Microbial Cell Factories

RESEARCH

Open Access

Translational landscape and protein biogenesis demands of the early secretory pathway in *Komagataella phaffii*

Troy R. Alva[1*], Melanie Riera[1] and Justin W. Chartron[1,2]

Abstract

Background: Eukaryotes use distinct networks of biogenesis factors to synthesize, fold, monitor, traffic, and secrete proteins. During heterologous expression, saturation of any of these networks may bottleneck titer and yield. To understand the flux through various routes into the early secretory pathway, we quantified the global and membrane-associated translatomes of *Komagataella phaffii*.

Results: By coupling Ribo-seq with long-read mRNA sequencing, we generated a new annotation of protein-encoding genes. By using Ribo-seq with subcellular fractionation, we quantified demands on co- and posttranslational translocation pathways. During exponential growth in rich media, protein components of the cell-wall represent the greatest number of nascent chains entering the ER. Transcripts encoding the transmembrane protein *PMA1* sequester more ribosomes at the ER membrane than any others. Comparison to *Saccharomyces cerevisiae* reveals conservation in the resources allocated by gene ontology, but variation in the diversity of gene products entering the secretory pathway.

Conclusion: A subset of host proteins, particularly cell-wall components, impose the greatest biosynthetic demands in the early secretory pathway. These proteins are potential targets in strain engineering aimed at alleviating bottlenecks during heterologous protein production.

Keywords: Ribosome profiling, Protein secretion, Resource allocation, *Pichia pastoris*

As microbial cell factories, yeasts offer many advantages for recombinant protein production including their natural properties and potential in synthetic biology. Yeasts grow rapidly to high densities in inexpensive media and are resistant to physical and chemical stress [1]. They also have an endomembrane system that is fundamentally conserved with higher eukaryotes [2]. This oxidative environment supports glycosylation and subsequent glycan modification, folding using ATP-driven molecular chaperones and protein disulfide isomerases, and protein quality control [3]. Compared to mammalian cells, yeasts have simpler genomes and can be more easily characterized and modified [4]. Combine these features with tools such as CRISPR/cas9, and the range of tractable species is expanding [5, 6]. *Komagataella phaffii* (one of two species previously known as *P. pastoris* [7–9]) stands out as a host for recombinant protein expression due to its high secretion capacity, its ability to metabolize methanol as its primary carbon source, its safety record as a source of biologics, and its extensive literature compared to other non-model yeasts [10, 11]. Thus, *K. phaffii* is an ideal chassis to rapidly implement changes designed to improve protein expression and secretion [4]. Indeed, recent work in *K. phaffii* has focused on systems-level analysis [12] and implementing design approaches of

*Correspondence: talva001@ucr.edu
[1] Department of Bioengineering, University of California, Riverside 92521, United States of America
Full list of author information is available at the end of the article

synthetic biology such as molecular parts lists and strain engineering [13, 14]. Such changes may accelerate product development and allow cheap, local production of pharmaceuticals [15, 16].

Identifying and relieving protein biogenesis bottlenecks is one strategy to improve yields of high-value, recombinant proteins [1, 17]. For secreted proteins expressed in *K. phaffii*, an early bottleneck is the translocation of newly made proteins from the cytoplasm into the lumen of the endoplasmic reticulum (ER) [18, 19]. Yeasts have multiple pathways for translocation, which use partially overlapping sets of biogenesis factors (reviewed in [2]). In the major pathway into the ER, translocation occurs through a membrane-embedded protein complex called the *sec* translocon. At least three major translocons exist in yeasts (the Ssh1 complex; two Sec61 complexes with, and without, Sec62p, Sec63p, Sec66p and Sec71p), which can accept proteins as they are synthesized by ribosomes (cotranslationally) or after synthesis of the polypeptide chain is complete (posttranslationally). Besides translocon architecture, co- and posttranslational pathways differ in their reliance on cytosolic molecular chaperones [20, 21]. Translocons bind hydrophobic amino acid motifs, called signal peptides, found at the amino termini of secreted proteins [22]. Some signal peptides are dependent upon a cytosolic factor, the Signal Recognition Particle (SRP), and the ER-bound SRP receptor to engage a translocon [23]; these tend to be longer or more hydrophobic than SRP independent signals [24, 25]. Binding of a signal peptide to a translocon opens the channel and allows the rest of the protein to pass into the lumen. In addition to secreted proteins, the *sec* translocon is a major point of entry for integral membrane proteins of the endomembrane system [26]. Integral membrane proteins that use a *sec* translocon require SRP for targeting to the ER over mitochondria [24].

For any production host, ribosomes, molecular chaperones, and *sec* translocons represent limited pools of resources that are distributed between heterologous proteins and the host proteome [27–29]. Unlike resources that are replenished enzymatically (like aminoacyl-tRNAs), ribosomes, translocons and chaperones only act on a single nascent chain at a time. While in use, they are sequestered and unavailable for other tasks. Although computational models that approximate these effects exist for bacteria [30], the complexity of eukaryotic translation is insufficiently understood to predict these allocations from transcriptomics alone. Accurate accounting of these resources could allow strains to be engineered in ways to relieve bottlenecks specific to a target. The secretome of *K. phaffii* has been characterized under several conditions [31], but the precise biosynthetic requirements of each protein remain unknown. Sequence features of secreted proteins, like glycosylation motifs, allow approximation of their direct biosynthetic costs such as ATP, carbohydrates, disulfide bonds, or GPI-anchors [32]. Per molecule costs can be coupled with measurements of gene expression to identify most expensive host proteins. Deletion of these proteins improves yields of secreted heterologous proteins in mammalian systems [33, 34]. However, while these analyses account for demands on global resources, they are limited by insufficient experimental data which links gene products to specific biogenesis subnetworks. For instance, overloading cotranslational translocons could limit secretory yields even if metabolic demands are met and posttranslational translocons are available. Quantification of global ribosome, cotranslational translocon and SRP use is available for *S. cerevisiae.* [24, 35, 36] However, these measurements are unavailable for other industrially significant species, including *K. phaffii.*

Which host proteins sequester the most biogenesis machinery in the early secretory pathway of *K. phaffii*? Which host genes produce the most nascent chains, competing for chaperones and sorting factors within the endomembrane system? To answer these questions, we quantified active translation globally and at the surface of the ER or mitochondria. Our analysis reveals the set of proteins that enter the secretory pathway cotranslationally and predicts the set that enter posttranslationally. In each set, we estimate demand for ribosomes and translocons. We distinguish between resources that act on a per nascent chain basis from machinery that is utilized based on elongation time.

Materials and methods

Strains and culture conditions

All experiments were performed using *Komagataella phaffii* GS115 (Invitrogen). For each Ribo-seq biological replicate, 500 ml liquid cultures of YPD (1% yeast extract, 2% peptone and 2% glucose) were grown to an $OD_{600\,nm}$ of 2 at 30°C with shaking in baffled 2 l flasks. Cells were harvested by vacuum filtration through a 0.8 μm filter. Immediately after filtering, cells were scraped off the filter using a chilled scoopula and submerged in a 50 ml conical tube containing liquid nitrogen. When indicated in order to match conditions of *S. cerevisiae* fractionated Ribo-seq data [35], cycloheximide (CHX) was added to 100 $\mu g\,ml^{-1}$ for 3 min prior to harvesting. CHX treatments longer than a few minutes can alter ribosome abundance near the start of transcripts [37]. Short incubations with CHX enhance targeting of translocation competent ribosome-nascent

chain complex while not perturbing non-secretory polysomes [36].

Lysis and subcellular fractionation

Cells were lysed in either soluble lysis buffer (50 mM MOPS, 25 mM potassium hydroxide, 100 mMolar potassium acetate, 2 mM magnesium acetate, 1 mM dithiothreitol and 100 $\mu g\,mL^{-1}$ CHX) or membrane lysis buffer (soluble lysis buffer with 1% Triton X-100). Lysis buffers for each sample were frozen by adding 2 ml dropwise to a 50 ml conical tube containing liquid nitrogen. For each biological replicate, $\frac{2}{3}$ frozen cells were mixed with 2 ml frozen soluble lysis and the remaining $\frac{1}{3}$ were mixed with 2 ml frozen membrane lysis buffer. Cell fractions were pulverized for 2 min in a 50 ml ball mill chamber with a single 2 cm steel ball (Retsch) and collected into 1.5 ml conical tubes. After thawing, lysates were centrifuged at 20,000×*g* for 10 min. Supernatants from samples lysed with membrane lysis buffer were collected and used as "total" fractions. Supernatants from samples lysed with soluble lysis buffer were collected and used as "soluble" fractions. The pellets from sample lysed with soluble lysis buffer were resuspended in 2 ml membrane lysis buffer and centrifuged. The supernatants were collected and used as "membrane" fractions. Triton-X 100 was added to 1% in soluble fractions, so that all three fractions were in equivalent buffers.

Ribo-Seq

Lysed samples were digested using 40 U of ribonuclease A (Ambion) for 1 h at room temperature. Digested samples were layered on a 10 to 50% sucrose gradient prepared in 50 mM Tris pH 7.5, 200 mM sodium chloride, and 2 mM magnesium acetate case using a Gradient Master (Biocomp). Gradients were centrifuged at 39, 000 rpm for 2.5 h in a TH-641 rotor (Thermo). After centrifugation, gradients were fractionated using a Piston Gradient Fractionator (Biocomp) and monosome peaks were retained. Total RNA was extracted using a standard phenol-chloroform method and alcohol precipitated. Ribosome protected footprints, corresponding to (18 nt to 34 nt), were excised from a TBE urea gel. RNA was collected from excised gel fragments using RNA gel extraction buffer (300 mM sodium acetate, 1 mM EDTA, and 0.25% SDS), precipitated, and resuspended in water containing 20 U/ml SUPERase•In (Invitrogen).

Purified fragments were used to prepare sequencing libraries as described in [38] with some modification. Linker ligations were allowed to proceed for 4 h, and afterwards, samples were pooled and purified by TBE-urea PAGE. The pooled library was depleted of ribosomal RNA using the Ribo-Zero Gold rRNA Removal Kit (Illumina), following manufacturer's instructions. Reverse transcriptions were performed using SuperScript II (Invitrogen). After circularization, PCR amplification and TBE PAGE purification, libraries were quantified using a Qubit 2.0 Fluorometer (Invitrogen) and sequenced using a HiSeq 4000 (Illumina.) Linker sequences were trimmed and libraries were demultiplexed using Cutadapt [39].

