein production in mammalian cells from

targeted metabolic engineering. *Current Opinion in Systems Biology* **6,** 1–6 (2017).

28. Jossé, L., Smales, C. M. & Tuite, M. F. Engineering the chaperone network of CHO cells for optimal recombinant protein production and authenticity. *Methods Mol. Biol.* **824,** 595–608 (2012).

29. Hansen, H. G., Pristovšek, N., Kildegaard, H. F. & Lee, G. M. Improving the secretory capacity of Chinese hamster ovary cells by ectopic expression of effector genes: Lessons learned and future directions. *Biotechnol. Adv.* **35,** 64–76 (2017).

30. Le Fourn, V., Girod, P.-A., Buceta, M., Regamey, A. & Mermod, N. CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. *Metab. Eng.* **21,** 91–102 (2014).

31. Kim, H. & Kim, J.-S. A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.* **15,** 321–334 (2014).

32. Yang, Z., Wang, S., Halim, A., Schulz, M. A., Frodin, M., Rahman, S. H., Vester-Christensen, M. B., Behrens, C., Kristensen, C., Vakhrushev, S. Y., Bennett, E. P., Wandall, H. H. & Clausen, H. Engineered CHO cells for production of diverse, homogeneous glycoproteins. *Nat. Biotechnol.* **33,** 842–844 (2015).

33. Wang, Q., Yin, B., Chung, C.-Y. & Betenbaugh, M. J. Glycoengineering of CHO Cells to Improve Product Quality. *Methods Mol. Biol.* **1603,** 25–44 (2017).

34. Chavez, A., Tuttle, M., Pruitt, B. W., Ewen-Campen, B., Chari, R., Ter-Ovanesyan, D., Haque, S. J., Cecchi, R. J., Kowal, E. J. K., Buchthal, J., Housden, B. E., Perrimon, N., Collins, J. J. & Church, G. Comparison of Cas9 activators in multiple species. *Nat. Methods* **13,** 563–567 (2016).

35. Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E. & Gersbach, C. A. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33,** 510–517 (2015).

36. Orellana, C. A., Marcellin, E., Schulz, B. L., Nouwens, A. S., Gray, P. P. & Nielsen, L. K. High-antibody-producing Chinese hamster ovary cells up-regulate intracellular protein transport and glutathione synthesis. *J. Proteome Res.* **14,** 609–618 (2015).

37. Vishwanathan, N., Yongky, A., Johnson, K. C., Fu, H.-Y., Jacob, N. M., Le, H., Yusufi, F. N. K., Lee, D. Y. & Hu, W.-S. Global insights into the Chinese hamster and CHO cell transcriptomes. *Biotechnol. Bioeng.* **112,** 965–976 (2015).

38. Kallehauge, T. B., Li, S., Pedersen, L. E., Ha, T. K., Ley, D., Andersen, M. R., Kildegaard, H. F., Lee, G. M. & Lewis, N. E. Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion. *Sci. Rep.* **7,** 40388 (2017).

39. Godfrey, C. L., Mead, E. J., Daramola, O., Dunn, S., Hatton, D., Field, R., Pettman, G. & Smales, C. M. Polysome profiling of mAb producing CHO cell lines links translational control of cell proliferation and recombinant mRNA loading onto ribosomes with global and recombinant protein synthesis. *Biotechnol. J.* **12,** (2017).

40. Mohmad-Saberi, S. E., Hashim, Y. Z. H.-Y., Mel, M., Amid, A., Ahmad-Raus, R. & Packeer-Mohamed, V. Metabolomics profiling of extracellular metabolites in CHO-K1 cells cultured in

different types of growth media. *Cytotechnology* **65,** 577–586 (2013).

41. Sellick, C. A., Croxford, A. S., Maqsood, A. R., Stephens, G. M., Westerhoff, H. V., Goodacre, R. & Dickson, A. J. Metabolite profiling of CHO cells: Molecular reflections of bioprocessing effectiveness. *Biotechnol. J.* **10,** 1434–1445 (2015).

42. Dietmair, S., Hodson, M. P., Quek, L.-E., Timmins, N. E., Chrysanthopoulos, P., Jacob, S. S., Gray, P. & Nielsen, L. K. Metabolite profiling of CHO cells with different growth characteristics. *Biotechnol. Bioeng.* **109,** 1404–1414 (2012).

43. Mulukutla, B. C., Kale, J., Kalomeris, T., Jacobs, M. & Hiller, G. W. Identification and control of novel growth inhibitors in fed-batch cultures of Chinese hamster ovary cells. *Biotechnol. Bioeng.* **114,** 1779–1790 (2017).

44. Dinnis, D. M. & James, D. C. Engineering mammalian cell factories for improved recombinant monoclonal antibody production: lessons from nature? *Biotechnol. Bioeng.* **91,** 180–189 (2005).

45. Jonikas, M. C., Collins, S. R., Denic, V., Oh, E., Quan, E. M., Schmid, V., Weibezahn, J., Schwappach, B., Walter, P., Weissman, J. S. & Schuldiner, M. Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science* **323,** 1693–1697 (2009).

46. Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A. & Weissman, J. S. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167,** 1867–1882.e21 (2016).

47. Lund, A. M., Kaas, C. S., Brandl, J., Pedersen, L. E., Kildegaard, H. F., Kristensen, C. & Andersen, M. R. Network reconstruction of the mouse secretory pathway applied on CHO cell transcriptome data. *BMC Syst. Biol.* **11,** 37 (2017).

48. Feizi, A., Gatto, F., Uhlen, M. & Nielsen, J. Human protein secretory pathway genes are expressed in a tissue-specific pattern to match processing demands of the secretome. *NPJ Syst Biol Appl* **3,** 22 (2017).

49. Chiang, A. W., Li, S., Spahn, P. N., Richelle, A., Kuo, C.-C., Samoudi, M. & Lewis, N. E. Modulating carbohydrate-protein interactions through glycoengineering of monoclonal antibodies to impact cancer physiology. *Curr. Opin. Struct. Biol.* **40,** 104–111 (2016).

50. Zhang, L., Luo, S. & Zhang, B. The use of lectin microarray for assessing glycosylation of therapeutic proteins. *MAbs* **8,** 524–535 (2016).

51. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161,** 971–987 (2015).

52. Yusufi, F. N. K., Lakshmanan, M., Ho, Y. S., Loo, B. L. W., Ariyaratne, P., Yang, Y., Ng, S. K., Tan, T. R. M., Yeo, H. C., Lim, H. L., Ng, S. W., Hiu, A. P., Chow, C. P., Wan, C., Chen, S., Teo, G., Song, G., Chin, J. X., Ruan, X., Sung, K. W. K., Hu, W.-S., Yap, M. G. S., Bardor, M., Nagarajan, N. & Lee, D.-Y. Mammalian Systems Biotechnology Reveals Global Cellular Adaptations in a Recombinant CHO Cell Line. *Cell Syst* **4,** 530–542.e6 (2017).

53. Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J. & Jirstrand, M. Kinetic models in industrial biotechnology – Improving cell factory performance. *Metab. Eng.* **24,** 38–60 (2014).

54. Spahn, P. N. & Lewis, N. E. Systems glycobiology for glycoengineering. *Curr. Opin. Biotechnol.* **30,** 218–224 (2014).

55. Krambeck, F. J., Bennun, S. V., Andersen, M. R. & Betenbaugh, M. J. Model-based analysis of N-glycosylation in Chinese hamster ovary cells. *PLoS One* **12,** e0175376 (2017).

56. Spahn, P. N., Hansen, A. H., Hansen, H. G., Arnsdorf, J., Kildegaard, H. F. & Lewis, N. E. A Markov chain model for N-linked protein glycosylation--towards a low-parameter tool for model-driven glycoengineering. *Metab. Eng.* **33,** 52–66 (2016).

57. Spahn, P. N., Hansen, A. H., Kol, S., Voldborg, B. G. & Lewis, N. E. Predictive glycoengineering of biosimilars using a Markov chain glycosylation model. *Biotechnol. J.* **12,** (2017).

58. Jimenez Del Val, I., Fan, Y. & Weilguny, D. Dynamics of immature mAb glycoform secretion during CHO cell culture: An integrated modelling framework. *Biotechnol. J.* **11,** 610–623 (2016).

59. Clarke, C., Doolan, P., Barron, N., Meleady, P., O'Sullivan, F., Gammell, P., Melville, M., Leonard, M. & Clynes, M. Predicting cell-specific productivity from CHO gene expression. *J. Biotechnol.* **151,** 159–165 (2011).

60. Pybus, L. P., James, D. C., Dean, G., Slidel, T., Hardman, C., Smith, A., Daramola, O. & Field, R. Predicting the expression of recombinant monoclonal antibodies in Chinese hamster ovary cells based on sequence features of the CDR3 domain. *Biotechnol. Prog.* **30,** 188–197 (2014).

61. Le, H., Kabbur, S., Pollastrini, L., Sun, Z., Mills, K., Johnson, K., Karypis, G. & Hu, W.-S. Multivariate analysis of cell culture bioprocess data--lactate consumption as process indicator. *J. Biotechnol.* **162,** 210–223 (2012).

62. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning* (eds. Hastie, T., Tibshirani, R. & Friedman, J.) 649–698 (Springer New York, 2009).

63. Gerstl, M. P., Hanscho, M., Ruckerbauer, D. E., Zanghellini, J. & Borth, N. CHOmine: an integrated data warehouse for CHO systems biology and modeling. *Database* **2017,** (2017).

64. Kremkow, B. G., Baik, J. Y., MacDonald, M. L. & Lee, K. H. CHOgenome.org 2.0: Genome resources and website updates. *Biotechnol. J.* **10,** 931–938 (2015).

65. Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D. C. & Lewis, N. E. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Syst* **4,** 318–329.e6 (2017).

66. Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A., Kuo, C.-C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., Qi, L., Ideker, T. & Mali, P. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions.

