Protein biogenesis demands of the early secretory pathway in Komagataella phaffii

Troy R Alva, Melanie Riera and Justin W Chartron*

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 a microbial cell factory, 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 robust to physical and chemical stress [1]. They also have an endomembrane system that is fundamentally conserved with higher eukaryotes. This oxidative environment supports glycosylation and subsequent glycan modification, folding using ATP-driven molecular chaperones and protein disulfide isomerases, and protein quality control [2]. Compared to mammalian cells, yeasts have simpler genomes and can be more easily characterized and modified [3]. Combine this with tools such as CRISPR/cas9 and the range of tractable organisms is expanding [4, 5]. Komagataella phaffii (one of two species previously known as P. pastoris) stands out for its high-protein secretion capacity, its ability to metabolize methanol as its primary carbon source, and its safety record as a source of biologics [3]. Thus, K. phaffii is an ideal chassis to rapidly implement changes designed to improve protein expression and secretion. Improving expression from a species like K. phaffii can accelerate product development and allow cheap, local production of pharmaceuticals [6, 7].
Identifying and relieving protein biogenesis bottlenecks is one strategy to improve yields of high-value, recombinant proteins [1, 8]. 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) [9, 10]. Yeasts have multiple pathways for translocation, which use partially overlapping sets of biogenesis factors (reviewed in [11]). 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 [12, 13]. Translocons bind hydrophobic amino acid motifs, called signal peptides, found at the amino termini of secreted proteins [14]. Some signal peptides are dependent upon a cytosolic factor, the Signal Recognition Particle, and the ER-bound SRP receptor to engage a translocon [15]; these tend to be longer or more hydrophobic than SRP independent signals [16, 17]. 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 [18]. Integral membrane proteins that use a sec translocon require SRP for targeting to the ER over mitochondria [16].

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 [19–21]. 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 [22], 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 [23], 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 [24]. 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 [25, 26]. 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*. [16, 27, 28] 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, which will compete 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.

1 Materials & Methods

1.1 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, cycloheximide (CHX) was added to 100µg ml$^{-1}$ for 3 min prior to harvesting.

1.2 Lysis and subcellular fractionation

Cells were lysed in either soluble lysis buffer (50 mM MOPS, 25 mM potassium hydroxide, 100 mM potassium acetate, 2 mM magnesium acetate, 1 mM dithiothreitol and 100µg 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 x g for 10 minutes. 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.

1.3 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
SUPERaseIn (Invitrogen).

Purified fragments were used to prepare sequencing libraries as described in [29]
with some modification. Linker ligations were allowed to proceed for 4 hours, and
afterwards, samples were pooled and purified by TBE-urea PAGE. The pooled li-
brary was depleted of ribosomal RNA using the Ribo-Zero Gold rRNA Removal Kit
(Illumina), following manufacturer’s instructions. Reverse transcriptions were per-
formed using SuperScript II (Invitrogen). After circularization, PCR amplification
and TBE PAGE purification, libraries were quantified using a Qubit 2.0 Fluorom-
eter (Invitrogen) and sequenced using a HiSeq 4000 (Illumina). Linker sequences
were trimmed and libraries were demultiplexed using Cutadapt [30].

1.4 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 minutes during incubation with TRI
reagent. After purifying RNA, a library was prepared using a PCR-cDNA kit ac-
cording to manufacturer’s instructions (SQK-PCS109, Oxford Nanopore Technolo-
gies) and sequenced using a minION R9.4.1 flow cell. Base calling was performed
using Guppy (Oxford Nanopore Technologies).

1.5 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 [31]. 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 [32] and Minimap2 [33] respectively.
Stringtie version 1.3.6 was used to assemble transcripts from Ribo-seq data, with
reads mapping to each strand processed separately [34]. Pinfish was used to assem-
ble transcripts from long reads (Oxford Nanopore Technologies). After transcript
assembly, PASA [35] was used to combine the Stringtie and Pinfish models into a
single transcriptome. Transdecoder [36] was then run twice: first, to identify candi-
date 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 [37] and CodingQuarry [38], which were used to provide de
novo predictions in the genome. EVidenceModeler [39] was used to incorporate pre-
dictions 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 [40].