Long read RNA sequencing

Cells were grown in YPD at 30 °C with agitation to an $OD_{600\,nm}$ of 2 and harvested by centrifugation. Total RNA was obtained using a Direct-Zol kit (Zymo Research). Cells were vortexed with glass beads for 2 min during incubation with TRI reagent. After purifying RNA, a library was prepared using a PCR-cDNA kit according to manufacturer's instructions (SQK-PCS109, Oxford Nanopore Technologies) and sequenced using a minION R9.4.1 flow cell. Base calling was performed using Guppy (Oxford Nanopore Technologies).

Transcript assembly

A novel transcriptome was assembled using data derived from Ribo-Seq, long-read RNA-Seq, and a prior genome sequence of strain GS115 [40]. A flowchart of the annotation pipeline is provided in Figure S2c. Ribo-seq reads and long reads were aligned to the reference genome using HISAT2 [41] and Minimap2 [42] respectively. Stringtie version 1.3.6 was used to assemble transcripts from Ribo-seq data, with reads mapping to each strand processed separately [43]. Pinfish was used to assemble transcripts from long reads (Oxford Nanopore Technologies). After transcript assembly, PASA [44] was used to combine the Stringtie and Pinfish models into a single transcriptome. Transdecoder [45] was then run twice: first, to identify candidate coding regions with PASA model with a lower limit of 100 amino acids, and second, to identify coding regions in just the Stringtie model with a lower limit of 40 amino acids. The latter run has a reduced risk of misannotating start codons in the 5′-UTR. Transdecoder annotated transcripts from $Transdecoder_{PASA}$ were used to train GlimmerHMM [46] and CodingQuarry [47], which were used to provide de novo predictions in the genome. EVidenceModeler [48] was used to incorporate predictions from PASA, $Transdecoder_{Stringtie}$, $Transdecoder_{PASA}$, GlimmerHMM and CodingQuarry. File processing, UTRs, and tRNAs annotations were provide by the update utility in the Funannotate package [49].

Mapping of ribosome protected reads to codons and masking

Ribo-seq reads were mapped to the genome of *Komagataella pastoris* GS115 [40] using HISAT2 [41, 50].

Alignments were converted from SAM to sorted and indexed BAM files using Samtools and only included reads with mapping quality threshold of 60 [51]. Mapped reads were loaded into R using the GenomicAlignments package from Bioconductor[52] and converted to their 3′ end positions before determining p-site offsets. P-site offsets were determined using the RiboProfiling package in Bioconductor [53]. Each read was mapped to a single codon. Masking files were created by first parsing the coding sequence (CDS) annotation file associated with the reference genome into a fasta file simulating every possible 28 nt combination (approximate length of a ribosome protected mRNA fragment). This fasta file was then aligned to reference genome twice, once to only include reads with mapping quality greater than or equal to 60 (unambiguously assigned), and another to include all reads (ambiguously assigned). Both alignment files were used to generate reads per codon per gene (RPCPG) data tables. The unambiguously assigned reads were subtracted from ambiguously assigned reads and codons with a nonzero difference were included in mask. The first and last five codons in genes' open reading frames (ORFs) were masked to correct for variable read quality at the beginning and ending of transcripts inherent to Ribo-Seq [54].

Metagene correction and quantification of metabolic demand

Read counts were normalized at the codon level using a metagene analysis that provides a global profile for each data set. First, for each ORF, reads at each codon position were scaled by the average reads per codon mapped ORF. Then, for codon position, either a mean or median value was calculated from all ORFs using the following scheme: for positions 1 to 100, a rolling mean with a window of 10 codons; for positions 100 to 1000, a rolling mean with a window of 100; for positions 1000 and onward, a rolling median with a window of 1000. In calculating corrected transcripts per million (cTPM), codon read counts were scaled by dividing the metagene-derived value at that position and normalized by their pseudo gene lengths (theoretical gene length minus number of masked codons) and a per million scaling factor unique to each data set. In calculating ribosomes per million (cRPM), a ribosome scaling factor was created for each gene by dividing the sum of the metagene-derived values at all codon positions by the sum of smoothed reads per codon with the mask applied (a gene with zero masked codons will have a ribosome scaling factor equal to one, while a gene that contains masked codons will have a scaling factor greater than one). The ribosome scaling factor is multiplied by unmasked gene read counts and normalized by a per million scaling factor unique to each data set to give RPM. Membrane enrichment is quantified for each gene as the $\log_2$ ratio of membrane cTPM scores or total cTPM scores to soluble cTPM scores.

Classification and annotation of ORFs

Gene names were hierarchically assigned to novel *K. phaffii* transcripts through homology. Firstly, transcripts were assigned names inherited from *S. cerevisiae* using BlastP [55] with an expected value less than 1e−5. For genes that were not predicted to be homologous, gene names were assigned common names using EggNOG 4.5 [56] using a taxonomic scope limited to ascomycetes. Genes that did not share homology with *S. cerevisiae* or known ascomycetes were assigned names inherited from *K. phaffii* GS115 [40] using BlastP with expected values less than 1e−5. Novel genes that were not assigned names using methods above were named after the moniker given during transcript assembly.

ORFs were classified by function, cellular location, and sequence features using various prediction software. Functions were assigned ontologically using clusters of orthologous groups (COG) and were prepared using EggNOG 4.5 [56]. Vironoi tessellations were created to quantitatively map the biosynthetic composition of these functions using COGs and expression metrics derived from Ribo-Seq cTPM [57]. DeepLoc was used to predict the subcellular localization associated with ORF products [58]. Sequence features such as signal sequences, transmembrane domains (TMD), and GPI anchors were identified using SignalP 5.0 [59], TOPCONS [60], and predGPI [61] respectively.

S. cerevisiae analysis

Ribo-seq data for total protein synthesis were taken from [62], and data obtained from soluble or membrane-bound ribosome fractions were obtained from [35]. All data were processed in the same way as *K. phaffi* using the S288C reference genome R64-2-1 [63].

Results

Ribo-seq and long-read RNA-seq improve open reading frames annotations

We sought to globally quantify several aspects of protein synthesis in *K. phaffii* GS115. We asked which genes were responsible for sequestering limited biosynthetic resources, such as ribosomes and ER translocons. We also asked which genes were responsible for producing the most nascent chains, which is critical for predicting amino acid usage, as well as modifications that act on a per chain basis (i.e., N-terminal acetylation, GPI anchoring, vesicular sorting). Ribo-seq

provides a snapshot of protein translation, allowing us to answer both of these questions [64]. It is a high throughput sequencing technique used to infer ribosome abundance at each codon of each transcript. In Ribo-seq, a non-specific ribonuclease generates 20 to 22 nt or 28 nt to 30 nt "footprints" of ribosome-protected mRNA depending on the translational conformation of the ribosome [65], which are then sequenced. We performed a series of Ribo-seq experiments to capture global translation and translation on the surface of organelles (Fig. 1). Our data sets captured footprint lengths from 15 to 42 nt (Additional file 1: Figure S1a). Nearly all (99%) footprints mapped within open reading frames (ORFs). Our profiling data also indicate active translation through the appearance of three nucleotide periodicity in read depth that is preserved across the transcriptome (Additional file 1: Figure S1b).

We noticed that ribosome-protected read patterns were often inconsistent with prior annotations of open reading frames (Additional file 2: Figure S2a). At many loci, Ribo-seq appeared to indicate that translation began at an alternate start codon. Inaccuracies in ORF structure are problematic, since the length of a reading frame is a critical parameter used for quantifying translation and the position of the start site is used in correction using global profiles (see below). We therefore sought to improve the GS115 annotation using Ribo-seq. Several methods that rely solely on Ribo-seq to annotate structure rely on the three nucleotide periodicity of reads to define reading frames [66]. They require substantial coverage for each gene; however, sparse Ribo-seq coverage could still support re-annotation if it were treated like stranded RNA-seq data. Moreover, de novo open reading frame predictors can be trained using verified translational start sites, and so improving the accuracy of annotations for a subset of the transcriptome was expected to improve overall prediction accuracy. We therefore adapted consensus methods used in gene prediction and annotation with standard RNA-seq data, with optimizations for fungi [48, 49]. Our approach uses Ribo-seq to construct transcript models, which are then used to train several de novo annotators.

Like other yeasts, *K. phaffii* has short intergenic sequences, leading to overlapping untranslated regions (UTRs), even on transcripts encoded on the same DNA strand. As a result, methods that construct transcripts from short-read sequencing merge data from adjacent genes into a single transcript. We therefore collected long-read data using Oxford Nanopore PCR-cDNA sequencing and developed a pipeline to integrate Ribo-seq, long-read RNA-seq, and de novo gene prediction (Additional file 2: Figure S2b, c). Our annotation is provided as Additional file 3. ORFs that were fully covered by Ribo-seq data were allowed to be as short as 40 amino acids, increasing the number of annotated genes compared to other annotations of *K. phaffii* (Table 1) [40, 67, 68]. Homologs between our annotation and prior annotations are provided as Additional file 4. Our annotation adjusted the translational start site of about 10% of ORFs compared to each previous model. Overall, Ribo-seq reads were mapped to 5303 genes in *K. phaffii* in the assembly presented here. We have named genes based on homology to prior annotations, to *S. cerevisiae* and to other ascomycetes.

Translational landscape of *K. phaffii*

Each read in Ribo-seq originates from a translating ribosome. Thus, by comparing the distribution of reads, we can answer our first question and identify which transcripts sequester ribosomes and ribosome-associated factors, like the *sec* translocon. As a method to predict the abundance of polypeptide chains, Ribo-seq has greater sensitivity than mass spectrometry, and more closely matches measurements of protein

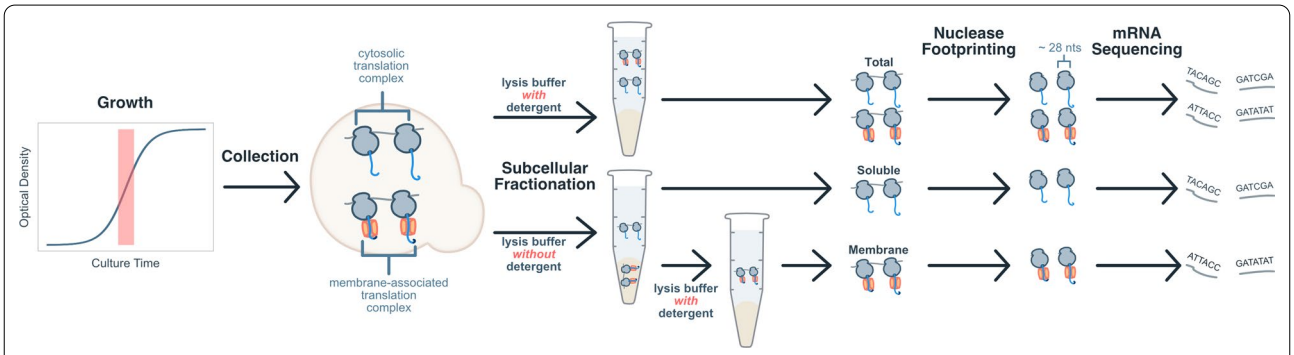

Fig. 1 Overview of Ribo-seq and subcellular fractionation. Ribosomes (grey) bound to a translocon (red) are only solubilized in the presence of detergent. The total sample has footprints originating from both membrane-bound and free-floating ribosomes. The soluble fraction is enriched in footprints from free-floating ribosomes. The membrane fraction is enriched in footprints from membrane-bound ribosomes

abundance than RNA-seq [69]. To answer our second question, the number of nascent polypeptide chains produced per unit time can be approximated using a modified form of the transcripts per million (TPM) metric used in RNA-seq. TPM has advantages over other metrics (RPKM or FPKM) for its intuitive interpretation during differential analysis and for its congruence with proteomics [70, 71]. In RNA-seq, reads are generally long enough to be unambiguously mapped to the transcriptome, and they can be assumed to equally cover a transcript. In Ribo-seq, however, these assumptions do not hold, and biases due to ambiguous mapping and unequal coverage must be corrected.