Nat. Methods **14,** 573–576 (2017).

67. Kim, J. Y., Kim, Y.-G. & Lee, G. M. CHO cells in biotechnology for production of recombinant proteins : current state and further potential. 917–930 (2012).

68. Walsh, G. Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.* **36,** 1136–1145 (2018).

69. Tegel, H., Dannemeyer, M., Kanje, S., Sivertsson, Å., Berling, A., Svensson, A.-S., Hober, A., Enstedt, H., Volk, A.-L., Lundqvist, M., Moradi, M., Afshari, D., Ekblad, S., Xu, L., Westin, M., Bidad, F., Schiavone, L. H., Davies, R., Mayr, L. M., Knight, S., Göpel, S. O., Voldborg, B. G., Edfors, F., Forsström, B., von Feilitzen, K., Zwahlen, M., Rockberg, J., Takanen, J. O., Uhlén, M. & Hober, S. High throughput generation of a resource of the human secretome in mammalian cells. *N. Biotechnol.* **58,** 45–54 (2020).

70. Uhlén, M., Karlsson, M. J., Hober, A., Svensson, A.-S., Scheffel, J., Kotol, D., Zhong, W., Tebani, A., Strandberg, L., Edfors, F., Sjöstedt, E., Mulder, J., Mardinoglu, A., Berling, A., Ekblad, S., Dannemeyer, M., Kanje, S., Rockberg, J., Lundqvist, M., Malm, M., Volk, A.-L., Nilsson, P., Månberg, A., Dodig-Crnkovic, T., Pin, E., Zwahlen, M., Oksvold, P., Feilitzen, K. von, Häussler, R. S., Hong, M.-G., Lindskog, C., Ponten, F., Katona, B., Vuu, J., Lindström, E., Nielsen, J., Robinson, J., Ayoglu, B., Mahdessian, D., Sullivan, D., Thul, P., Danielsson, F., Stadler, C., Lundberg, E., Bergström, G., Gummesson, A., Voldborg, B. G., Tegel, H., Hober, S., Forsström, B., Schwenk, J. M., Fagerberg, L. & Sivertsson, Å. The human secretome. *Sci. Signal.* **12,** (2019).

71. Le Fourn, V., Girod, P.-A., Buceta, M., Regamey, A. & Mermod, N. CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. *Metab. Eng.* **21,** 91–102 (2014).

72. Jiang, H. & Zhu, Z. in *Update on Production of Recombinant Therapeutic Protein: Transient Gene Expression* 17–79 (Smithers Rapra Technology Ltd, 2013).

73. Pham, P. L., Kamen, A. & Durocher, Y. Large-Scale Transfection of Mammalian Cells for the Fast Production of Recombinant Protein. *Mol. Biotechnol.* **34,** 225–238 (2006).

74. Böhm, E., Seyfried, B. K., Dockal, M., Graninger, M., Hasslacher, M., Neurath, M., Konetschny, C., Matthiessen, P., Mitterer, A. & Scheiflinger, F. Differences in N-glycosylation of recombinant human coagulation factor VII derived from BHK, CHO, and HEK293 cells. *BMC Biotechnol.* **15,** 87 (2015).

75. Croset, A., Delafosse, L., Gaudry, J.-P., Arod, C., Glez, L., Losberger, C., Begue, D., Krstanovic, A., Robert, F., Vilbois, F., Chevalet, L. & Antonsson, B. Differences in the glycosylation of recombinant proteins expressed in HEK and CHO cells. *J. Biotechnol.* **161,** 336–348 (2012).

76. Goh, J. B. & Ng, S. K. Impact of host cell line choice on glycan profile. *Crit. Rev. Biotechnol.* **38,** 851–867 (2018).

77. Dumont, J., Euwart, D., Mei, B., Estes, S. & Kshirsagar, R. Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. *Crit. Rev. Biotechnol.* **36,** 1110–1122 (2016).

78. Silla, T., Ha, I., Geimanen, J., Janikson, K., Abroi, A., Ustav, E. & Ustav, M. Episomal Maintenance of Plasmids with Hybrid Origins in Mouse Cells. **79,** 15277–15288 (2005).

79. Edfors, F., Boström, T., Forsström, B., Zeiler, M., Johansson, H., Lundberg, E., Hober, S., Lehtiö, J., Mann, M. & Uhlen, M. Immunoproteomics Using Polyclonal Antibodies and Stable Isotope–labeled Affinity-purified Recombinant Proteins. *Mol. Cell. Proteomics* **13,** 1611–1624 (2014).

80. Lasunskaia, E. B., Fridlianskaia, I. I., Darieva, Z. A., Da Silva, M. S. R., Kanashiro, M. M. & Margulis, B. A. Transfection of NS0 myeloma fusion partner cells with HSP70 gene results in higher hybridoma yield by improving cellular resistance to apoptosis. *Biotechnol. Bioeng.* **81,** 496–504 (2003).

81. Meleady, P., Doolan, P., Henry, M., Barron, N., Keenan, J., O'Sullivan, F., Clarke, C., Gammell, P., Melville, M. W., Leonard, M. & Clynes, M. Sustained productivity in recombinant Chinese hamster ovary (CHO) cell lines: proteome analysis of the molecular basis for a process-related phenotype. *BMC Biotechnol.* **11,** 78 (2011).

82. Ishaque, A., Thrift, J., Murphy, J. E. & Konstantinov, K. Over-expression of Hsp70 in BHK-21 cells engineered to produce recombinant factor VIII promotes resistance to apoptosis and enhances secretion. *Biotechnol. Bioeng.* **97,** 144–155 (2007).

83. Lee, Y. Y., Wong, K. T. K., Tan, J., Toh, P. C., Mao, Y., Brusic, V. & Yap, M. G. S. Overexpression of heat shock proteins (HSPs) in CHO cells for extended culture viability and improved recombinant protein production. *J. Biotechnol.* **143,** 34–43 (2009).

84. Orellana, C. A., Marcellin, E., Palfreyman, R. W., Munro, T. P., Gray, P. P. & Nielsen, L. K. RNA-Seq Highlights High Clonal Variation in Monoclonal Antibody Producing CHO Cells. *Biotechnol. J.* **13,** 1700231 (2018).

85. Sommeregger, W., Mayrhofer, P., Steinfellner, W., Reinhart, D., Henry, M., Clynes, M., Meleady, P. & Kunert, R. Proteomic differences in recombinant CHO cells producing two similar antibody fragments. *Biotechnol. Bioeng.* **113,** 1902–1912 (2016).

86. Ohya, T., Hayashi, T., Kiyama, E., Nishii, H., Miki, H., Kobayashi, K., Honda, K., Omasa, T. & Ohtake, H. Improved production of recombinant human antithrombin III in Chinese hamster ovary cells by ATF4 overexpression. *Biotechnol. Bioeng.* **100,** 317–324 (2008).

87. Haredy, A. M., Nishizawa, A., Honda, K., Ohya, T., Ohtake, H. & Omasa, T. Improved antibody production in Chinese hamster ovary cells by ATF4 overexpression. *Cytotechnology* **65,** 993–1002 (2013).

88. Hwang, S. O., Chung, J. Y. & Lee, G. M. Effect of doxycycline-regulated ERp57 expression on specific thrombopoietin productivity of recombinant CHO cells. *Biotechnol. Prog.* **19,** (2003).

89. Hansen, H. G., Nilsson, C. N., Lund, A. M., Kol, S., Grav, L. M., Lundqvist, M., Rockberg, J., Lee, G. M., Andersen, M. R. & Kildegaard, H. F. Versatile microscale screening platform for improving recombinant protein productivity in Chinese hamster ovary cells. *Sci. Rep.* **5,** 18016 (2015).

90. Feichtinger, J., Hernández, I., Fischer, C., Hanscho, M., Auer, N., Hackl, M., Jadhav, V., Baumann, M., Krempl, P. M., Schmidl, C., Farlik, M., Schuster, M., Merkel, A., Sommer, A., Heath, S., Rico, D., Bock, C., Thallinger, G. G. & Borth, N. Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. *Biotechnol. Bioeng.* **113,** 2241–2253 (2016).

91. Lin, Y.-C., Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., Reumers, J., Moisse, M., Plaisance, S., Drmanac, R., Chen, J., Speleman, F., Lambrechts, D., Van de Peer, Y., Tavernier, J. & Callewaert, N. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat. Commun.* **5,** 4767 (2014).

92. Stepanenko, A. A. & Dmitrenko, V. V. HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene* **569,** 182–190 (2015).

93. Vcelar, S., Jadhav, V., Melcher, M., Auer, N., Hrdina, A., Sagmeister, R., Heffner, K., Puklowski, A., Betenbaugh, M., Wenger, T., Leisch, F., Baumann, M. & Borth, N. Karyotype variation of CHO host cell lines over time in culture characterized by chromosome counting and chromosome painting. *Biotechnol. Bioeng.* **115,** 165–173 (2018).

94. Wurm, F. M. CHO Quasispecies—Implications for Manufacturing Processes. *Processes* **1,** 296–311 (2013).

95. Malm, M., Saghaleyni, R., Lundqvist, M., Giudici, M., Chotteau, V., Field, R., Varley, P. G., Hatton, D., Grassi, L., Svensson, T., Nielsen, J. & Rockberg, J. Evolution from adherent to suspension: systems biology of HEK293 cell line development. *Sci. Rep.* **10,** (2020).

96. Jäger, V., Büssow, K. & Schirrmann, T. in *Animal Cell Culture* (ed. Al-Rubeai, M.) 27–64 (Springer, Cham, 2015).

97. Gutierrez, J. M., Feizi, A., Li, S., Kallehauge, T. B., Hefzi, H., Grav, L. M., Ley, D., Baycin Hizal, D., Betenbaugh, M. J., Voldborg, B., Faustrup Kildegaard, H., Min Lee, G., Palsson, B. O., Nielsen, J. & Lewis, N. E. Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. *Nat. Commun.* **11,** 68 (2020).