1.6 Mapping of ribosome protected reads to codons and masking
Ribo-seq reads were mapped to the genome of Komagataella pastoris GS115 [31]
using HISAT2 [32, 41]. Alignments were converted from SAM to sorted and indexed
BAM files using Samtools and only included reads with mapping quality threshold of 60 [42]. Mapped reads were loaded into R using the GenomicAlignments package from Bioconductor [43] and converted to their 3’ end positions before determining p-site offsets. P-site offsets were determined using the RiboProfiling package in Bioconductor [44]. 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 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 [45].

1.7 Metagene correction and quantification of metabolic demand

Read counts were normalized at the codon level using a metagene analysis 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.

1.8 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 [46] 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 [47] 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 [31] 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 [47]. Vironoi
tessellations were created to quantitatively map the biosynthetic composition of
these functions using COGs and expression metrics derived from Ribo-Seq cTPM
[48]. DeepLoc was used to predict the subcellular localization associated with ORF
products [49]. Sequence features such as signal sequences, transmembrane domains
(TMD), and GPI anchors were identified using SignalP 5.0 [50], TOPCONS [51],
and predGPI [52] respectively.

1.9 **S. cerevisiae** analysis

Ribo-seq data for total protein synthesis were taken from [53], and data obtained
from soluble or membrane-bound ribosome fractions were obtained from [27]. All
data were processed in the same way as *K. phaffii* using the S288C reference genome
R64-2-1 [54].

2 Results

2.1 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 biosyn-
thetic resources, such as ribosomes and ER translocons. We also asked which genes
were responsible for producing the most nascent chains, which is critical for predict-
ing 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 [55].
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 nt to 22 nt or 28 nt to 30 nt “footprints” of ribosome-protected mRNA depending
on the translational conformation of the ribosome [56], which are then sequenced.
We performed a series of Ribo-seq experiments to capture global translation and
translation on the surface of organelles (*Figure 1*) Our data sets captured foot-
print lengths from 15 nt to 42 nt (*Figure S1a*). Nearly all (99%) footprints mapped
within open reading frames (ORFs). Our profiling data also indicate active trans-
lation through the appearance of three nucleotide periodicity in read depth that is
preserved across the transcriptome (*Figure S1b*).

We noticed that ribosome-protected read patterns were often inconsistent with
prior annotations of open reading frames, especially for highly expressed proteins
(*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 metagene correction (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 [57]. 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 [39, 40]. 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 (Figure S2b, c). Our annotation is provided as Additional File 1. 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) [31, 58, 59]. Homologs between our annotation and prior annotations are provided as Additional File 2. Our annotation adjusted the translational start site of about 10% of ORFs compared to each previous model. Overall, Ribo-seq reads were mapped to 5,303 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.

### 2.2 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 abundance than RNA-seq [60]. 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 [61, 62]. 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.

---

**Table 1 Comparison of ORF annotations**

| Annotation¹ | Total ORFs | Homologs² | Length differences³ |
|-------------|------------|-----------|---------------------|
| Current study | 5329       |           |                     |
| GS115 (PRJNA304976) | 5064       | 5035      | 514                 |
| GS115 (PRJEA37871) | 5040       | 5100      | 697                 |
| CBS7435 (PRJEA62483) | 5291       | 5198      | 604                 |

¹NCBI bioproject numbers located in parenthesis.
²BlastP matches from current study to prior study.
³Number of homologs with different predicted lengths.
Ribosome protected fragments are small, 22 nt to 30 nt, and can often map to multiple locations when the transcriptome contains homologous stretches of mRNA. Previous efforts to address multi-mapped reads introduce other biases when calculating gene expression. Multi-mapped reads that are thrown out altogether \[63–66\] can depreciate read counts for highly expressed genes while randomly assigning reads to ORFs with the highest percentage of alignment \[55, 67, 68\] can overestimate read counts for lowly expressed genes. Recent efforts have used computational masks to exclude homologous ranges of codons, and these results more strongly agree with protein abundancies compared to previous methods\[53\]. Using this approach, ORF segments with high percentage identity to other ORFs will therefore have large segments without read counts. Here, we adapt the method of Taggart et al. \[53\] and calculated a mask over the \textit{K. phaffii} transcriptome accounting for all possible 28 nt reads. This method’s results exclude only 3% of codon positions available for counting reads in \textit{K. phaffii}. Metrics that predict gene expression using sequencing reads collected across a transcript must be scaled by transcript length. Incorrectly annotated transcriptional start or stop sites, or other masked, empty segments will result in biases in the calculation of gene expression. A computationally shortened sequence length is used instead of actual ORF length in normalized calculations of expression. Shortened lengths, however, can be problematic because of ribosome protected reads are not evenly distributed across transcripts.