Ribosome protected fragments are small, 22 nt to 30 nt, and may map to multiple mRNA sequences when the transcriptome contains homologous stretches. Ambiguously mapped reads can be handled in one of several ways, often with shortcomings. Discarding multi-mapped reads [72–75] depreciates read counts for highly expressed genes. Randomly assigning reads to ORFs with equivalent percentage of alignment [64, 76, 77] overestimates read counts for lowly expressed genes. Here, we adapt the method of Taggart et al. [62], who used computational masks to exclude homologous segments of the predicted transcriptome. We calculated a mask over the *K. phaffii* transcriptome accounting for all possible 28 nt reads, excluding 3% of codon positions available. To estimate gene expression via TPM, reads must be scaled by ORF length. Unlike discarding or randomly assigning reads, masking adjusts the gene length to reflect mRNA positions available for analysis. However, masking alone is insufficient because ribosome protected reads are not evenly distributed across transcripts.

Ribosome-protected reads are more abundant near the 5′ end of ORFs [64, 78]. This effect may be due slower elongation rates at the beginning of translation [79] or abortive translation [62]. Regardless of the mechanism, the positional bias is observed in nearly every transcript and results in a global read profile that is conserved across the translatome (Fig. 2a). As a result, estimates of the expression of short ORFs will appear inflated (and long ORFs deflated), since only the ribosome-rich region of the global profile is sampled. We again adapt the method of Taggart el al. [62], where the positional bias is removed by scaling reads at each codon by the empirical global profile. (Fig. 2b). We use corrected TPM (cTPM), with masking and scaling, as a measure of the rate at which nascent chains are produced. For example, transcripts of *RPL5* and *YEF3* display similar numbers of ribosomes at the start of their ORFs (Fig. 2c), suggesting similar initiation rates. However, because *YEF3* is a longer ORF, its standard TPM is smaller than the TPM of *RPL5*. Here, we assume that if *RPL5* were as long as *YEF3*, then its translational profile will be similar to the global profile, resulting in similar cTPM scores.

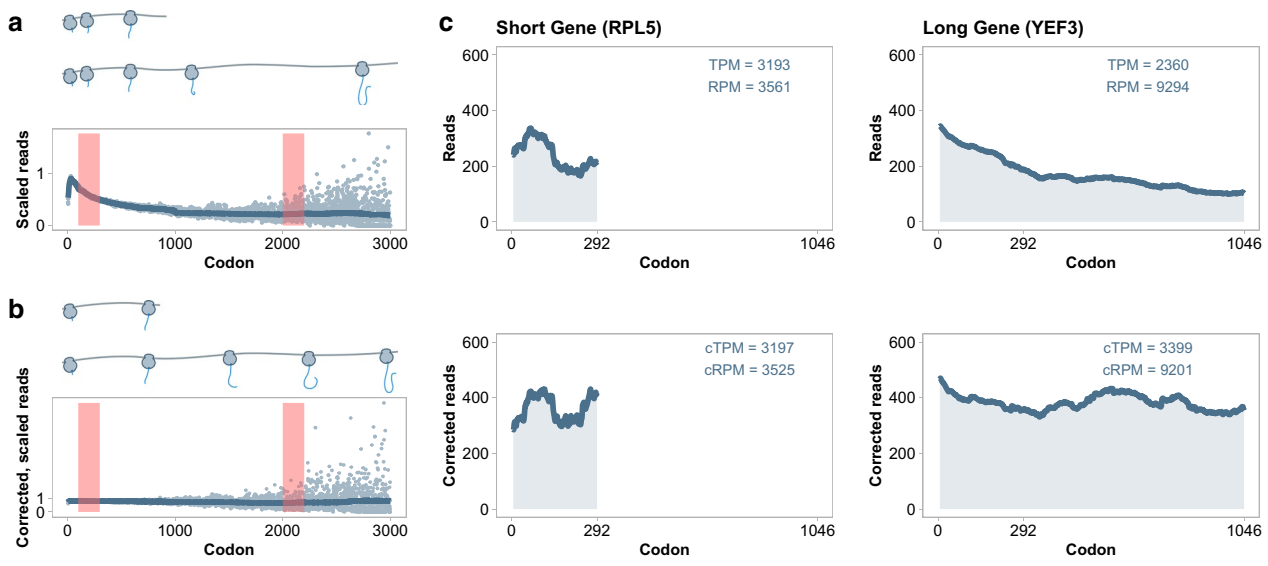

Fig. 2 Corrections applied to Ribo-seq data. **a** Ribosome-protected read counts at each codon were scaled by the total reads mapping to the ORF. Dots represent individual codons, and the line represents a composite of rolling means and medians see Methods. Regions in orange are the same width and are used to demonstrate that masked codons at the beginning of ORFs have a greater influence of calculated expression than masked codons at the end of ORFs. **b** Data from **a** after metagene correction. **c** Comparison of ribosome-protected reads per codon for highly expressed genes of different length. TPM for *RPL5* gene is approximately 135% greater than TPM for *YEF3* while producing approximately 38% as many ribosome-protected reads. After metagene correction cTPM scores are similar preserving the same difference in ribosome sequestration

While cTPM estimates the number of nascent polypeptide chains, it does not answer our question regarding ribosome sequestration. Longer transcripts sequester a greater number of ribosomes in order to produce the same number of nascent chains as a shorter transcript. If ribosomes accumulate near the start codon in vivo, then it is important to include this effect while measuring allocation. cTPM, therefore, is an inappropriate metric. If ribosome-protected reads could be unambiguously mapped to the transcriptome, then simple read counts estimate ribosome usage per gene. However, when masking is applied, the position of the mask becomes important (Fig. 2a, b). Two masks of the same length, applied at different positions, will hide different amounts of ribosomes based on the global profile. To correct for this, we introduce a ribosome scaling factor that accounts for masking of each gene. The factor represents the fraction of ribosomes expected to be observed when the gene-specific mask is applied to the global translational profile. We generate a new metric for each gene, corrected ribosomes per million (cRPM), which is practically equivalent to reads per million (RPM) in standard RNA-seq. In our example in Fig. 2c, cRPM and RPM are almost identical, as expected since there are no masks applied to *RPL5* or *YEF3*. Read counts, cTPM and cRPM for each gene in each dataset are provided as Additional file 5.

After applying corrections, we find that the majority of nascent chains synthesized in *K. phaffii* are from genes involved in translation, ribosomal structure and biogenesis (see Table 2 and Fig. 3a), as expected for log-phase growth. The majority of nascent chains encoded by genes of unknown function are predicted to be extracellular, where they are likely components of the cell wall. We consider endomembrane lumenal and secreted proteins to be those with (i) predicted N-terminal signal sequences, (ii) are not predicted to be localized to the mitochondria, and (iii) contain less than or equal to one transmembrane domain, as these are frequently GPI anchors. Some single-pass, type I transmembrane proteins will be misannotated by this definition. The number of genes containing these predictive features and the relative percentage of nascent chains they produce are summarized in Table 2. A majority of nascent chains for genes containing a signal sequence also contain GPI anchors, suggesting that this structural class represents the majority of products that will be processed by the secretory pathway.

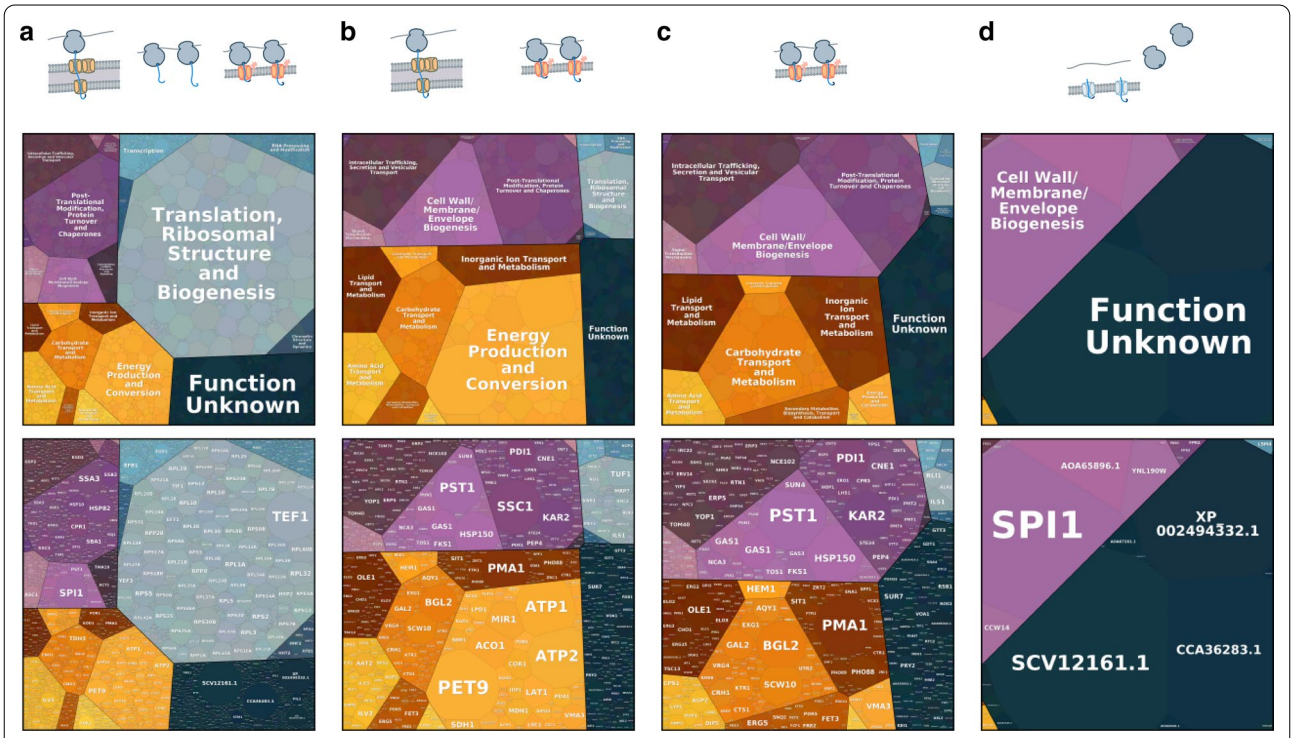

Fig. 3 Protein expression and trafficking in *K. phaffi*. Tessellations are calculated using cTPM from the total fraction of a CHX treated culture and represent relative quantities of nascent chains produced from each gene. **a** Nascent chains produced by all ribosomes. **b** Nascent chains from genes showing twofold membrane enrichment. This includes mitochondrial and ER destined proteins. **c** Nascent chains from genes showing twofold membrane enrichment that are not predicted to be mitochondrial. **d** Nascent chains from genes showing less than twofold membrane enrichment but with a predicted ER signal sequence