98. Hussain, H., Maldonado-Agurto, R. & Dickson, A. J. The endoplasmic reticulum and unfolded protein response in the control of mammalian recombinant protein production. *Biotechnol. Lett.* **36,** 1581–1593 (2014).

99. Prashad, K. & Mehra, S. Dynamics of unfolded protein response in recombinant CHO cells. *Cytotechnology* **67,** 237–254 (2015).

100. Ikawa, M., Wada, I., Kominami, K., Watanabe, D., Toshimori, K., Nishimune, Y. & Okabe, M. The putative chaperone calmegin is required for sperm fertility. *Nature* **387,** 607–611 (1997).

101. Leung-Hagesteijn, C., Erdmann, N., Cheung, G., Keats, J. J., Stewart, A. K., Reece, D. E., Chung, K. C. & Tiedemann, R. E. Xbp1s-Negative Tumor B Cells and Pre-Plasmablasts Mediate Therapeutic Proteasome Inhibitor Resistance in Multiple Myeloma. *Cancer Cell* **24,** 289–304 (2013).

102. Sheng, J., Flick, H. & Feng, X. Systematic Optimization of Protein Secretory Pathways in Saccharomyces cerevisiae to Increase Expression of Hepatitis B Small Antigen. *Front. Microbiol.* **8,** 875 (2017).

103. Roth, R. A. & Koshland, M. E. Role of disulfide interchange enzyme in immunoglobulin synthesis. *Biochemistry* **20,** 6594–6599 (1981).

104. Graf, M., Deml, L. & Wagner, R. Codon-optimized genes that enable increased heterologous expression in mammalian cells and elicit efficient immune responses in mice after vaccination of naked DNA. *Methods Mol. Med.* **94,** 197–210 (2004).

105. Hung, F., Deng, L., Ravnikar, P., Condon, R., Li, B., Do, L., Saha, D., Tsao, Y.-S., Merchant, A., Liu, Z. & Shi, S. mRNA stability and antibody production in CHO cells: Improvement through gene optimization. *Biotechnol. J.* **5,** 393–401 (2010).

106. Scholten, K. B. J., Kramer, D., Kueter, E. W. M., Graf, M., Schoedl, T., Meijer, C. J. L. M., Schreurs, M. W. J. & Hooijberg, E. Codon modification of T cell receptors allows enhanced functional expression in transgenic human T cells. *Clin. Immunol.* **119,** 135–145 (2006).

107. Dalton, A. C. & Barton, W. A. Over-expression of secreted proteins from mammalian cell lines. *Protein Sci.* **23,** 517–525 (2014).

108. Güler-Gane, G., Kidd, S., Sridharan, S., Vaughan, T. J., Wilkinson, T. C. I. & Tigue, N. J. Overcoming the Refractory Expression of Secreted Recombinant Proteins in Mammalian Cells through Modification of the Signal Peptide and Adjacent Amino Acids. *PLoS One* **11,** e0155340 (2016).

109. Dorai, H., Santiago, A., Campbell, M., Tang, Q. M., Lewis, M. J., Wang, Y., Lu, Q.-Z., Wu, S.-L. & Hancock, W. Characterization of the proteases involved in the N-terminal clipping of glucagon-like-peptide-1-antibody fusion proteins. *Biotechnol. Prog.* **27,** 220–231 (2011).

110. Gao, S. X., Zhang, Y., Stansberry-Perkins, K., Buko, A., Bai, S., Nguyen, V. & Brader, M. L. Fragmentation of a highly purified monoclonal antibody attributed to residual CHO cell protease activity. *Biotechnol. Bioeng.* **108,** 977–982 (2011).

111. Goldman, M. H., James, D. C., Ison, A. P. & Bull, A. T. Monitoring proteolysis of recombinant human interferon-gamma during batch culture of Chinese hamster ovary cells. *Cytotechnology* **23,** 103–111 (1997).

112. Saibil, H. Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.* **14,** 630–642 (2013).

113. Butz, J. A., Niebauer, R. T. & Robinson, A. S. Co-expression of molecular chaperones does not improve the heterologous expression of mammalian G-protein coupled receptor expression in yeast. *Biotechnol. Bioeng.* **84,** 292–304 (2003).

114. Yoshida, Y., Adachi, E., Fukiya, K., Iwai, K. & Tanaka, K. Glycoprotein-specific ubiquitin ligases recognize N-glycans in unfolded substrates. *EMBO Rep.* **6,** 239–244 (2005).

115. Chapple, S. D. J., Crofts, A. M., Shadbolt, S. P., McCafferty, J. & Dyson, M. R. Multiplexed expression and screening for recombinant protein production in mammalian cells. *BMC Biotechnol.* **6,** 49 (2006).

116. Volk, A.-L. ;., Hu, F. J., Berglund, M. M., Nordling, E. ;., Strömberg, P. ;., Uhlén, M. ;. & Rockberg, J. Stratification of responders towards eculizumab using a structural epitope mapping strategy. *Citation* (2016). doi:10.1038/srep31365

117. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results

for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047–3048 (2016).

118. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

119. Rupp, O., MacDonald, M. L., Li, S., Dhiman, H., Polson, S., Griep, S., Heffner, K., Hernandez, I., Brinkrolf, K., Jadhav, V., Samoudi, M., Hao, H., Kingham, B., Goesmann, A., Betenbaugh, M. J., Lewis, N. E., Borth, N. & Lee, K. H. A reference genome of the Chinese hamster based on a hybrid assembly strategy. *Biotechnol. Bioeng.* **115,** 2087–2100 (2018).

120. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14,** 417–419 (2017).

121. Liebermeister, W., Noor, E., Flamholz, A., Davidi, D., Bernhardt, J. & Milo, R. Visual account of protein investment in cellular functions. *Proceedings of the National Academy of Sciences* **111,** 8488–8493 (2014).

122. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

123. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4,** 1521 (2016).

124. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35,** 2084–2092 (2019).

125. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

126. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P. & Mesirov, J. P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).

127. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

128. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N. & Sergushichev, A. Fast gene set enrichment analysis. *bioRxiv* 060012 (2021).

129. Narimatsu, Y., Joshi, H. J., Nason, R., Van Coillie, J., Karlsson, R., Sun, L., Ye, Z., Chen, Y.-H., Schjoldager, K. T., Steentoft, C., Furukawa, S., Bensing, B. A., Sullam, P. M., Thompson, A. J., Paulson, J. C., Büll, C., Adema, G. J., Mandel, U., Hansen, L., Bennett, E. P., Varki, A., Vakhrushev, S. Y., Yang, Z. & Clausen, H. An Atlas of Human Glycosylation Pathways Enables Display of the Human Glycome by Gene Engineered Cells. *Mol. Cell* **75,** 394–407.e5 (2019).

130. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

131. McElreath, R. *Statistical Rethinking*. (Chapman and Hall/CRC, 2020).

132. Uhlen, M., Tegel, H., Sivertsson, Å., Kuo, C.-C., Gutierrez, J. M., Lewis, N. E., Forsström, B., Dannemeyer, M., Fagerberg, L., Malm, M., Vunk, H., Edfors, F., Hober, A., Sjöstedt, E., Kotol, D., Mulder, J., Mardinoglu, A., Schwenk, J. M., Nilsson, P., Zwahlen, M., Takanen, J. O., von Feilitzen, K., Stadler, C., Lindskog, C., Ponten, F., Nielsen, J., Palsson, B. O., Volk, A.-L., Lundqvist, M., Berling, A., Svensson, A.-S., Kanje, S., Enstedt, H., Afshari, D., Ekblad, S., Scheffel, J., Katona, B., Vuu, J., Lindström, E., Xu, L., Mihai, R., Bremer, L., Westin, M., Muse, M., Mayr, L. M., Knight, S., Göpel, S., Davies, R., Varley, P., Hatton, D., Fields, R., Voldborg, B. G., Rockberg, J., Schiavone, L. H. & Hober, S. The human secretome – the proteins secreted from human cells. *bioRxiv* (2018). doi:10.1101/465815

133. Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., Oliver-Williams, C., Kamat, M. A., Prins, B. P., Wilcox, S. K., Zimmerman, E. S., Chi, A., Bansal, N., Spain, S. L., Wood, A. M., Morrell, N. W., Bradley, J. R., Janjic, N., Roberts, D. J., Ouwehand, W. H., Todd, J. A., Soranzo, N., Suhre, K., Paul, D. S., Fox, C. S., Plenge, R. M., Danesh, J., Runz, H. & Butterworth, A. S. Genomic atlas of the human plasma proteome. *Nature* **558,** 73–79 (2018).

134. Uhlen, M., Karlsson, M. J., Zhong, W., Tebani, A., Pou, C., Mikes, J., Lakshmikanth, T., Forsström, B., Edfors, F., Odeberg, J., Mardinoglu, A., Zhang, C., von Feilitzen, K., Mulder, J., Sjöstedt, E., Hober, A., Oksvold, P., Zwahlen, M., Ponten, F., Lindskog, C., Sivertsson, Å., Fagerberg, L. & Brodin, P. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366,** (2019).

135. Butler, M. & Spearman, M. The choice of mammalian cell host and possibilities for glycosylation engineering. *Curr. Opin. Biotechnol.* **30,** 107–112 (2014).

136. Kunert, R. & Reinhart, D. Advances in recombinant antibody manufacturing. *Appl. Microbiol. Biotechnol.* **100,** 3451–3461 (2016).