Ribosome-protected reads accumulate at the 5’ end of ORFs for a variety of reasons, such as having a potentially slower elongation rate at the beginning of translation \[53, 55\]. This bias affects every transcript and has a conserved profile (Figure 2a). As a result, long ORFs will appear to be underrepresented in single-value metrics of expression (like TPM) and short ORFs will appear to be over represented, even if their initiation and elongation kinetics are the same. We adapt the method of Taggart et al., where a composite “metagene” profile is calculated and used to scale reads at each codon position. This procedure effectively removes the positional global bias, so that long and short genes can be directly compared (Figure 2b). For example, transcripts of \textit{RPL5} and \textit{YEF3} display similar numbers of ribosomes at the start of their ORFs (Figure 2c). However, because \textit{YEF3} is a longer transcript, the 5’ bias causes the uncorrected TPM to be smaller than the TPM of \textit{RPL5}. Since the 5’ bias is strongly conserved across transcriptomes, we assume that if \textit{RPL5} were as long as \textit{YEF3}, the genes would have similar TPM. After correcting for this bias with metagene normalization, the corrected TPM (cTPM) scores are similar between the two genes. Thus, we use cTPM, which includes both masking and metagene scaling, as a measure of gene expression or nascent chain production.

While cTPM provides an estimate of the number of nascent polypeptide chains, it does not answer our question of ribosome sequestration. Longer transcripts will sequester a greater number of ribosomes, a shared cellular resource, in order to produce the same number of nascent chains as a shorter transcript. If the 5’ bias is due to slower elongation kinetics at the start of transcripts and occurs \textit{in vivo}, then it is important to include this effect while measuring global ribosome allocation, as these ribosomes remain unavailable to other transcripts. If reads could be unambiguously mapped to the transcriptome, then simply scaling reads per gene by library size would indicate ribosome usage per gene. However, when masking is
Table 2 Nascent chains produced in K. phaffii

| Ontological Functions                                                                 | Nascent chains (%) | Genes (n) |
|--------------------------------------------------------------------------------------|--------------------|-----------|
| Translation, Ribosomal 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                        | 6.0%               | 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                                                       | 8%                 | 266       |
| GPI Anchors                                                                          | 79%                | 117       |
| Transmembrane proteins                                                               | 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.

applied to correct for ambiguous reads, the position of the mask becomes important (Figure 2a, b). Two masks of the same length, but applied at different positions along transcripts of the same length, will hide different amounts of ribosomes based on the metagene profile. To correct for this, we calculate a ribosome scaling factor for each gene which accounts for both the position of the mask and the metagene profile. We present a new metric for each gene, corrected ribosomes per million (cRPM), which is the product of the masked ribosome protected reads and the ribosome scaling factor. This product is then scaled by library size. The acronym is chosen because it is functionally equivalent to reads per million (RPM) in standard RNA-seq. In our example in Figure 2c, cRPM and RPM are approximately identical, as expected since there are no masks applied to RPL5 or YEF3. cRPM is different between these two genes, yet is not linearly related to length due to the bias in ribosome distribution. Read counts, cTPM and cRPM for each gene in each dataset are provided as Additional File 3.

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 Figure 3a), as expected for log-phase growth. While the majority of genes with unknown ontological functions are predicted to be localized in the nucleus, the majority of nascent chains produced by this category of genes are predicted to have extracellular localizations. We consider endomembrane lumenal and secreted proteins to be those with predicted N-terminal signal sequences, are not predicted to be localized to the mitochondria, and 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.
2.3 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 (Figure 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 [27] (Figure 4). Membrane enrichment scores were calculated as the $\log_2$ ratio of cTPM for membrane and soluble fractions and were reproducible (Figure S3a and provided in Additional File 3). The magnitude of membrane enrichment scores depends on the efficiency of fractionation, and if a gene falls below the diagonal line in Figure 4, it will have a negative enrichment score. As in *S. cerevisiae*, membrane enrichment scores are limited by the length of the ORF when transcripts encode signal-sequence bearing proteins [27, 28] (Figure 5). This effect is due to a kinetic competition between trafficking rate and translation elongation rate. Figure 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 (Figure 3b). To assess entry into the ER, we filtered out transcripts encoding proteins predicted to localize in the mitochondria by DeepLoc (Figure 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 2-fold membrane enrichment (Figure 3d). Posttranslationally trafficked membrane proteins rely on other mechanisms, such as the GET pathway [14].