Biogenesis demands in the early secretory pathway

We next investigated the global demands for machinery needed for translocation into the ER. Subcellular fractionation was used to separate membrane-bound ribosomes from free floating, soluble ribosomes. Membrane-bound ribosomes were detergent solubilized, and then samples from both soluble and membrane fractions were subject to Ribo-Seq (Fig. 1). As in *S. cerevisiae*, libraries derived from the membrane fractions are enriched in ribosome-protected footprints originating from transcripts that encode proteins destined for the ER or mitochondria [35] (Fig. 4). Membrane enrichment scores were calculated as the log_2 ratio of cTPM for membrane and soluble fractions and were reproducible (Additional file 6: Figure S3a and provided in Additional file 5). The magnitude of membrane enrichment scores depends on the efficiency of fractionation, and if a gene falls below the diagonal line in Fig. 4, it will have a negative enrichment score. As in *S. cerevisae*, membrane enrichment scores are limited by the length of the ORF when transcripts encode signal-sequence bearing proteins [35, 36] (Fig. 5). This effect is due to a kinetic competition between trafficking rate and translation elongation rate. Fig. 5 also reveals that a membrane enrichment score of 2 effectively separates two populations, and so we define genes with scores greater than 2 as cotranslationally translocated into either the ER or mitochondria. The set of cotranslationally translocated nascent polypeptides is enriched for those involved in energy production and conversion, cell wall and membrane biogenesis, and various transporters (Fig. 3b). To assess entry into the ER, we filtered out transcripts encoding proteins predicted to localize

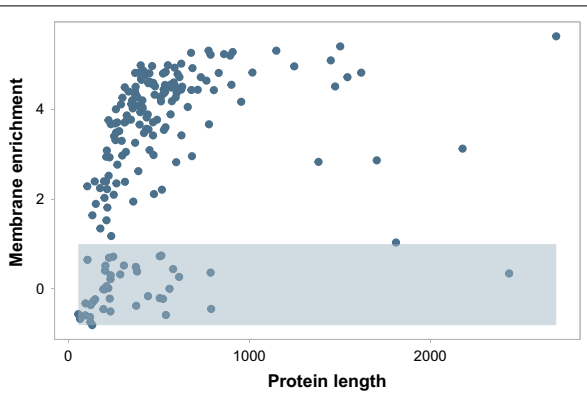

Fig. 5 Nascent peptide length and membrane enrichment for secreted, lumenal, or GPI-anchored proteins. Proteins have a predicted N-terminal signal sequence. GPI anchors are included. The shaded box is drawn over genes with less than twofold membrane enrichment, which are considered posttranslationally targeted

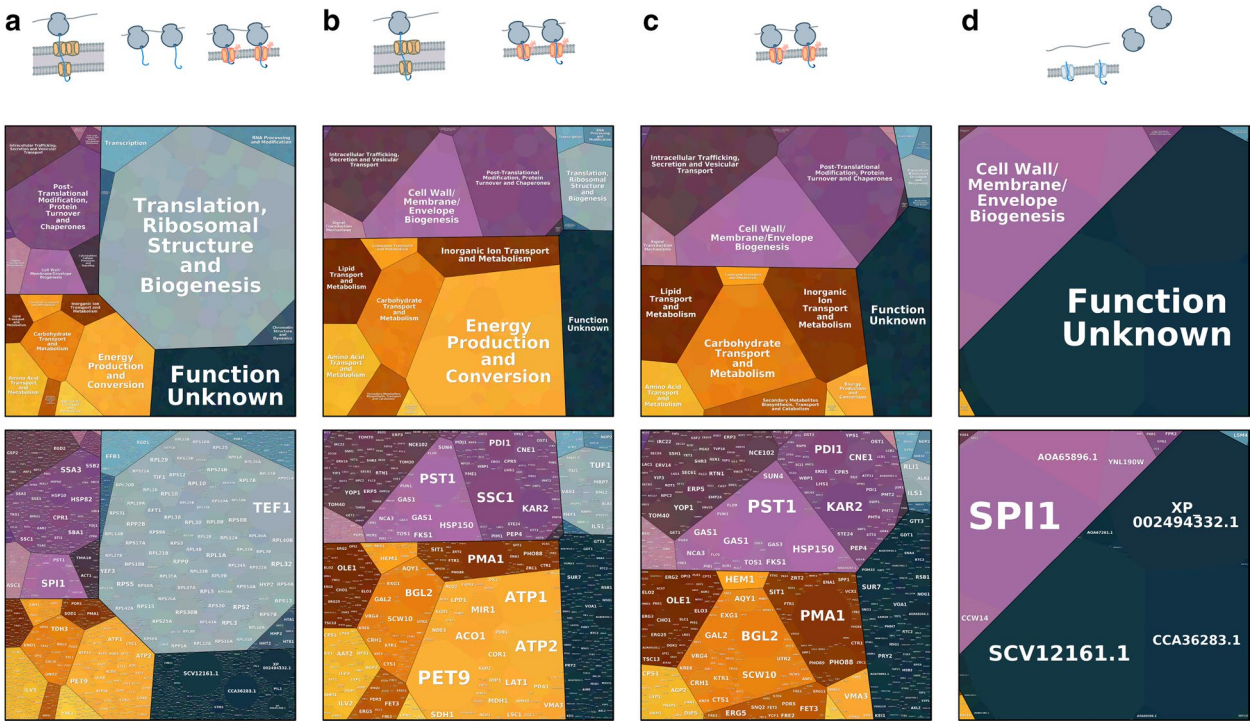

Fig. 4 Comparison of translation from samples of membrane-bound and soluble fraction. Values are calculated using fractions obtained after incubation with CHX

Table 1 Comparison of ORF annotations

Annotation[a]	Total ORFs	Homologs[b]	Length differences[c]
Current study	5329		
GS115 (PRJNA304976)	5064	5035	514
GS115 (PRJEA37871)	5040	5100	697
CBS7435 (PRJEA62483)	5291	5198	604

[a] NCBI bioproject numbers located in parenthesis

[b] BlastP matches from current study to prior study

[c] Number of homologs with different predicted lengths

in the mitochondria by DeepLoc (Fig. 3c). Finally, we define proteins that enter the ER through a posttranslational *sec* translocon as those having a predicted N-terminal signal sequence and less than twofold membrane enrichment (Fig. 3d). Posttranslationally trafficked membrane proteins rely on other mechanisms, such as the GET pathway [22].

A more diverse group of proteins enter the ER through cotranslational translocons than those that enter posttranslationally (Fig. 3c,d and Table 3). While the diversity of functions for proteins that enter the ER posttranslationally is relatively small (mostly unknown function and then cell wall and membrane biogenesis), we find that posttranslational translocation handles a majority of total nascent chains entering the ER. These genes encode primarily small proteins such as SCV12161.1p or cell wall proteins processed with GPI-anchors, such as Spi1p. Although its function is unknown, Spi1p is also predicted to be GPI-anchored, and both *SPI1* and *SCV12161.1* produce among most nascent proteins within the cell under conditions tested here (Fig. 3a). We then classified the genes of unknown function that entered the ER by their predicted final location. The majority of these gene products, approximately four fifths, are predicted to be localized extracellularly and have an unusual discrepancy between their relative ribosomal usage, nascent chains produced, and average gene length compared to unknown genes predicted to localize elsewhere (Additional file 7: Table S1).

Comparing the translational landscape between *K. phaffii* and *S. cerevisiae*

Of the 5329 *K. phaffii* genes annotated here, 73% have a homolog in *S. cerevisiae*. Unlike *K. phaffii*, *S. cerevisiae* is thought to have undergone a whole-genome duplication, and so many *S. cerevisiae* genes have paralogs [80]. The influence of paralogy is evident in how these two species allocate translational throughput. We calculated of cTPM and cRPM in *S. cerevisiae* (Additional file 8) using prior data acquired under similar growth conditions [35, 62]. The overall distribution of cTPM by ontological category is similar between species (Additional file 9: Figure S4). Under the conditions tested here (glucose-containing rich media), *TEF1*, encoding translational elongation factor 1 alpha, is the most translated protein in *K. phaffii*. The TEF1 promoter is used to drive constitutive expression in *K. phaffi* [81], and our results suggest that the native TEF1 ORF is translated more than the ORFs linked to other promoters used for expression in glucose, such *GAP* (here, *TDH3*) and *PGK1* [11]. *S. cerevisiae* generates a similar amount of nascent chains to the same function, but it does so using a combination of its paralogous genes *TEF1* and *TEF2*. Unsurprisingly, Crabtree-positive *S. cerevisiae* generates three times more polypeptides involved in glycolysis and fermentation than *K. phaffii* (e.g., *ENO1/2, GPM1, FBA1, TDH2/3, TPI1, PGK1, PDC1, ADH1*).