137. Tegel, H., Dannemeyer, M., Kanje, S., Sivertsson, Å., Berling, A., Svensson, A.-S., Hober, A., Enstedt, H., Volk, A.-L., Lundqvist, M., Moradi, M., Afshari, D., Ekblad, S., Xu, L., Westin, M., Bidad, F., Schiavone, L. H., Davies, R., Mayr, L. M., Knight, S., Göpel, S. O., Voldborg, B. G., Edfors, F., Forsström, B., von Feilitzen, K., Zwahlen, M., Rockberg, J., Takanen, J. O., Uhlén, M. & Hober, S. High throughput generation of a resource of the human secretome in mammalian cells. *N. Biotechnol.* **58,** 45–54 (2020).

138. Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).

139. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).

140. Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*

34, 267–273 (2003).

141. Weinberg, F., Hamanaka, R., Wheaton, W. W., Weinberg, S., Joseph, J., Lopez, M., Kalyanaraman, B., Mutlu, G. M., Budinger, G. R. S. & Chandel, N. S. Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 8788–8793 (2010).

142. Kuo, C.-C., Chiang, A. W. T., Baghdassarian, H. M. & Lewis, N. E. Dysregulation of the secretory pathway connects Alzheimer's disease genetics to aggregate formation. *Cell Syst.* (2021). doi:10.1016/j.cels.2021.06.001

143. Samoudi, M., Kuo, C.-C., Robinson, C. M., Shams-Ud-Doha, K., Schinn, S.-M., Kol, S., Weiss, L., Petersen Bjorn, S., Voldborg, B. G., Rosa Campos, A. & Lewis, N. E. In situ detection of protein interactions for recombinant therapeutic enzymes. *Biotechnol. Bioeng.* (2020). doi:10.1002/bit.27621

144. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43,** D234–9 (2015).

145. Sastry, A., Monk, J., Tegel, H., Uhlen, M., Palsson, B. O., Rockberg, J. & Brunk, E. Machine learning in computational biology to accelerate high-throughput protein expression. *Bioinformatics* **33,** 2487–2495 (2017).

146. Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. & Sabeti, P. C. Detecting novel associations in large data sets. *Science* **334,** 1518–1524 (2011).

147. Efron, B. & Tibshirani, R. Improvements on Cross-Validation: The .632+ Bootstrap Method. *J. Am. Stat. Assoc.* **92,** 548 (1997).

148. Matasci, M., Hacker, D. L., Baldi, L. & Wurm, F. M. Recombinant therapeutic protein production in cultivated mammalian cells: current status and future prospects. *Drug Discov. Today Technol.* **5,** e37–42 (2008).

149. Jenkins, N., Murphy, L. & Tyther, R. Post-translational Modifications of Recombinant Proteins: Significance for Biopharmaceuticals. *Molecular Biotechnology* **39,** 113–118 (2008).

150. Novick, P., Ferro, S. & Schekman, R. Order of events in the yeast secretory pathway. *Cell* **25,** 461–469 (1981).

151. Reynaud, E. G. & Simpson, J. C. Navigating the secretory pathway: conference on exocytosis membrane structure and dynamics. *EMBO Rep.* **3,** 828–833 (2002).

152. Young, C. L., Yuraszeck, T. & Robinson, A. S. Decreased Secretion and Unfolded Protein Response Upregulation. *Methods in Enzymology* 235–260 (2011). doi:10.1016/b978-0-12-385928-0.00014-6

153. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **196,** 801–810 (2012).

154. Kim, D. I., Birendra, K. C., Zhu, W., Motamedchaboki, K., Doye, V. & Roux, K. J. Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc. Natl. Acad.*

Sci. U. S. A. **111,** E2453–61 (2014).

155. Rhee, H.-W., Zou, P., Udeshi, N. D., Martell, J. D., Mootha, V. K., Carr, S. A. & Ting, A. Y. Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339,** 1328–1331 (2013).

156. Varnaitė, R. & MacNeill, S. A. Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. *Proteomics* **16,** 2503–2518 (2016).

157. Firat-Karalar, E. N. & Stearns, T. Probing mammalian centrosome structure using BioID proximity-dependent biotinylation. *Methods Cell Biol.* **129,** 153–170 (2015).

158. Gupta, G. D., Coyaud, É., Gonçalves, J., Mojarad, B. A., Liu, Y., Wu, Q., Gheiratmand, L., Comartin, D., Tkach, J. M., Cheung, S. W. T., Bashkurov, M., Hasegan, M., Knight, J. D., Lin, Z.-Y., Schueler, M., Hildebrandt, F., Moffat, J., Gingras, A.-C., Raught, B. & Pelletier, L. A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. *Cell* **163,** 1484–1499 (2015).

159. Dong, J.-M., Tay, F. P.-L., Swa, H. L.-F., Gunaratne, J., Leung, T., Burke, B. & Manser, E. Proximity biotinylation provides insight into the molecular composition of focal adhesions at the nanometer scale. *Sci. Signal.* **9,** rs4 (2016).

160. Hoffman, A. M., Chen, Q., Zheng, T. & Nicchitta, C. V. Heterogeneous translational landscape of the endoplasmic reticulum revealed by ribosome proximity labeling and transcriptome analysis. *J. Biol. Chem.* **294,** 8942–8958 (2019).

161. Kim, D. I., Jensen, S. C., Noble, K. A., Kc, B., Roux, K. H., Motamedchaboki, K. & Roux, K. J. An improved smaller biotin ligase for BioID proximity labeling. *Mol. Biol. Cell* **27,** 1188–1196 (2016).

162. Schindelin, J., Rueden, C. T., Hiner, M. C. & Eliceiri, K. W. The ImageJ ecosystem: An open platform for biomedical image analysis. *Mol. Reprod. Dev.* **82,** 518–529 (2015).

163. Li, Q. A Syntaxin 1, G o, and N-Type Calcium Channel Complex at a Presynaptic Nerve Terminal: Analysis by Quantitative Immunocolocalization. *Journal of Neuroscience* **24,** 4070–4081 (2004).

164. Manders, E. M. M., Verbeek, F. J. & Aten, J. A. Measurement of co-localization of objects in dual-colour confocal images. *Journal of Microscopy* **169,** 375–382 (1993).

165. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11,** 2301–2319 (2016).

166. Smyth, G. K. limma: Linear Models for Microarray Data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420 doi:10.1007/0-387-29362-0_23

167. Chawade, A., Alexandersson, E. & Levander, F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res.* **13,** 3114–3120 (2014).

168. Tyanova, S. & Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. *Methods Mol. Biol.* **1711,** 133–148 (2018).

169. Zhang, X., Smits, A. H., van Tilburg, G. B., Ovaa, H., Huber, W. & Vermeulen, M.

Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nat. Protoc.* **13,** 530–550 (2018).

170. Blatch, G. L. & Lässle, M. The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays* **21,** 932–939 (1999).

171. Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V. & Skrzypek, E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43,** D512–20 (2015).

172. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. (CRC Press, 2020).

173. Tytgat, H. L. P., Schoofs, G., Driesen, M., Proost, P., Van Damme, E. J. M., Vanderleyden, J. & Lebeer, S. Endogenous biotin-binding proteins: an overlooked factor causing false positives in streptavidin-based protein detection. *Microb. Biotechnol.* **8,** 164–168 (2015).

174. Mayer, M. P. Gymnastics of molecular chaperones. *Mol. Cell* **39,** 321–331 (2010).

175. Calakos, N., Bennett, M. K., Peterson, K. E. & Scheller, R. H. Protein-protein interactions contributing to the specificity of intracellular vesicular trafficking. *Science* **263,** 1146–1149 (1994).

176. Pearl, L. H. & Prodromou, C. Structure and mechanism of the Hsp90 molecular chaperone machinery. *Annu. Rev. Biochem.* **75,** 271–294 (2006).

177. Watanabe, S., Amagai, Y., Sannino, S., Tempio, T., Anelli, T., Harayama, M., Masui, S., Sorrentino, I., Yamada, M., Sitia, R. & Inaba, K. Zinc regulates ERp44-dependent protein quality control in the early secretory pathway. *Nat. Commun.* **10,** 603 (2019).

178. Bonifacino, J. S. & Glick, B. S. The Mechanisms of Vesicle Budding and Fusion. *Cell* **116,** 153–166 (2004).

179. Nauseef, W. M., McCormick, S. J. & Clark, R. A. Calreticulin functions as a molecular chaperone in the biosynthesis of myeloperoxidase. *J. Biol. Chem.* **270,** 4741–4747 (1995).

180. Ferris, S. P., Jaber, N. S., Molinari, M., Arvan, P. & Kaufman, R. J. UDP-glucose:glycoprotein glucosyltransferase (UGGT1) promotes substrate solubility in the endoplasmic reticulum. *Mol. Biol. Cell* **24,** 2597–2608 (2013).

181. Sakono, M., Seko, A., Takeda, Y. & Ito, Y. PDI family protein ERp29 forms 1:1 complex with lectin chaperone calreticulin. *Biochem. Biophys. Res. Commun.* **452,** 27–31 (2014).

182. Tannous, A., Pisoni, G. B., Hebert, D. N. & Molinari, M. N-linked sugar-regulated protein folding and quality control in the ER. *Semin. Cell Dev. Biol.* **41,** 79–89 (2015).

183. Ferris, S. P., Kodali, V. K. & Kaufman, R. J. Glycoprotein folding and quality-control mechanisms in protein-folding diseases. *Dis. Model. Mech.* **7,** 331–341 (2014).

184. Ninagawa, S., Okada, T., Sumitomo, Y., Kamiya, Y., Kato, K., Horimoto, S., Ishikawa, T., Takeda, S., Sakuma, T., Yamamoto, T. & Mori, K. EDEM2 initiates mammalian glycoprotein ERAD by catalyzing the first mannose trimming step. *J. Cell Biol.* **206,** 347–356 (2014).

185. Kozlov, G., Määttänen, P., Thomas, D. Y. & Gehring, K. A structural overview of the PDI family of proteins. *FEBS J.* **277,** 3924–3936 (2010).

186. Ng, D. T., Watowich, S. S. & Lamb, R. A. Analysis in vivo of GRP78-BiP/substrate interactions and their role in induction of the GRP78-BiP gene. *Mol. Biol. Cell* **3,** 143–155 (1992).