A more diverse group of proteins enter the ER through cotranslational translocons than those that enter posttranslationally (Figure 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 (Figure 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 (Table S1).

2.4 Comparison with *S. cerevisiae*

Of the 5,329 *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 dupli-
Table 3 Comparison of translocon demands by ontological function.

| Function Unknown | Cotranslationally Translocated<sup>c</sup> | Posttranslationally Translocated<sup>d</sup> |
|------------------|-------------------------------------------|------------------------------------------|
| 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% |

<sup>a</sup>Calculated as percent of total cTPM for all proteins predicted to be ER destined.
<sup>b</sup>Calculated as percent of total cRPM for all proteins predicted to be ER destined.
<sup>c</sup>Proteins with greater than 2-fold membrane enrichment and not predicted to be mitochondrial.
<sup>d</sup>Proteins with less than 2-fold membrane enrichment and not predicted to be mitochondrial and contained a predicted signal sequence.

cation, and so many *S. cerevisiae* genes have paralogs [69]. The influence of paralogy becomes evident in the way these two species use resources. Our calculations of cTPM and cRPM are presented in Additional File 4, using previously acquired data [27, 53]. The overall distribution of cTPM by ontological category is similar between species (Figure S4). Under the conditions tested here, TEF1, encoding translational elongation factor 1 alpha, is the highest expressed protein in *K. phaffii*. *S. cerevisiae* generates a similar amount of nascent chains to this function but does so using an combination of its paralogous genes TEF1 and TEF2. *S. cerevisiae* is Crabtree-negative and undergoes alcoholic fermentation even under aerobic conditions [70]. Unsurprisingly, *S. cerevisiae* generates three times more polypeptides involved in carbohydrate transport and metabolism than *K. phaffii*, particularly genes involved in glycolysis and gluconeogenesis.

Indeed, these two species also show divergence in energy production with regards to cotranslational mitochondrial import (Figure 6). Our subcellular fractionation 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 [27]. 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 (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 (Figure 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 luminal 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* (Figure 6c).
Table 4 Membrane enrichment for secreted, luminal and GPI-anchored proteins in K. phaffii and S. cerevisiae

| Gene      | Product                                                                 | K. phaffii | S. cerevisiae |
|-----------|--------------------------------------------------------------------------|------------|---------------|
| Increased |                                                                          |            |               |
| FLO9      | Lectin-like protein, flocculin (isoform 2)                               | 5.32       | 1.06          |
| ZPS1      | Putative GPI-anchored protein                                             | 5.80       | 2.54          |
| SGA1      | Sporulation-specific glucosylase                                          | 4.49       | 1.32          |
| B1G1      | 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 |                                                                          |            |               |
| YKL077W   | Uncharacterized protein                                                  | 1.39       | 3.49          |
| PDI1      | Protein disulfide isomerase                                               | 2.21       | 4.35          |
| MNL1      | Uncharacterized protein                                                  | 1.53       | 3.81          |
| KRES      | Beta-1,6-glucan biosynthesis protein (isoform 2)                         | 2.84       | 5.47          |

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 (Figure 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 luminal, 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 luminal chaperones to synthesize and transport nascent chains into the ER.

3 Discussion

Our Ribo-seq analysis reveals several of the biosynthetic demands imposed by the host proteome on the early secretory pathway of *K. phaffii*. The yields of engineered, recombinant proteins are restricted by bottlenecks in protein biogenesis [1]. These bottlenecks may be metabolic, or they may be due to insufficient cellular biogenesis machinery, including polymerases, ribosomes, translocons, and molecular chaperones. Engineered proteins compete with the host proteome for these resources, and by globally quantifying host usage we can predict design principles for strain engineering.