Indeed, these two species also show divergence in energy production with regards to cotranslational mitochondrial import (Fig. 6). Our subcellular fractionation

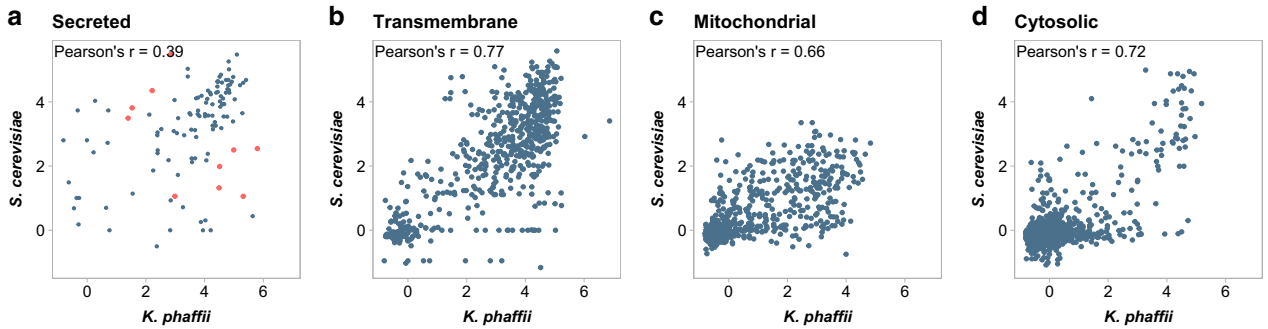

Fig. 6 Correlation of membrane enrichment scores between species. Scores are deteremined using the membrane-bound and soluble fractions of ribosomes from cultures treated with CHX. **a** Enrichment scores restricted to signal sequence bearing proteins. Contrast dots represent genes found in Table 2. **b** Enrichment scores restricted to non-mitochondrial transmembrane proteins. **c** Enrichment scores restricted to mitochondrial proteins. **d** Enrichment scores restricted to cytosolic proteins

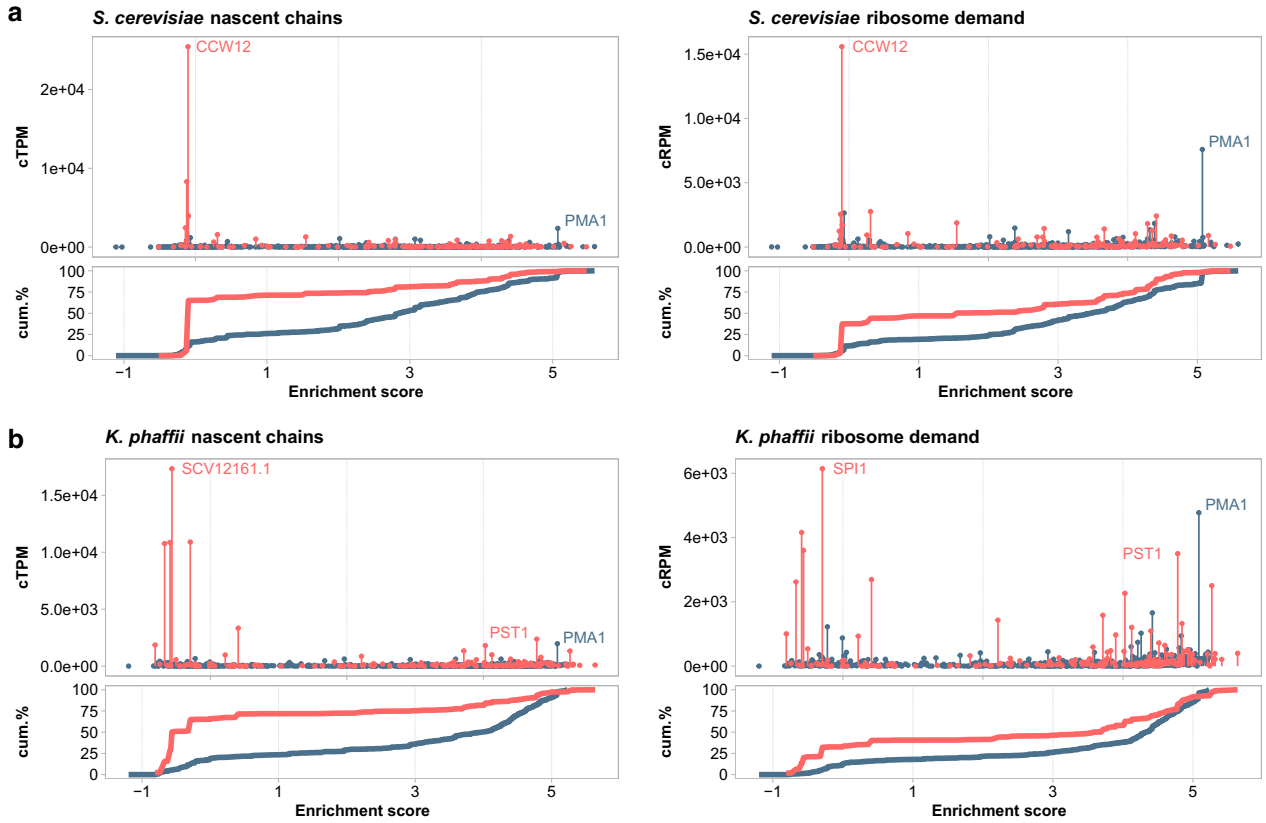

Fig. 7 Demands imposed on secretion pathway. Blue lines represent membrane proteins and orange lines represent secreted, lumenal or GPI-anchored proteins. **a** Demands in *S. cerevisiae*. **b** Demands in *K. phaffii*

assay recovers all membrane-bound ribosomes, including those attached to the mitochondria. A greater number of nuclear-encoded mitochondrial proteins undergo membrane-localized translation in *K. phaffii*. Recovery of membrane associated mRNA strongly depends on active translation [35]. Therefore, less active translation of mitochondrially destined proteins may become reflected in lower membrane-enrichment scores.

We next asked whether ER translocation pathways are conserved between the two species. Between homologs, membrane enrichment scores correlated with a Pearson's r of 0.85 (Additional file 6: Figure S3b). Genes encoding transmembrane proteins or cytosolic proteins which lack ER or mitochondrial targeting sequences had the highest correlation. Signal-sequence bearing proteins, including GPI-anchored proteins, however, had lower correlation (Fig. 6a). There were several genes which only showed cotranslational membrane enrichment in one species, and in some cases this was due to loss of a signal peptide in one of the homologs. The ten genes that showed the greatest difference in magnitude, while still showing evidence for membrane enrichment in both species, are reported in Table 4. Notably, this list includes *PDI1*, encoding an ER lumenal protein-disulfide isomerase that is essential for ER homeostasis. Mitochondrially localized proteins have greater membrane enrichment in *K. phaffii*, which may be related to the greater use of aerobic respiration compared to *S. cerevisiae* (Fig. 6c).

Finally, we explored the relationship between the burden imposed by production of polypeptide chains (cTPM), ribosome demand (cRPM) and translocation pathway (membrane enrichment score) for ER destined proteins within the two species (Fig. 7). In *S. cerevisiae*, most of these chains originate from a single gene, *CCW12*, while in *K. phaffii*, there are a wider variety of genes, with *SCV12161.1* being the most dominant. Strikingly, posttranslational targeting is used for about two-thirds of lumenal, secreted or GPI-anchored nascent chains in both species. *K. phaffii*, however, is distinguished by at least one major cell wall protein, Pst1p, which enters the ER cotranslationally. In both species, Pma1p is the dominant membrane protein passing into the ER. In terms of ribosome sequestration, the trend reverses; cotranslational translocation is responsible for sequestering two thirds of ribosomes used to produce secreted or GPI-anchored proteins. While *PST1* yields

slightly more nascent chains than *PMA1*, *PMA1* is more than twice as long as *PST1* and sequesters 1.36 times more ribosomes. Thus, *PMA1* represents a significant burden to the secretory systems of both *S. cerevisiae* and *K. phaffii* as it is predicted to sequester more ribosomes, cotranslational translocons, and lumenal chaperones to synthesize and transport nascent chains into the ER.

Discussion

The yields of engineered, recombinant proteins are restricted by bottlenecks in biogenesis [1]. Certain bottlenecks are metabolic, including insufficient ATP or other high-energy compounds, nucleotides for mRNA synthesis, amino acids, carbohydrates for glycosylation, and reducing equivalents. A promising systems-level approach to remove bottlenecks is to identify and delete host proteins with the greatest demand for metabolic resources. Indeed, the Lewis lab has elegantly demonstrated in CHO cells that deleting expensive proteins (in terms of ATP equivalents) increases the yield of heterologous secreted proteins [33, 34, 82]. Similar modeling of metabolic demand has been performed by the Nielsen lab for the secretome of *S. cerevisiae* [32]. Other bottlenecks are due to insufficient cellular protein biosynthetic machinery, such as polymerases, ribosomes, translocons, and molecular chaperones. Focusing on metabolic demand will likely relieve pressure on machinery with tightly coupled–and therefore accurately predicted–energetic requirements (e.g., cycles of translation elongation by the ribosome). However, it only approximates demand for chaperones and translocons, which gate entry into the ER. Compared to tightly coupled complexes, chaperones and translocons are ambiguous in their energetic demand. Chaperones perform cycles of binding and rebinding that depend on the folding pathways of client proteins [83]. Translocation into the ER is driven by ATP-hydrolyzing chaperones, translation elongation, or a combination of the two in a client dependent manner [84, 85]. Engineering of the early secretory pathway, such as the optimization of signal sequences for protein targeting [86] and reducing the effect of the ERAD system [19], provides varying degrees of success. These approaches are contingent on the complexity of the protein product and must be empirically optimized [87, 88]. Our data and analysis may augment these efforts by accounting for capacity of translation, co- and post-translational translocation.

Despite the ability of Ribo-seq to accurately quantify gene expression, our study has several caveats that limit interpretation. First, we have only considered yeast undergoing log phase growth in liter scale, aerated shaking cultures using rich media. This design enabled comparison to several published data sets using *S. cerevisiae* that were collected under identical conditions [35, 62]. We chose strain GS115, a commonly used commercially available strain that is a histidine auxotroph (*his4*). Even under rich media with abundant extracellular histidine, this auxotrophy may influence gene expression compared to strains which supply *HIS4*. Future work involves quantifying demands at industrial scale in stirred bioreactors under induction of a heterologous protein. Second, we assume that elongation rates are relatively constant across genes. However, if the elongation rate is altered for a transcript, it may result in greater or fewer ribosome protected reads. We argue that on the whole, our assumption is valid, given that Ribo-seq accurately predicts mature protein stoichiometry [62, 89]. Third, Ribo-seq does not account for protein degradation; indeed, some proteins are cotranslationally ubiquitinated [90]. Our results should therefore not be interpreted as revealing steady-state protein levels in *K. phaffii*. However, our goal was to quantify the costs of protein synthesis, and so we argue that Ribo-seq is a more appropriate tool than mass spectrometry. Despite these limitations, our approach allowed us to interrogate protein translocation into the ER.