187. Yu, M., Haslam, R. H. & Haslam, D. B. HEDJ, an Hsp40 co-chaperone localized to the endoplasmic reticulum of human cells. *J. Biol. Chem.* **275,** 24984–24992 (2000).

188. Anelli, T., Alessio, M., Bachi, A., Bergamelli, L., Bertoli, G., Camerini, S., Mezghrani, A., Ruffato, E., Simmen, T. & Sitia, R. Thiol-mediated protein retention in the endoplasmic reticulum: the role of ERp44. *EMBO J.* **22,** 5015–5022 (2003).

189. Mezghrani, A., Fassio, A., Benham, A., Simmen, T., Braakman, I. & Sitia, R. Manipulation of oxidative protein folding and PDI redox state in mammalian cells. *EMBO J.* **20,** 6288–6296 (2001).

190. Zito, E. PRDX4, an endoplasmic reticulum-localized peroxiredoxin at the crossroads between enzymatic oxidative protein folding and nonenzymatic protein oxidation. *Antioxid. Redox Signal.* **18,** 1666–1674 (2013).

191. Siegenthaler, K. D., Pareja, K. A., Wang, J. & Sevier, C. S. An unexpected role for the yeast nucleotide exchange factor Sil1 as a reductant acting on the molecular chaperone BiP. *Elife* **6,** (2017).

192. Kittler, R., Heninger, A.-K., Franke, K., Habermann, B. & Buchholz, F. Production of endoribonuclease-prepared short interfering RNAs for gene silencing in mammalian cells. *Nat. Methods* **2,** 779–784 (2005).

193. Golovin, A. & Henrick, K. MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics* **9,** 312 (2008).

194. Imperiali, B., Shannon, K. L. & Rickert, K. W. Role of peptide conformation in asparagine-linked glycosylation. *Journal of the American Chemical Society* **114,** 7942–7944 (1992).

195. Imperiali, B., Shannon, K. L., Unno, M. & Rickert, K. W. Mechanistic proposal for asparagine-linked glycosylation. *Journal of the American Chemical Society* **114,** 7944–7945 (1992).

196. Bonito-Oliva, A., Barbash, S., Sakmar, T. P. & Graham, W. V. Nucleobindin 1 binds to multiple types of pre-fibrillar amyloid and inhibits fibrillization. *Sci. Rep.* **7,** 42880 (2017).

197. Doig, A. J., Stapley, B. J., Macarthur, M. W. & Thornton, J. M. Structures of N-termini of helices in proteins. *Protein Science* **6,** 147–155 (2008).

198. Aurora, R. & Rosee, G. D. Helix capping. *Protein Science* **7,** 21–38 (1998).

199. Nyfeler, B., Michnick, S. W. & Hauri, H.-P. Capturing protein interactions in the secretory pathway of living cells. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 6350–6355 (2005).

200. Schreiber, G., Haran, G. & Zhou, H.-X. Fundamental aspects of protein-protein association kinetics. *Chem. Rev.* **109,** 839–860 (2009).

201. Sears, R. M., May, D. G. & Roux, K. J. BioID as a Tool for Protein-Proximity Labeling in Living Cells. *Methods Mol. Biol.* **2012,** 299–313 (2019).

202. Mathias, S., Wippermann, A., Raab, N., Zeh, N., Handrick, R., Gorr, I., Schulz, P., Fischer, S., Gamer, M. & Otte, K. Unraveling what makes a monoclonal antibody difficult-to-express: From intracellular accumulation to incomplete folding and degradation via ERAD. *Biotechnology and Bioengineering* **117,** 5–16 (2020).

203. Gidalevitz, T., Stevens, F. & Argon, Y. Orchestration of secretory protein folding by ER chaperones. *Biochim. Biophys. Acta* **1833,** 2410–2424 (2013).

204. Kim, D. I. & Roux, K. J. Filling the Void: Proximity-Based Labeling of Proteins in Living Cells. *Trends Cell Biol.* **26,** 804–817 (2016).

205. Branon, T. C., Bosch, J. A., Sanchez, A. D., Udeshi, N. D., Svinkina, T., Carr, S. A., Feldman, J. L., Perrimon, N. & Ting, A. Y. Author Correction: Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.* **38,** 108 (2020).

206. Chen, C.-L. & Perrimon, N. Proximity-dependent labeling methods for proteomic profiling in living cells. *Wiley Interdiscip. Rev. Dev. Biol.* **6,** (2017).

207. Bar, D. Z., Atkatsh, K., Tavarez, U., Erdos, M. R., Gruenbaum, Y. & Collins, F. S. Biotinylation by antibody recognition—a method for proximity labeling. *Nature Methods* **15,** 127–133 (2018).

208. Gutierrez, J. M., Feizi, A., Li, S., Kallehauge, T. B., Hefzi, H., Grav, L. M., Ley, D., Hizal, D. B., Betenbaugh, M. J., Voldborg, B. & Others. Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. *Nat. Commun.* **11,** 1–10 (2020).

209. Barnes, L. M., Bentley, C. M. & Dickson, A. J. Molecular definition of predictive indicators of stable protein expression in recombinant NS0 myeloma cells. *Biotechnol. Bioeng.* **85,** 115–121 (2004).

210. Uhlen, M., Tegel, H., Sivertsson, Å., Kuo, C.-C., Gutierrez, J. M., Lewis, N. E., Forsström, B., Dannemeyer, M., Fagerberg, L., Malm, M. & Others. The human secretome--the proteins secreted from human cells. *bioRxiv* 465815 (2018).

211. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J. & Pontén, F. Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).

212. Hukelmann, J. L., Anderson, K. E., Sinclair, L. V., Grzes, K. M., Murillo, A. B., Hawkins, P. T., Stephens, L. R., Lamond, A. I. & Cantrell, D. A. The cytotoxic T cell proteome and its shaping by the kinase mTOR. *Nat. Immunol.* **17,** 104–112 (2016).

213. Tan, H., Yang, K., Li, Y., Shaw, T. I., Wang, Y., Blanco, D. B., Wang, X., Cho, J.-H., Wang, H., Rankin, S., Guy, C., Peng, J. & Chi, H. Integrative Proteomics and Phosphoproteomics Profiling Reveals Dynamic Signaling Networks and Bioenergetics

Pathways Underlying T Cell Activation. *Immunity* **46,** 488–503 (2017).

214. Feizi, A., Österlund, T., Petranovic, D., Bordel, S. & Nielsen, J. Genome-scale modeling of the protein secretory machinery in yeast. *PLoS One* **8,** e63284 (2013).

215. Vassar, R., Bennett, B. D., Babu-Khan, S., Kahn, S., Mendiaz, E. A., Denis, P., Teplow, D. B., Ross, S., Amarante, P., Loeloff, R., Luo, Y., Fisher, S., Fuller, J., Edenson, S., Lile, J., Jarosinski, M. A., Biere, A. L., Curran, E., Burgess, T., Louis, J. C., Collins, F., Treanor, J., Rogers, G. & Citron, M. Beta-secretase cleavage of Alzheimer's amyloid precursor protein by the transmembrane aspartic protease BACE. *Science* **286,** 735–741 (1999).

216. Lammich, S., Kojro, E., Postina, R., Gilbert, S., Pfeiffer, R., Jasionowski, M., Haass, C. & Fahrenholz, F. Constitutive and regulated alpha-secretase cleavage of Alzheimer's amyloid precursor protein by a disintegrin metalloprotease. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 3922–3927 (1999).

217. De Strooper, B., Saftig, P., Craessaerts, K., Vanderstichele, H., Guhde, G., Annaert, W., Von Figura, K. & Van Leuven, F. Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391,** 387–390 (1998).

218. Thinakaran, G. & Koo, E. H. Amyloid precursor protein trafficking, processing, and function. *J. Biol. Chem.* **283,** 29615–29619 (2008).

219. Wang, X., Zhou, X., Li, G., Zhang, Y., Wu, Y. & Song, W. Modifications and Trafficking of APP in the Pathogenesis of Alzheimer's Disease. *Front. Mol. Neurosci.* **10,** 294 (2017).

220. Lee, M.-S., Kao, S.-C., Lemere, C. A., Xia, W., Tseng, H.-C., Zhou, Y., Neve, R., Ahlijanian, M. K. & Tsai, L.-H. APP processing is regulated by cytoplasmic phosphorylation. *J. Cell Biol.* **163,** 83–95 (2003).

221. McFarlane, I., Breen, K. C., Giamberardino, L. D. & Moya, K. L. Inhibition of N-glycan processing alters axonal transport of synaptic glycoproteins in vivo. *Neuroreport* **11,** 1543–1547 (2000).

222. McFarlane, I., Georgopoulou, N., Coughlan, C. M., Gillian, A. M. & Breen, K. C. The role of the protein glycosylation state in the control of cellular transport of the amyloid β precursor protein. *Neuroscience* **90,** 15–25 (1999).

223. Joshi, G. & Wang, Y. Golgi defects enhance APP amyloidogenic processing in Alzheimer's disease. *Bioessays* **37,** 240–247 (2015).

224. Schedin-Weiss, S., Winblad, B. & Tjernberg, L. O. The role of protein glycosylation in Alzheimer disease. *FEBS J.* **281,** 46–62 (2014).

225. Jiang, S., Li, Y., Zhang, X., Bu, G., Xu, H. & Zhang, Y.-W. Trafficking regulation of proteins in Alzheimer's disease. *Mol. Neurodegener.* **9,** 6 (2014).

226. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15,** 384–396 (2014).

227. Dourlen, P., Kilinc, D., Malmanche, N., Chapuis, J. & Lambert, J.-C. The new genetic landscape of Alzheimer's disease: from amyloid cascade to genetically driven synaptic failure hypothesis? *Acta Neuropathol.* **138,** 221–236 (2019).

228. Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., van der Lee, S. J., Amlie-Wolf, A., Bellenguez, C., Frizatti, A., Chouraki, V., Martin, E. R., Sleegers, K., Badarinarayan, N., Jakobsdottir, J., Hamilton-Nelson, K. L., Moreno-Grau, S., Olaso, R., Raybould, R., Chen, Y., Kuzma, A. B., Hiltunen, M., Morgan, T., Ahmad, S., Vardarajan, B. N., Epelbaum, J., Hoffmann, P., Boada, M., Beecham, G. W., Garnier, J.-G., Harold, D., Fitzpatrick, A. L., Valladares, O., Moutet, M.-L., Gerrish, A., Smith, A. V., Qu, L., Bacq, D., Denning, N., Jian, X., Zhao, Y., Del Zompo, M., Fox, N. C., Choi, S.-H., Mateo, I., Hughes, J. T., Adams, H. H., Malamon, J., Sanchez-Garcia, F., Patel, Y., Brody, J. A., Dombroski, B. A., Naranjo, M. C. D., Daniilidou, M., Eiriksdottir, G., Mukherjee, S., Wallon, D., Uphill, J., Aspelund, T., Cantwell, L. B., Garzia, F., Galimberti, D., Hofer, E., Butkiewicz, M., Fin, B., Scarpini, E., Sarnowski, C., Bush, W. S., Meslage, S., Kornhuber, J., White, C. C., Song, Y., Barber, R. C., Engelborghs, S., Sordon, S., Voijnovic, D., Adams, P. M., Vandenberghe, R., Mayhaus, M., Cupples, L. A., Albert, M. S., De Deyn, P. P., Gu, W., Himali, J. J., Beekly, D., Squassina, A., Hartmann, A. M., Orellana, A., Blacker, D., Rodriguez-Rodriguez, E., Lovestone, S., Garcia, M. E., Doody, R. S., Munoz-Fernadez, C., Sussams, R., Lin, H., Fairchild, T. J., Benito, Y. A., Holmes, C., Karamujić-Čomić, H., Frosch, M. P., Thonberg, H., Maier, W., Roshchupkin, G., Ghetti, B., Giedraitis, V., Kawalia, A., Li, S., Huebinger, R. M., Kilander, L., Moebus, S., Hernández, I., Kamboh, M. I., Brundin, R., Turton, J., Yang, Q., Katz, M. J., Concari, L., Lord, J., Beiser, A. S., Keene, C. D., Helisalmi, S., Kloszewska, I., Kukull, W. A., Koivisto, A. M., Lynch, A., Tarraga, L., Larson, E. B., Haapasalo, A., Lawlor, B., Mosley, T. H., Lipton, R. B., Solfrizzi, V., Gill, M., Longstreth, W. T., Jr, Montine, T. J., Frisardi, V., Diez-Fairen, M., Rivadeneira, F., Petersen, R. C., Deramecourt, V., Alvarez, I., Salani, F., Ciaramella, A., Boerwinkle, E., Reiman, E. M., Fievet, N., Rotter, J. I., Reisch, J. S., Hanon, O., Cupidi, C., Andre Uitterlinden, A. G., Royall, D. R., Dufouil, C., Maletta, R. G., de Rojas, I., Sano, M., Brice, A., Cecchetti, R., George-Hyslop, P. S., Ritchie, K., Tsolaki, M., Tsuang, D. W., Dubois, B., Craig, D., Wu, C.-K., Soininen, H., Avramidou, D., Albin, R. L., Fratiglioni, L., Germanou, A., Apostolova, L. G., Keller, L., Koutroumani, M., Arnold, S. E., Panza, F., Gkatzima, O., Asthana, S., Hannequin, D., Whitehead, P., Atwood, C. S., Caffarra, P., Hampel, H., Quintela, I., Carracedo, Á., Lannfelt, L., Rubinsztein, D. C., Barnes, L. L., Pasquier, F., Frölich, L., Barral, S., McGuinness, B., Beach, T. G., Johnston, J. A., Becker, J. T., Passmore, P., Bigio, E. H., Schott, J. M., Bird, T. D., Warren, J. D., Boeve, B. F., Lupton, M. K., Bowen, J. D., Proitsi, P., Boxer, A., Powell, J. F., Burke, J. R., Kauwe, J. S. K., Burns, J. M., Mancuso, M., Buxbaum, J. D., Bonuccelli, U., Cairns, N. J., McQuillin, A., Cao, C., Livingston, G., Carlson, C. S., Bass, N. J., Carlsson, C. M., Hardy, J., Carney, R. M., Bras, J., Carrasquillo, M. M., Guerreiro, R., Allen, M., Chui, H. C., Fisher, E., Masullo, C., Crocco, E. A., DeCarli, C., Bisceglio, G., Dick, M., Ma, L., Duara, R., Graff-Radford, N. R., Evans, D. A., Hodges, A., Faber, K. M., Scherer, M., Fallon, K. B., Riemenschneider, M., Fardo, D. W., Heun, R., Farlow, M. R., Kölsch, H., Ferris, S., Leber, M., Foroud, T. M., Heuser, I., Galasko, D. R., Giegling, I., Gearing, M., Hüll, M., Geschwind, D. H., Gilbert, J. R., Morris, J., Green, R. C., Mayo, K., Growdon, J. H., Feulner, T., Hamilton, R. L., Harrell, L. E., Drichel, D., Honig, L. S., Cushion, T. D., Huentelman, M. J., Hollingworth, P., Hulette, C. M., Hyman, B. T., Marshall, R., Jarvik, G. P., Meggy, A., Abner, E., Menzies, G. E., Jin, L.-W., Leonenko, G., Real, L. M., Jun, G. R., Baldwin, C. T., Grozeva, D., Karydas, A., Russo, G., Kaye, J. A., Kim, R., Jessen, F., Kowall, N. W., Vellas, B., Kramer, J. H., Vardy, E., LaFerla, F. M., Jöckel, K.-H., Lah, J. J., Dichgans, M., Leverenz, J. B., Mann, D., Levey, A. I., Pickering-Brown, S., Lieberman, A. P., Klopp, N., Lunetta, K. L., Wichmann, H.-E., Lyketsos, C. G., Morgan, K., Marson, D. C., Brown, K., Martiniuk, F., Medway, C., Mash, D. C., Nöthen, M. M., Masliah, E., Hooper, N. M., McCormick, W. C., Daniele, A., McCurry, S. M., Bayer, A., McDavid, A. N., Gallacher, J., McKee, A. C., van den Bussche, H., Mesulam, M., Brayne, C., Miller, B. L., Riedel-Heller, S., Miller, C. A., Miller, J. W., Al-Chalabi, A.,

Morris, J. C., Shaw, C. E., Myers, A. J., Wiltfang, J., O'Bryant, S., Olichney, J. M., Alvarez, V., Parisi, J. E., Singleton, A. B., Paulson, H. L., Collinge, J., Perry, W. R., Mead, S., Peskind, E., Cribbs, D. H., Rossor, M., Pierce, A., Ryan, N. S., Poon, W. W., Nacmias, B., Potter, H., Sorbi, S., Quinn, J. F., Sacchinelli, E., Raj, A., Spalletta, G., Raskind, M., Caltagirone, C., Bossù, P., Orfei, M. D., Reisberg, B., Clarke, R., Reitz, C., Smith, A. D., Ringman, J. M., Warden, D., Roberson, E. D., Wilcock, G., Rogaeva, E., Bruni, A. C., Rosen, H. J., Gallo, M., Rosenberg, R. N., Ben-Shlomo, Y., Sager, M. A., Mecocci, P., Saykin, A. J., Pastor, P., Cuccaro, M. L., Vance, J. M., Schneider, J. A., Schneider, L. S., Slifer, S., Seeley, W. W., Smith, A. G., Sonnen, J. A., Spina, S., Stern, R. A., Swerdlow, R. H., Tang, M., Tanzi, R. E., Trojanowski, J. Q., Troncoso, J. C., Van Deerlin, V. M., Van Eldik, L. J., Vinters, H. V., Vonsattel, J. P., Weintraub, S., Welsh-Bohmer, K. A., Wilhelmsen, K. C., Williamson, J., Wingo, T. S., Woltjer, R. L., Wright, C. B., Yu, C.-E., Yu, L., Saba, Y., Pilotto, A., Bullido, M. J., Peters, O., Crane, P. K., Bennett, D., Bosco, P., Coto, E., Boccardi, V., De Jager, P. L., Lleo, A., Warner, N., Lopez, O. L., Ingelsson, M., Deloukas, P., Cruchaga, C., Graff, C., Gwilliam, R., Fornage, M., Goate, A. M., Sanchez-Juan, P., Kehoe, P. G., Amin, N., Ertekin-Taner, N., Berr, C., Debette, S., Love, S., Launer, L. J., Younkin, S. G., Dartigues, J.-F., Corcoran, C., Ikram, M. A., Dickson, D. W., Nicolas, G., Campion, D., Tschanz, J., Schmidt, H., Hakonarson, H., Clarimon, J., Munger, R., Schmidt, R., Farrer, L. A., Van Broeckhoven, C., C O'Donovan, M., DeStefano, A. L., Jones, L., Haines, J. L., Deleuze, J.-F., Owen, M. J., Gudnason, V., Mayeux, R., Escott-Price, V., Psaty, B. M., Ramirez, A., Wang, L.-S., Ruiz, A., van Duijn, C. M., Holmans, P. A., Seshadri, S., Williams, J., Amouyel, P., Schellenberg, G. D., Lambert, J.-C., Pericak-Vance, M. A., Alzheimer Disease Genetics Consortium (ADGC), European Alzheimer's Disease Initiative (EADI), Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE), & Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES),. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51,** 414–430 (2019).

229. Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witoelar, A., Stringer, S., Aarsland, D., Almdahl, I. S., Andersen, F., Bergh, S., Bettella, F., Bjornsson, S., Brækhus, A., Bråthen, G., de Leeuw, C., Desikan, R. S., Djurovic, S., Dumitrescu, L., Fladby, T., Hohman, T. J., Jonsson, P. V., Kiddle, S. J., Rongve, A., Saltvedt, I., Sando, S. B., Selbæk, G., Shoai, M., Skene, N. G., Snaedal, J., Stordal, E., Ulstein, I. D., Wang, Y., White, L. R., Hardy, J., Hjerling-Leffler, J., Sullivan, P. F., van der Flier, W. M., Dobson, R., Davis, L. K., Stefansson, H., Stefansson, K., Pedersen, N. L., Ripke, S., Andreassen, O. A. & Posthuma, D. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51,** 404–413 (2019).

230. Ridge, P. G., Hoyt, K. B., Boehme, K., Mukherjee, S., Crane, P. K., Haines, J. L., Mayeux, R., Farrer, L. A., Pericak-Vance, M. A., Schellenberg, G. D., Kauwe, J. S. K. & Alzheimer's Disease Genetics Consortium (ADGC). Assessment of the genetic variance of late-onset Alzheimer's disease. *Neurobiol. Aging* **41,** 200.e13–200.e20 (2016).

231. Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., Zhang, C., Xie, T., Tran, L., Dobrin, R., Fluder, E., Clurman, B., Melquist, S., Narayanan, M., Suver, C., Shah, H., Mahajan, M., Gillis, T., Mysore, J., MacDonald, M. E., Lamb, J. R., Bennett, D. A., Molony, C., Stone, D. J., Gudnason, V., Myers, A. J., Schadt, E. E., Neumann, H., Zhu, J. & Emilsson, V. Integrated systems approach identifies genetic nodes

and networks in late-onset Alzheimer's disease. *Cell* **153,** 707–720 (2013).

232. Greenfield, J. P., Tsai, J., Gouras, G. K., Hai, B., Thinakaran, G., Checler, F., Sisodia, S. S., Greengard, P. & Xu, H. Endoplasmic reticulum and trans-Golgi network generate distinct populations of Alzheimer beta-amyloid peptides. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 742–747 (1999).

233. Hartmann, T., Bieger, S. C., Brühl, B., Tienari, P. J., Ida, N., Allsop, D., Roberts, G. W., Masters, C. L., Dotti, C. G., Unsicker, K. & Beyreuther, K. Distinct sites of intracellular production for Alzheimer's disease A beta40/42 amyloid peptides. *Nat. Med.* **3,** 1016–1020 (1997).

234. Wang, M., Beckmann, N. D., Roussos, P., Wang, E., Zhou, X., Wang, Q., Ming, C., Neff, R., Ma, W., Fullard, J. F., Hauberg, M. E., Bendl, J., Peters, M. A., Logsdon, B., Wang, P., Mahajan, M., Mangravite, L. M., Dammer, E. B., Duong, D. M., Lah, J. J., Seyfried, N. T., Levey, A. I., Buxbaum, J. D., Ehrlich, M., Gandy, S., Katsel, P., Haroutunian, V., Schadt, E. & Zhang, B. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Scientific Data* **5,** 180185 (2018).

235. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett, D. A., Kellis, M. & Tsai, L.-H. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570,** 332–337 (2019).

236. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5,** e1000598 (2009).

237. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31,** 3718–3720 (2015).

238. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).

239. Anelli, T. & Sitia, R. Protein quality control in the early secretory pathway. *EMBO J.* **27,** 315–327 (2008).

240. Bonifacino, J. S. & Glick, B. S. The Mechanisms of Vesicle Budding and Fusion. *Cell* **116,** 153–166 (2004).

241. Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerrière, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M., Dubas, F., Frebourg, T. & Campion, D. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.* **38,** 24–26 (2006).

242. Bushman, D. M., Kaeser, G. E., Siddoway, B., Westra, J. W., Rivera, R. R., Rehen, S. K., Yung, Y. C. & Chun, J. Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. *Elife* **4,** (2015).

243. Matsui, T., Ingelsson, M., Fukumoto, H., Ramasamy, K., Kowa, H., Frosch, M. P., Irizarry, M. C. & Hyman, B. T. Expression of APP pathway mRNAs and proteins in Alzheimer's disease. *Brain Res.* **1161,** 116–123 (2007).

244. De Strooper, B., Vassar, R. & Golde, T. The secretases: enzymes with therapeutic potential in Alzheimer disease. *Nat. Rev. Neurol.* **6,** 99–107 (2010).

245. Laird, F. M., Cai, H., Savonenko, A. V., Farah, M. H., He, K., Melnikova, T., Wen, H., Chiang, H.-C., Xu, G., Koliatsos, V. E., Borchelt, D. R., Price, D. L., Lee, H.-K. & Wong, P. C. BACE1, a major determinant of selective vulnerability of the brain to amyloid-beta amyloidogenesis, is essential for cognitive, emotional, and synaptic functions. *J. Neurosci.* **25,** 11693–11709 (2005).

246. Frost, G. R. & Li, Y.-M. The role of astrocytes in amyloid production and Alzheimer's disease. *Open Biol.* **7,** (2017).

247. Phatnani, H. & Maniatis, T. Astrocytes in neurodegenerative disease. *Cold Spring Harb. Perspect. Biol.* **7,** (2015).

248. Sofroniew, M. V. & Vinters, H. V. Astrocytes: biology and pathology. *Acta Neuropathol.* **119,** 7–35 (2010).

249. Liddelow, S. A. & Barres, B. A. Reactive Astrocytes: Production, Function, and Therapeutic Potential. *Immunity* **46,** 957–967 (2017).

250. Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., DeStafano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thorton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., Vardarajan, B. N., Kamatani, Y., Lin, C. F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M. L., Ruiz, A., Bihoreau, M. T., Choi, S. H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez, O. L., De Jager, P. L., Deramecourt, V., Johnston, J. A., Evans, D., Lovestone, S., Letenneur, L., Morón, F. J., Rubinsztein, D. C., Eiriksdottir, G., Sleegers, K., Goate, A. M., Fiévet, N., Huentelman, M. W., Gill, M., Brown, K., Kamboh, M. I., Keller, L., Barberger-Gateau, P., McGuiness, B., Larson, E. B., Green, R., Myers, A. J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D. W., Yu, L., Tsolaki, M., Bossù, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N. C., Hardy, J., Deniz Naranjo, M. C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer's Disease, Alzheimer's Disease Genetic Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J. R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M. M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S. G., Coto, E., Hamilton-Nelson, K. L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M. J., Faber, K. M., Jonsson, P. V., Combarros, O., O'Donovan, M. C., Cantwell, L. B., Soininen, H., Blacker, D., Mead, S., Mosley, T. H., Jr, Bennett, D. A., Harris, T. B., Fratiglioni, L., Holmes, C., de Bruijn, R. F., Passmore, P., Montine, T. J., Bettens, K., Rotter, J. I., Brice, A., Morgan, K., Foroud, T. M., Kukull, W. A., Hannequin, D., Powell, J. F., Nalls, M. A., Ritchie, K., Lunetta, K. L., Kauwe, J. S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltuenen, M., Martin, E. R., Schmidt, R., Rujescu, D., Wang, L. S., Dartigues, J. F., Mayeux, R., Tzourio, C., Hofman, A., Nöthen, M. M., Graff, C., Psaty, B. M., Jones, L., Haines, J. L., Holmans, P. A., Lathrop, M., Pericak-Vance, M. A., Launer, L. J., Farrer, L. A., van Duijn, C. M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G. D. & Amouyel, P. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–1458 (2013).

251. Lardenoije, R., Iatrou, A., Kenis, G., Kompotis, K., Steinbusch, H. W. M., Mastroeni, D., Coleman, P., Lemere, C. A., Hof, P. R., van den Hove, D. L. A. & Rutten, B. P. F. The epigenetics of aging and neurodegeneration. *Prog. Neurobiol.* **131,** 21–64 (2015).

252. Liu, X., Jiao, B. & Shen, L. The Epigenetics of Alzheimer's Disease: Factors and Therapeutic Implications. *Front. Genet.* **9,** 579 (2018).

253. Klein, H.-U., McCabe, C., Gjoneska, E., Sullivan, S. E., Kaskow, B. J., Tang, A., Smith, R. V., Xu, J., Pfenning, A. R., Bernstein, B. E., Meissner, A., Schneider, J. A., Mostafavi, S., Tsai, L.-H., Young-Pearse, T. L., Bennett, D. A. & De Jager, P. L. Epigenome-wide study uncovers large-scale changes in histone acetylation driven by tau pathology in aging and Alzheimer's human brains. *Nat. Neurosci.* **22,** 37–46 (2019).

254. Nativio, R., Lan, Y., Donahue, G., Sidoli, S., Berson, A., Srinivasan, A. R., Shcherbakova, O., Amlie-Wolf, A., Nie, J., Cui, X., He, C., Wang, L.-S., Garcia, B. A., Trojanowski, J. Q., Bonini, N. M. & Berger, S. L. An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat. Genet.* **52,** 1024–1035 (2020).

255. Nativio, R., Donahue, G., Berson, A., Lan, Y., Amlie-Wolf, A., Tuzer, F., Toledo, J. B., Gosai, S. J., Gregory, B. D., Torres, C., Trojanowski, J. Q., Wang, L.-S., Johnson, F. B., Bonini, N. M. & Berger, S. L. Dysregulation of the epigenetic landscape of normal aging in Alzheimer's disease. *Nat. Neurosci.* **21,** 497–505 (2018).

256. Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H. & Kellis, M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518,** 365–369 (2015).

257. Cataldo, A. M., Peterhoff, C. M., Troncoso, J. C., Gomez-Isla, T., Hyman, B. T. & Nixon, R. A. Endocytic Pathway Abnormalities Precede Amyloid β Deposition in Sporadic Alzheimer's Disease and Down Syndrome. *Am. J. Pathol.* **157,** 277–286 (2000).