Most secreted proteins, including high-value targets like antibodies, will enter the ER via a *sec* translocon [2]. The translocon subunits Sec62p, Sec63p, Sec66p and Sec72p are required for the translocation of certain proteins, particularly those with shorter or less hydrophobic signal peptides [21, 25, 36]. Molecular chaperones are also implicated in protein translocation, through binding of proteins in the cytoplasm (Ssa1p) [20] or the ER lumen (Kar2p) [84]. However, many gene products are able to associate with more than one class of translocon [25, 36]. In addition, while recent structural work suggests that the heptameric Sec61 complex cannot directly bind a ribosome [91, 92], there is a preponderance of evidence demonstrating that the proteins dependent on this complex are translated at the ER membrane [24, 35, 36, 93, 94]. Further, even if a protein does not strictly require particular machinery, like SRP, it may nonetheless sequester it in vivo, reducing availability for proteins that do require these factors [35, 93]. Because of these complexities, it is unsurprising that it has remained difficult to precisely tune a translocon for a specific engineered protein. Rather, optimization will likely require understanding the needs of the target, what the target will sequester, and how this will relate to the balance of resources in the host.

Our calculations for nascent chains produced, ribosomes used, and predicted translocation pathways suggest that each gene presents a unique combination of challenges to the cellular biosynthetic capacity. For instance, long, cotranslationally translocated proteins

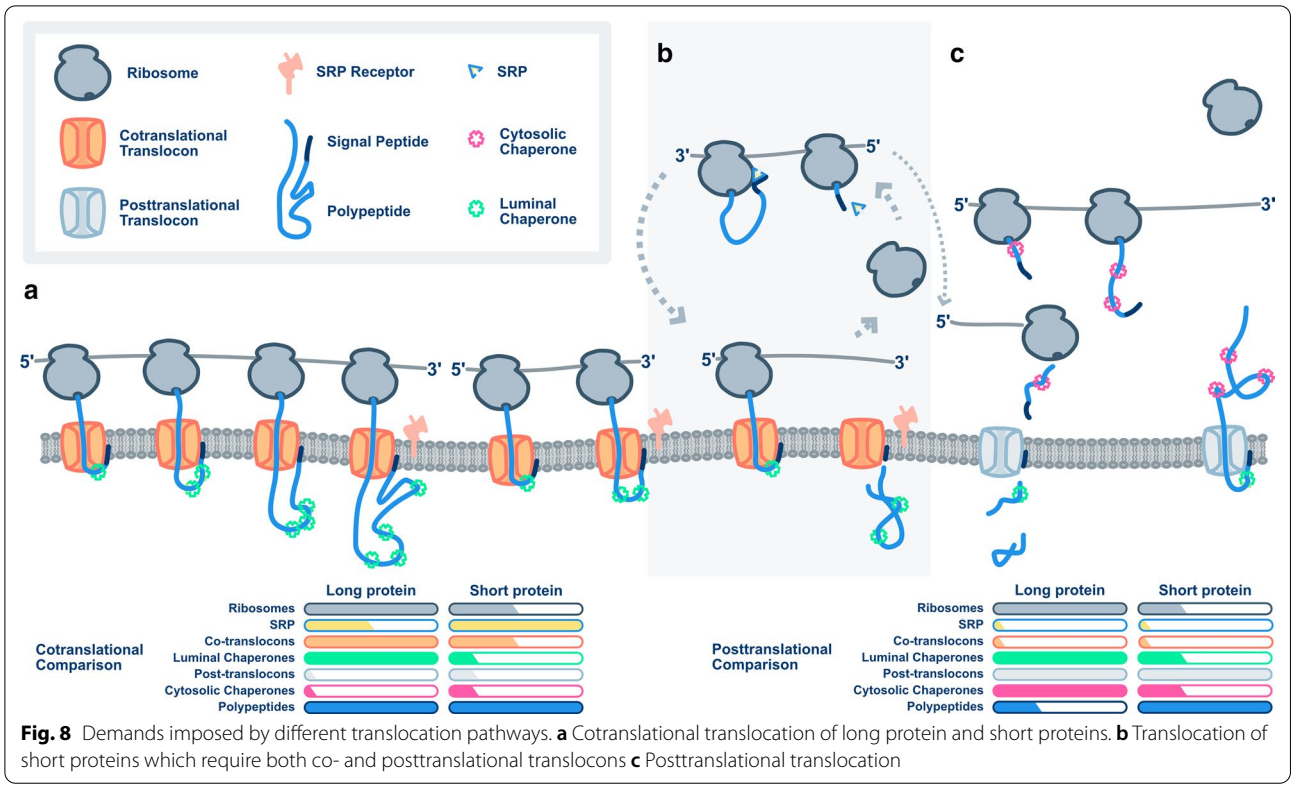

Fig. 8 Demands imposed by different translocation pathways. **a** Cotranslational translocation of long protein and short proteins. **b** Translocation of short proteins which require both co- and posttranslational translocons **c** Posttranslational translocation

Table 2 Nascent chains produced in *K. phaffii*

	Nascent chains (%)[a]	Genes (n)
Ontological functions		
Translation, rbosomal structure and biogenesis	44.0	366
Function unknown	11.0	1602
Post-translational modification, protein turnover and chaperones	9.0	409
Energy production and conversion	8.0	207
Intracellular trafficking, secretion and vesicular transport	4.0	382
Carbohydrate transport and metabolism	3.0	218
Cell wall/membrane/envelope biogenesis	3.0	85
Amino acid transport and metabolism	3.0	191
Transcription	2.0	355
RNA processing and modification	2.0	242
Predicted features of ER destined proteins		
Lumenal and secreted proteins[b]	8	266
GPI Anchors	79[c]	117
Transmembrane proteins[d]	7	960

[a] Nascent chains are percentage of the total cTPM represented by each category

[b] Total number of genes with an N-terminal signal sequence and may include a GPI anchor

[c] Percentage of nascent chains containing signal sequences that also contain a predicted GPI anchor

[d] Transmembrane proteins either have no signal sequence but one transmembrane domain (TMD), or two or more TMDs

will impart little demand on cytoplasmic chaperones, but will sequester ribosomes, translocons, and lumenal chaperones for extended periods of time (Fig. 8a). However, because of sustained translation on the surface of the ER, fewer instances of SRP targeting are required. A shorter cotranslational protein will require fewer ribosomes,

Table 3 Comparison of translocon demands by ontological function

	Genes (n)	Nascent[a] (%)	Ribosomes[b] (%)
Cotranslationally translocated[c]			
Function unknown	261	8.00	11.0
Cell wall/membrane/envelope Biogenesis	41	7.00	12.0
Post-translational modification, protein turnover and chaperones	89	7.00	12.0
Carbohydrate transport and metabolism	114	7.00	9.0
Intracellular trafficking, secretion and vesicular transport	95	6.00	7.0
Inorganic Ion transport and metabolism	82	5.00	10.0
Lipid transport and metabolism	72	4.00	5.0
Posttranslationally translocated[d]			
Function unknown	30	36	11.0
Cell wall/membrane/envelope Biogenesis	10	15	10.0
Post-translational modification, protein turnover and chaperones	5	0	0.0

[a] Calculated as percent of total cTPM for all proteins predicted to be ER destined

[b] Calculated as percent of total cRPM for all proteins predicted to be ER destined

[c] Proteins with greater than twofold membrane enrichment and not predicted to be mitochondrial

[d] Proteins with less than twofold membrane enrichment and not predicted to be mitochondrial and contained a predicted signal sequence

Table 4 Membrane enrichment for secreted, lumenal and GPI-anchored proteins in *K. phaffii* and *S. cerevisiae*

Gene	Product	*K. phaffii*	*S. cerevisiae*
Increased enrichment			
FLO9	Lectin-like protein, flocculin (isoform 2)	5.32	1.06
ZPS1	Putative GPI-anchored protein	5.80	2.54
SGA1	Sporulation-specific glucoamylase	4.49	1.32
BIG1	Cell wall beta-1,6-glucan level regulator	4.51	1.99
GDA1	Guanosine-diphosphatase	4.99	2.50
FLO9	Lectin-like protein, flocculin (isoform 1)	2.99	1.06
Decreased enrichment			
YKL077W	Uncharacterized protein	1.39	3.49
PDI1	Protein disulfide isomerase	2.21	4.35
MNL1	Uncharacterized protein	1.53	3.81
KRE5	Beta-1,6-glucan biosynthesis protein (isoform 2)	2.84	5.47

translocons, and lumenal chaperones to produce the same number of polypeptide chains. However, if the gene is short enough to fail to sustain translation at the membrane (Figs. 5, 8b), then it may require multiple rounds of SRP targetting to get there. If sufficient nascent chains are exposed to the cytosol, the gene may also require cytosolic chaperones. If translation terminates prior to membrane attachment, then posttranslational translocons may be needed as well. Long, posttranslationally translocated proteins will also sequester ribosomes, but will require both lumenal and cytosolic chaperones (Fig. 8c). There are few genes in *K. phaffii* in this category (Fig. 5). Finally, short, posttranslationally translocated proteins will sequester few ribosomes, no cotranslational translocons, and some cytosolic and lumenal chaperones. Our experimental approach cannot measure transit time through posttranslational translocons; we speculate that it will be correlated to polypeptide length.

Some resources used in biogenesis of ER proteins are dependent on chain number, rather than elongation time. For instance, GPI-anchored proteins each receive a single lipid anchor [95], retrograde transport is mediated by the K/HDEL recognition [96], and protein sorting in the secretory pathway involves interactions between cargo and receptors, such as Sec24p [97]. In optimizing these systems, cTPM may be the appropriate metric to consider, and strain engineering efforts could focus on deleting or downregulating highly expressed host proteins. In yeasts, GPI-anchored cell wall proteins present the greatest burden by cTPM. Other aspects are dependent on

total polypeptide length, such as the potential ratcheting mechanism provided by Kar2p during translocation [84]. Although not considered here, cTPM scaled by protein length may be the appropriate metric used in engineering. A third aspect is the availability of resources such as ribosomes or translocons, which are sequestered while in operation. cRPM is an appropriate metric to understand ribosome sequestration. For cotranslational translocation, we propose that cRPM could be used as a proxy, as one ribosome binds one translocon during import. In *S. cerevisiae* and *K. phaffii*, expression of *PMA1* appears to be a major ribosome sink, and therefore also a translocon sink. In *K. phaffii*, *PST1* is a second major sink for ribosomes and translocons.

Although fungi are genetically and physiologically diverse, most mechanistic knowledge about secretion is derived from studies in *S. cerevisiae* [2]. Based on a recent molecular dating using 332 genomes [98], *K. phaffii* and *S. cerevisiae* diverged roughly 230 million years ago, whereas the *S. cerevisiae* whole-genome duplication occurred roughly 90 million years ago. Thus, sequence variation is found in nearly all of the proteins conserved in the two species, and due to the paralogy in *S. cerevisiae*, additional differences exist in the regulation of gene expression. Our comparison of *K. phaffii* and *S. cerevisiae* suggests that the path a conserved protein takes to the ER is not necessarily the same between species, even for essential genes critical to health of the secretory pathway, like *PDI1*. However, we find that even though the number and diversity of genes differ between the species, categorically there is conservation in the biosynthetic demand. For instance, our data suggest that *K. phaffii* can provide more nuanced engineering of the cell wall, as it is composed by a greater number of genes. Optimizing fungal species separately may increase protein secretion yields in ways not predicted through analysis of model organisms alone. These results call for a more thorough understanding of industrially used fungal secretion systems for rationally engineering cellular factories during bioproduction.