258. Israel, M. A., Yuan, S. H., Bardy, C., Reyna, S. M., Mu, Y., Herrera, C., Hefferan, M. P., Van Gorp, S., Nazor, K. L., Boscolo, F. S., Carson, C. T., Laurent, L. C., Marsala, M., Gage, F. H., Remes, A. M., Koo, E. H. & Goldstein, L. S. B. Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. *Nature* **482,** 216–220 (2012).

259. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

260. Krämer, A., Green, J., Pollard, J., Jr & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30,** 523–530 (2014).

261. Citron, B. A., Saykally, J. N., Cao, C., Dennis, J. S., Runfeldt, M. & Arendash, G. W. Transcription factor Sp1 inhibition, memory, and cytokines in a mouse model of Alzheimer's disease. *Am. J. Neurodegener. Dis.* **4,** 40–48 (2015).

262. Santpere, G., Nieto, M., Puig, B. & Ferrer, I. Abnormal Sp1 transcription factor expression in Alzheimer disease and tauopathies. *Neurosci. Lett.* **397,** 30–34 (2006).

263. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D. & Cohen, D. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017,** (2017).

264. Allen, M., Carrasquillo, M. M., Funk, C., Heavner, B. D., Zou, F., Younkin, C. S., Burgess, J. D., Chai, H.-S., Crook, J., Eddy, J. A., Li, H., Logsdon, B., Peters, M. A., Dang, K. K., Wang, X., Serie, D., Wang, C., Nguyen, T., Lincoln, S., Malphrus, K., Bisceglio, G., Li, M., Golde, T. E., Mangravite, L. M., Asmann, Y., Price, N. D., Petersen, R. C., Graff-Radford, N. R., Dickson, D. W., Younkin, S. G. & Ertekin-Taner, N. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data* **3,** 160089 (2016).

265. De Jager, P. L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B. N., Felsky, D., Klein, H.-U., White, C. C., Peters, M. A., Lodgson, B., Nejad, P., Tang, A., Mangravite, L. M., Yu, L., Gaiteri, C., Mostafavi, S., Schneider, J. A. & Bennett, D. A. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci Data* **5,** 180142 (2018).

266. Bertram, L., Lill, C. M. & Tanzi, R. E. The genetics of Alzheimer disease: back to the future. *Neuron* **68,** 270–281 (2010).

267. Walther, D. M., Kasturi, P., Zheng, M., Pinkert, S., Vecchi, G., Ciryam, P., Morimoto, R. I., Dobson, C. M., Vendruscolo, M., Mann, M. & Hartl, F. U. Widespread Proteome Remodeling and Aggregation in Aging C. elegans. *Cell* **168,** 944 (2017).

268. Brehme, M., Voisine, C., Rolland, T., Wachi, S., Soper, J. H., Zhu, Y., Orton, K., Villella, A., Garza, D., Vidal, M., Ge, H. & Morimoto, R. I. A chaperome subnetwork safeguards proteostasis in aging and neurodegenerative disease. *Cell Rep.* **9,** 1135–1150 (2014).

269. Hipp, M. S., Kasturi, P. & Hartl, F. U. The proteostasis network and its decline in ageing. *Nat. Rev. Mol. Cell Biol.* **20,** 421–435 (2019).

270. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F. & Parkinson, H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).

271. Burrinha, T., Gomes, R., Terrasso, A. P. & Almeida, C. G. Neuronal aging potentiates beta-amyloid generation via amyloid precursor protein endocytosis. *Neuroscience* 1633 (2019).

272. Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science* **297,** 353–356 (2002).

273. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **8,** 595–608 (2016).

274. Hardy, J. A. & Higgins, G. A. Alzheimer's disease: the amyloid cascade hypothesis. *Science* **256,** 184–185 (1992).

275. Lewis, J., Dickson, D. W., Lin, W. L., Chisholm, L., Corral, A., Jones, G., Yen, S. H., Sahara, N., Skipper, L., Yager, D., Eckman, C., Hardy, J., Hutton, M. & McGowan, E. Enhanced neurofibrillary degeneration in transgenic mice expressing mutant tau and APP. *Science* **293,** 1487–1491 (2001).

276. Hardy, J., Duff, K., Hardy, K. G., Perez-Tur, J. & Hutton, M. Genetic dissection of Alzheimer's disease and related dementias: amyloid and its relationship to tau. *Nat.*

Neurosci. **1,** 355–358 (1998).

277. Bateman, R. J., Xiong, C., Benzinger, T. L. S., Fagan, A. M., Goate, A., Fox, N. C., Marcus, D. S., Cairns, N. J., Xie, X., Blazey, T. M., Holtzman, D. M., Santacruz, A., Buckles, V., Oliver, A., Moulder, K., Aisen, P. S., Ghetti, B., Klunk, W. E., McDade, E., Martins, R. N., Masters, C. L., Mayeux, R., Ringman, J. M., Rossor, M. N., Schofield, P. R., Sperling, R. A., Salloway, S., Morris, J. C. & Dominantly Inherited Alzheimer Network. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* **367,** 795–804 (2012).

278. Jack, C. R., Jr, Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., Shaw, L. M., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Pankratz, V. S., Donohue, M. C. & Trojanowski, J. Q. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol.* **12,** 207–216 (2013).

279. Meex, R. C., Hoy, A. J., Morris, A., Brown, R. D., Lo, J. C. Y., Burke, M., Goode, R. J. A., Kingwell, B. A., Kraakman, M. J., Febbraio, M. A., Greve, J. W., Rensen, S. S., Molloy, M. P., Lancaster, G. I., Bruce, C. R. & Watt, M. J. Fetuin B Is a Secreted Hepatocyte Factor Linking Steatosis to Impaired Glucose Metabolism. *Cell Metab.* **22,** 1078–1089 (2015).

280. Gorden, D. L., Myers, D. S., Ivanova, P. T., Fahy, E., Maurya, M. R., Gupta, S., Min, J., Spann, N. J., McDonald, J. G., Kelly, S. L., Duan, J., Sullards, M. C., Leiker, T. J., Barkley, R. M., Quehenberger, O., Armando, A. M., Milne, S. B., Mathews, T. P., Armstrong, M. D., Li, C., Melvin, W. V., Clements, R. H., Washington, M. K., Mendonsa, A. M., Witztum, J. L., Guan, Z., Glass, C. K., Murphy, R. C., Dennis, E. A., Merrill, A. H., Jr, Russell, D. W., Subramaniam, S. & Brown, H. A. Biomarkers of NAFLD progression: a lipidomics approach to an epidemic. *J. Lipid Res.* **56,** 722–736 (2015).

281. Robinson, J. L., Feizi, A., Uhlén, M. & Nielsen, J. A Systematic Investigation of the Malignant Functions and Diagnostic Potential of the Cancer Secretome. *Cell Rep.* **26,** 2622–2635.e5 (2019).

282. Wang, D., Eraslan, B., Wieland, T., Hallström, B., Hopf, T., Zolg, D. P., Zecha, J., Asplund, A., Li, L.-H., Meng, C., Frejno, M., Schmidt, T., Schnatbaum, K., Wilhelm, M., Ponten, F., Uhlen, M., Gagneur, J., Hahne, H. & Kuster, B. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15,** e8503 (2019).

283. Huang, J. K., Carlin, D. E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P. & Ideker, T. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell Syst* **6,** 484–495.e5 (2018).

284. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43,** W39–49 (2015).

285. Page, L., Brin, S., Motwani, R. & Winograd, T. The Pagerank Citation Ranking: Bringing Order to the web. *technical report* (1998).

286. Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., Lee, S., Lindskog, C., Mulder, J., Mulvey, C. M., Nilsson, P., Oksvold, P., Rockberg, J., Schutten, R., Schwenk, J. M.,

Sivertsson, Å., Sjöstedt, E., Skogs, M., Stadler, C., Sullivan, D. P., Tegel, H., Winsnes, C., Zhang, C., Zwahlen, M., Mardinoglu, A., Pontén, F., von Feilitzen, K., Lilley, K. S., Uhlén, M. & Lundberg, E. A subcellular map of the human proteome. *Science* **356,** (2017).

287. Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S. & Schneider, J. A. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers. Dis.* **64,** S161–S189 (2018).

288. Kendall, M. G. & Stuart, A. *The Advanced Theory of Statistics: Inference and relationship*. (Hafner Press, 1977).

289. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11,** e1004219 (2015).

290. Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y. & Cherry, J. M. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46,** D794–D801 (2018).

291. Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R. & Ma'ayan, A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26,** 2438–2444 (2010).

292. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W. & Ma'ayan, A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–7 (2016).

293. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. *Bioinformatics* 471 (2016).

294. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

295. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** D91–4 (2004).

296. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8,** R24 (2007).

297. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6,** e21800 (2011).

298. Zhang, J., Petersen, S. D., Radivojevic, T., Ramirez, A., Pérez-Manríquez, A., Abeliuk, E., Sánchez, B. J., Costello, Z., Chen, Y., Fero, M. J., Martin, H. G., Nielsen, J., Keasling, J. D. & Jensen, M. K. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* **11,** 4880 (2020).

299. Kol, S., Ley, D., Wulff, T., Decker, M., Arnsdorf, J., Gutierrez, J. M., Chiang, A. W. T., Pedersen, L. E., Kildegaard, H. F., Lee, G. M. & Lewis, N. E. Multiplex secretome engineering enhances recombinant protein production and purity. *Nature Communications*

doi:10.1101/647214

300. Kuo, C.-C., Chiang, A. W., Shamie, I., Samoudi, M., Gutierrez, J. M. & Lewis, N. E. The emerging role of systems biology for engineering protein production in CHO cells. *Curr. Opin. Biotechnol.* **51,** (2018).