Conclusions

Protein biogenesis is a complex phenomena that not only requires raw materials (energy and amino acids), but also access to specialized cellular machinery. Our analysis in *K. phaffii* reveals several principles about these pathways that will be useful in strain engineering. First, we find that a small number of host genes are responsible for most of the protein entering the secretory pathway. Second, GPI-anchored protein components of the cell wall represent the greatest number of nascent chains within the secretory pathway. Third, cotranslational translocation pathways must accommodate a wider set of proteins than posttranslational pathways. Fourth, orthologs may enter the endoplasmic reticulum through different translocation pathways. Fifth, despite differences in the number of genes associated with biological function, the amount of nascent chains entering the ER are similar between *K. phaffii* and *S. cerevisiae*. Finally, we provide an updated genome annotation based on both Ribo-seq and long-read RNA-seq.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12934-020-01489-9.

Additional file 1: Figure S1. Ribo-Seq models active translation. a. Distribution of reads for different length RNA fragments. b. P-site offset for 30 nt fragment reveals active translation.

Additional file 2: Figure S2. Ribo-seq and long-read RNA-seq improve transcriptome annotation. Images are screen captures from Integrated Genome Viewer (MIT). a. Ribo-seq reads are stranded. In the top register, ribosome-protected footprint reads mapped to transcripts translated left to right are in red, and reads mapped transcripts translighted right to left are in blue. The middle register shows a prior annotation of transcripts and ORFs. The arrows indicate genes where the annotated translational start site disagrees with Ribo-seq. In both cases, an alternate start codon is used. The bottom register shows the annotation developed here using RNA-seq and long-read RNA-seq data. b. In an example transcript, Ribo-seq (top register) and long-read RNA-seq (bottomregister) reveal both the open reading frame and the untranslated regions (UTRs). c. Flow-chart of the annotation pipeline.

Additional file 3 Annotation of *K. phaffii* GS115 transcriptome. GFF3 annotation file containing transcript structures derived from Ribo-seq and long-read RNA-seq analysis.

Additional file 4. Comparison with prior *K. phaffii* annotations. Comma seperated value (CSV) file that links open reading frames defined in the current study with prior annotations of the *K. phaffii* transcriptome. Includes predicted protein sequence from each annotation. Additional details are provided in the file header.

Additional file 5. Combined results of Ribo-seq analysis of *K. phaffii*. Read counts, corrected transcripts per million (cTPM) and corrected ribosomes per million (cRPM) scores for *K. phaffii* GS115 open reading frame (ORFs). The file also includes experimental membrane enrichment scores, predictions of ORF features (localization, signal peptides, GPI-anchors and transmembrane domains), ORF sequences, homology to *S. cerevisiae*, and associations with prior annotations of the *K. phaffii* transcriptome. Additional details are provided in the file header.

Additional file 6: Figure S3. Comparison of membrane enrichment between data sets. a. Comparing membrane enrichment in two Ribo-Seq data sets in *K. phaffii*. b. Comparing membrane enrichment in Ribo-Seq data sets in *K. phaffii* and *S. cerevisiae*.

Additional file 7: Table S1. Biosynthetic demands for proteins with unknown functions by predicted subcellular localization.

Additinal file 8. Combined results of Ribo-seq analysis of *S. cerevisiae*. Read counts, corrected transcripts per million (cTPM) and corrected ribosomes per million (cRPM) scores for *S. cerevisiae* open reading frame (ORFs). The file also includes experimental membrane enrichment scores, predictions of ORF features (localization, signal peptides, GPI-anchors and transmembrane domains), and ORF sequences. Additional details are provided in the file header.

Additional file 9: Figure S4. Comparison of metabolic burden for *K. phaffii* and *S. cerevisiae*. a. Total nascent chains for *K. phaffii*. b. Total nascent chains for *S. cerevisiae*.

Acknowledgements
We thank Josh Kittleson, Gustavo Pesce, and Thomas Stevens (Bolt Threads) and Chris Love (MIT) for useful discussions. We also thank our colleagues in the Department of Bioengineering at UC Riverside.

Authors' contributions
TRA and JWC designed experiments. TRA, MR and JWC performed experiments. TRA and JWC performed analysis and wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by a gift from Bolt Threads Inc. (Emeryville, CA), the Bourns College of Engineering at the University of California, Riverside, and NSF CBET 1951942.

Availability of data and materials
The datasets generated and analysed during the current study are available as NCBI Bioproject PRJNA669501.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that the current study was funded in part by a gift from Bolt Threads Inc. (Emeryville, CA).

Author details
[1] Department of Bioengineering, University of California, Riverside 92521, United States of America. [2] Present Address: Protabit LLC, 1010 E Union St Suite 110, Pasadena, California 91106, United States of America.

Received: 14 August 2020 Accepted: 29 November 2020
Published online: 20 January 2021

References

1. Wang G, Huang M, Nielsen J. Exploring the potential of *Saccharomyces cerevisiae* for biopharmaceutical protein production. Curr Opin Biotechnol. 2017;48:77–84.
2. Delic M, Valli M, Graf AB, Pfeffer M, Mattanovich D, Gasser B. The secretory pathway: exploring yeast diversity. FEMS Microbiol Rev. 2013;37(6):872–914.
3. Kim H, Yoo SJ, Kang HA. Yeast synthetic biology for the production of recombinant therapeutic proteins. FEMS Yeast Res. 2015;15(1):1–16.
4. Love KR, Dalvie NC, Love JC. The yeast stands alone: the future of protein biologic production. Curr Opin Biotechnol. 2017;53:50–8.
5. Lopes H, Rocha I. Genome-scale modeling of yeast: chronology, applications and critical perspectives. FEMS Yeast Res. 2017;17(5):fox050.
6. Cai P, Gao J, Zhou Y. CRISPR-mediated genome editing in non-conventional yeasts for biotechnological applications. Microb Cell Fact. 2019;18(1):63.
7. Yamada Y, Matsuda M, Maeda K, et al. The phylogenetic relationships of methanol-assimilating yeasts based on the partial sequences of 18S and 26S ribosomal RNAs: The proposal of komagataella gen. nov. (Saccharomycetaceae). Biosci Biotechnol Biochem. 1995;59(3):439–44.
8. Kurtzman CP. Description of *Komagataella phaffii* sp. nov. and the transfer of Pichia pseudopastoris to the methylotrophic yeast genus komagataella. Int J Syst Evol Microbiol. 2005;55(2):973–6.
9. Kurtzman CP. Biotechnological strains of komagataella (pichia) pastoris are *komagataella phaffii* as determined from multigene sequence analysis. J Ind Microbiol Biotechnol. 2009;36(11):1435–8.
10. Karbalaei M, Rezaee SA, Farsiani H. *Pichia pastoris:* A highly successful expression system for optimal synthesis of heterologous proteins. J Cell Physiol. 2020;235(9):5867–81.
11. Ahmad M, Hirz M, Pichler H, Schwab H. Protein expression in *Pichia pastoris*: recent achievements and perspectives for heterologous protein production. Appl Microbiol Biotechnol. 2014;98(12):5301–17.
12. Zahrl RJ, Peña DA, Mattanovich D, Gasser B. Systems biotechnology for protein production in *Pichia pastoris*. FEMS Yeast Res. 2017;17(7):fox068.
13. Fischer JE, Glieder A. Current advances in engineering tools for *Pichia pastoris*. Curr Opin Biotechnol. 2019;59:175–81.
14. Kang Z, Huang H, Zhang Y, Du G, Chen J. Recent advances of molecular toolbox construction expand *Pichia pastoris* in synthetic biology applications. World J Microbiol Biotechnol. 2017;33(1):19.
15. Jiang H, Horwitz AA, Wright C, Tai A, Znameroski EA, Tsegaye Y, Warbington H, Bower BS, Alves C, Co C, Jonnalagadda K, Platt D, Walter JM, Natarajan V, Ubersax JA, Cherry JR, Love JC. Challenging the workhorse: Comparative analysis of eukaryotic microorganisms for expressing monoclonal antibodies. Bioeng: Biotechnol; 2019.
16. Crowell LE, Lu AE, Love KR, Stockdale A, Timmick SM, Wu D, Wang YA, Doherty W, Bonnyman A, Vecchiarello N, Goodwine C, Bradbury L, Brady JR, Clark JJ, Colant NA, Cvetkovic A, Dalvie NC, Liu D, Liu Y, Mascarenhas CA, Matthews CB, Mozdzierz NJ, Shah KA, Wu S-L, Hancock WS, Braatz RD, Cramer SM, Love JC. On-demand manufacturing of clinical-quality biopharmaceuticals. Nat. Biotechnol. 2018.
17. Zhou Y, Raju R, Alves C, Gilbert A. Debottlenecking protein secretion and reducing protein aggregation in the cellular host. Curr Opin Biotechnol. 2018;53:151–7.
18. Love KR, Politano TJ, Panagiotou V, Jiang B, Stadheim TA, Christopher Love J. Systematic single-cell analysis of Pichia pastoris reveals secretory capacity limits productivity. PLoS ONE. 2012;7(6):37915.
19. Zahrl RJ, Mattanovich D, Gasser B. The impact of ERAD on recombinant protein secretion in *Pichia pastoris* (syn komagataella spp.). Microbiology. 2018;164(4):453–63.
20. Deshaies RJ, Koch BD, Werner-Washburne M, Craig EA, Schekman R. A subfamily of stress proteins facilitates translocation of secretory and mitochondrial precursor polypeptides. Nature. 1988;332(6167):800–5.
21. Ast T, Cohen G, Schuldiner M. A network of cytosolic factors targets SRP-independent proteins to the endoplasmic reticulum. Cell. 2013;152(5):1134–45.
22. Aviram N, Schuldiner M. Targeting and translocation of proteins to the endoplasmic reticulum at a glance. J Cell Sci. 2017;130(24):4079–85.
23. Keenan RJ, Freymann DM, Stroud RM, Walter P. The signal recognition particle. Annu Rev Biochem. 2001;70:755–75.
24. Costa EA, Subramanian K, Nunnari J, Weissman JS. Defining the physiological role of SRP in protein-targeting efficiency and specificity. Science. 2018;359(6376):689–92.
25. Ng DT, Brown JD, Walter P. Signal sequences specify the targeting route to the endoplasmic reticulum membrane. J Cell Biol. 1996;134(2):269–78.
26. Shao S, Hegde RS. Membrane protein insertion at the endoplasmic reticulum. Annu Rev Cell Dev Biol. 2011;27:25–56.
27. Metzl-Raz E, Kafri M, Yaakov G, Soifer I, Gurvich Y, Barkai N. Principles of cellular resource allocation revealed by condition-dependent proteome profiling. Elife. 2017;6:e28034.
28. Klepsch MM, Persson JO, de Gier J-WL. Consequences of the overexpression of a eukaryotic membrane protein, the human KDEL receptor, in Escherichia coli. J Mol Biol. 2011;407(4):532–42.
29. Farkas Z, Kalapis D, Bódi Z, Szamecz B, Daraba A, Almási K, Kovács K, Boross G, Pál F, Horváth P, Balassa T, Molnár C, Pettkó-Szandtner A, Klement É, Rutkai E, Szvetnik A, Papp B, Pál C. Hsp70-associated chaperones have a critical role in buffering protein production costs. Elife. 2018;7.
30. Yang L, Yurkovich JT, King ZA, Palsson BO. Modeling the multi-scale mechanisms of macromolecular resource allocation. Curr Opin Microbiol. 2018;45:8–15.
31. Burgard J, Grünwald-Gruber C, Altmann F, Zanghellini J, Valli M, Mattanovich D, Gasser B. The secretome of *Pichia pastoris* in fed-batch cultivations is largely independent of the carbon source but changes quantitatively over cultivation time. Microb Biotechnol. 2020;13(2):479–94.
32. Feizi A, Österlund T, Petranovic D, Bordel S, Nielsen J. Genome-scale modeling of the protein secretory machinery in yeast. PLoS ONE. 2013;8(5):63284.
33. Gutierrez JM, Feizi A, Li S, Kallehauge TB, Hefzi H, Grav LM, Ley D, Baycin Hizal D, Betenbaugh MJ, Voldborg B, Faustrup Kildegaard H, Min Lee G, Palsson BO, Nielsen J, Lewis NE. Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. Nat. Commun. 2020;(1), 351387
34. Kol S, Ley D, Wulff T, Decker M, Arnsdorf J, Schoffelen S, Hansen AH, Jensen TL, Gutierrez JM, Chiang AWT, Masson HO, Palsson BO, Voldborg

BG, Pedersen LE, Kildegaard HF, Lee GM, Lewis NE. Multiplex secretome engineering enhances recombinant protein production and purity. Nat Commun. 2020;11(1):1908.
35. Chartron JW, Hunt KCL, Frydman J. Cotranslational signal-independent SRP preloading during membrane targeting. Nature. 2016;536(7615):224–8.
36. Jan CH, Williams CC, Weissman JS. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. Science. 2014;346(6210):1257521.
37. Gerashchenko MV, Gladyshev VN. Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res. 2014;42(17):e134.
38. McGlincy NJ, Ingolia NT. Transcriptome-wide measurement of translation by ribosome profiling. Methods. 2017;126:112–29.
39. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
40. Love KR, Shah KA, Whittaker CA, Wu J, Bartlett MC, Ma D, Leeson RL, Priest M, Borowsky J, Young SK, Love JC. Comparative genomics and transcriptomics of Pichia pastoris. BMC Genomics. 2016;17:550.
41. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
42. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
43. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
44. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–66.
45. Haas B, Papanicolaou A, et al. Transdecoder (find coding regions within transcripts). Github, nd https://github.com/TransDecoder/TransDecoder (accessed May 17, 2018) (2015)
46. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878–9.
47. Testa AC, Hane JK, Ellwood SR, Oliver RP. CodingQuarry: highly accurate hidden markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics. 2015;16:170.
48. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9(1):7.
49. Palmer J, Stajich J. nextgenusfs/funannotate: funannotate v1.5.3 2019.
50. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
51. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
52. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):1003118.
53. Popa A, Lebrigand K, Paquet A, Nottet N, Robbe-Sermesant K, Waldmann R, Barbry P. RiboProfiling: a bioconductor package for standard ribo-seq pipeline processing. F1000Res. 2016;5:1309.
54. Mohammad F, Green R, Buskirk AR. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. Elife. 2019;8:e42591.
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
56. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44(D1):286–93.
57. Liebermeister W, Noor E, Flamholz A, Davidi D, Bernhardt J, Milo R. Visual account of protein investment in cellular functions. Proc Natl Acad Sci USA. 2014;111(23):8488–93.
58. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33(21):3387–95.
59. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37(4):420–3.
60. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 2015;43(W1):401–7.
61. Pierleoni A, Martelli PL, Casadio R. PredGPI: a GPI-anchor predictor. BMC Bioinform. 2008;9:392.
62. Taggart JC, Li G-W. Production of Protein-Complex components is stoichiometric and lacks general feedback regulation in eukaryotes. Cell Syst. 2018;7(6):580–5894.
63. Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, Weng S, Wong ED, Lloyd P, Skrzypek MS, Miyasato SR, Simison M, Cherry JM. The reference genome sequence of saccharomyces cerevisiae: then and now. G3. 2014;4(3):389–98.
64. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–23.
65. Lareau LF, Hite DH, Hogan GJ, Brown PO. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. Elife. 2014;3:01257.
66. Ingolia NT. Ribosome footprint profiling of translation throughout the genome. Cell. 2016;165(1):22–33.
67. Valli M, Tatto NE, Peymann A, Gruber C, Landes N, Ekker H, Thallinger GG, Mattanovich D, Gasser B, Graf AB. Curation of the genome annotation of *Pichia pastoris* (*komagataella phaffii*) CBS7435 from gene level to protein function. FEMS Yeast Res. 2016;16(6).
68. De Schutter K, Lin Y-C, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouzé P, Van de Peer Y, Callewaert N. Genome sequence of the recombinant protein production host *Pichia pastoris*. Nat Biotechnol. 2009;27(6):561–6.
69. Blevins WR, Tavella T, Moro SG, Blasco-Moreno B, Closa-Mosquera, A, Díez J, Carey LB, Mar Albà M. Extensive post-transcriptional buffering of gene expression in the response to oxidative stress in baker's yeast 2019.
70. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F. French StatOmique Consortium: a comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14(6):671–83.
71. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131(4):281–5.
72. Anders S, Pyl PT, Huber W. HTSeq-a python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
73. Baudin-Baillieu A, Legendre R, Kuchly C, Hatin I, Demais S, Mestdagh C, Gautheret D, Namy O. Genome-wide translational changes induced by the prion [PSI+]. Cell Rep. 2014;8(2):439–48.
74. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010;466(7308):835–40.
75. Xiao Z, Zou Q, Liu Y, Yang X. Genome-wide assessment of differential translations with ribosome profiling data. Nat Commun. 2016;7:11194.
76. Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. Measurement of average decoding rates of the 61 sense codons in vivo. Elife. 2014;3:e03735.
77. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
78. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet. 2014;15(3):205–13.
79. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010;141(2):344–54.
80. Scannell DR, Butler G, Wolfe KH. Yeast genome evolution-the origin of the species. Yeast. 2007;24(11):929–42.
81. Ahn J, Hong J, Lee H, Park M, Lee E, Kim C, Choi E, Jung J, Lee H. Translation elongation factor 1-alpha gene from pichia pastoris: molecular

cloning, sequence, and use of its promoter. Appl Microbiol Biotechnol. 2007;74(3):601–8.
82. Kallehauge TB, Li S, Pedersen LE, Ha TK, Ley D, Andersen MR, Kildegaard HF, Lee GM, Lewis NE. Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion. Sci Rep. 2017;7:40388.
83. Balchin D, Hayer-Hartl M, Hartl FU. In vivo aspects of protein folding and quality control. Science. 2016;353(6294):4354.
84. Matlack KE, Misselwitz B, Plath K, Rapoport TA. BiP acts as a molecular ratchet during posttranslational transport of prepro-alpha factor across the ER membrane. Cell. 1999;97(5):553–64.
85. Brodsky JL, Goeckeler J, Schekman R. BiP and sec63p are required for both co- and posttranslational protein translocation into the yeast endoplasmic reticulum. Proc Natl Acad Sci USA. 1995;92(21):9643–6.
86. Mori A, Hara S, Sugahara T, Kojima T, Iwasaki Y, Kawarasaki Y, Sahara T, Ohgiya S, Nakano H. Signal peptide optimization tool for the secretion of recombinant protein from Saccharomyces cerevisiae. J Biosci Bioeng. 2015;120(5):518–25.
87. Sumi A, Okuyama K, Kobayashi K, Ohtani W, Ohmura T, Yokoyama K. Purification of recombinant human serum albumin efficient purification using STREAMLINE. Bioseparation. 1999;8(1–5):195–200.
88. Potgieter TI, Cukan M, Drummond JE, Houston-Cummings NR, Jiang Y, Li F, Lynaugh H, Mallem M, McKelvey TW, Mitchell T, Nylen A, Rittenhour A, Stadheim TA, Zha D, d'Anjou M. Production of monoclonal antibodies by glycoengineered *Pichia pastoris*. J Biotechnol. 2009;139(4):318–25.
89. Li G-W, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014;157(3):624–35.
90. Duttler S, Pechmann S, Frydman J. Principles of cotranslational ubiquitination and quality control at the ribosome. Mol Cell. 2013;50(3):379–93.
91. Wu X, Cabanos C, Rapoport TA. Structure of the post-translational protein translocation machinery of the ER membrane. Nature. 2018;566(7742):136–9.
92. Itskanov S, Park E. Structure of the posttranslational sec protein-translocation channel complex from yeast. Science. 2019;363(6422):84–7.
93. del Alamo M, Hogan DJ, Pechmann S, Albanese V, Brown PO, Frydman J. Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with Ribosome-Nascent chain complexes. PLoS Biol. 2011;9(7):1001100.
94. Diehn M, Eisen MB, Botstein D, Brown PO. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat Genet. 2000;25(1):58–62.
95. Mayor S, Riezman H. Sorting GPI-anchored proteins. Nat Rev Mol Cell Biol. 2004;5(2):110–20.
96. Semenza JC, Hardwick KG, Dean N, Pelham HR. ERD2, a yeast gene required for the receptor-mediated retrieval of luminal ER proteins from the secretory pathway. Cell. 1990;61(7):1349–57.
97. Geva Y, Schuldiner M. The back and forth of cargo exit from the endoplasmic reticulum. Curr Biol. 2014;24(3):130–6.
98. Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, Boudouris JT, Schneider RM, Langdon QK, Ohkuma M, Endoh R, Takashima M, Manabe R-I, Čadež N, Libkind D, Rosa CA, DeVirgilio J, Hulfachor AB, Groenewald M, Kurtzman CP, Hittinger CT, Rokas A. Tempo and mode of genome evolution in the budding yeast subphylum. Cell. 2018;175(6):1533–154520.

Publisher's Note