Deleting metabolically expensive host proteins relieves bottlenecks in protein production. Indeed, the Lewis lab has elegantly demonstrated that deleting highly expressed proteins in CHO cells increases the yield of heterologous secreted proteins [25, 26]. Modification of the secretory pathway, such as the optimization of signal sequences for protein targeting [71] and reducing the effect of the ERAD system [10], provides varying degrees of success and is contingent on the complexity and optimization of the protein product [72, 73]. Our data and analysis may augment efforts by accounting for biogenesis capacity.
Despite the ability of Ribo-seq to accurately quantify gene expression, our study has several caveats. First, we have only considered log phase growth in liter scale, aerated shaking cultures. We selected our growth conditions to enable comparison to *S. cerevisiae*. Future work could involve quantifying demands at industrial scale in stirred bioreactors under induction. 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 [53, 74]. Third, Ribo-seq does not account for protein degradation; indeed, some proteins are cotranslationally ubiquitinated [75]. 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 [11]. 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 [13, 17, 28]. Molecular chaperones are also implicated in protein translocation, through binding of proteins in the cytoplasm (Ssa1p) [12] or the ER lumen (Kar2p) [76]. However, many gene products are able to associate with more than one class of translocon [17, 28]. In addition, while recent structural work suggests that the heptameric Sec61 complex cannot directly bind a ribosome [77, 78], there is a preponderance of evidence demonstrating that the proteins dependent on this complex are translated at the ER membrane [16, 27, 28, 79, 80]. 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 [27, 79]. 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 will impart little demand on cytoplasmic chaperones, but will sequester ribosomes, translocons, and lumenal chaperones for extended periods of time (*Figure 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, translocons, and luminal chaperones to produce the same number of polypeptide chains. However, if the gene is short enough to fail to sustain translation at the membrane (*Figures 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 (Figure 8c). There are few genes in *K. phaffii* in this category (Figure 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 [81], retrograde transport is mediated by the K/HDEL recognition [82], and protein sorting in the secretory pathway involves interactions between cargo and receptors, such as Sec24p [83]. 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 [76]. 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* [11]. *K. phaffii* and *S. cerevisiae* have diverged so that subtle differences exist in the sequence of conserved proteins, and due to paralogy, larger differences exist in the regulation of gene expression. Our results between *K. phaffii* and *S. cerevisiae* suggest 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.

### 3.1 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.

**Declarations**

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets generated and analysed during the current study are available in the NCBI Gene Expression Omnibus under accession number XXXXXXX.

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

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.

Authors’ contributions
TRA and JWC designed experiments. TRA, MR and JWC performed experiments. TRA and JWC performed analysis and wrote the manuscript.

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.

Additional files
Additional file 1 — GS115_CRG.gff
GFF3-format annotation of *K. phaffii* strain GS115.

Additional file 2 — annotation_homology.csv
Comparison between the annotation presented in the current work to prior annotations.

Additional file 3 — kphaffii.csv
Comma-separated values containing data and bioinformatic predictions for *K. phaffii* strain GS115.

Additional file 4 — scerevisiae.csv
Comma-separated values containing data and bioinformatic predictions for *S. cerevisiae*, based on reference strain S288C.

References
1. Wang, G., Huang, M., Nielsen, J.: Exploring the potential of saccharomyces cerevisiae for biopharmaceutical protein production. Curr. Opin. Biotechnol. 48, 77–84 (2017)
2. Kim, H., Yoo, S.J., Kang, H.A.: Yeast synthetic biology for the production of recombinant therapeutic proteins. FEMS Yeast Res. 15(1), 1–16 (2015)
3. Love, K.R., Dalvie, N.C., Love, J.C.: The yeast stands alone: the future of protein biologic production. Curr. Opin. Biotechnol. 53, 50–58 (2017)
4. Lopes, H., Rocha, I.: Genome-scale modeling of yeast: chronology, applications and critical perspectives. FEMS Yeast Res. 17(5) (2017)
5. Cai, P., Gao, J., Zhou, Y.: CRISPR-mediated genome editing in non-conventional yeasts for biotechnological applications. Microb. Cell Fact. 18(1), 63 (2019)
34. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L.: StringTie enables
24. Feizi, A., Østerlund, T., Petranovic, D., Bordel, S., Nielsen, J.: Genome-scale modeling of the protein secretory
27. Chartron, J.W., Hunt, K.C.L., Frydman, J.: Cotranslational signal-independent SRP preloading during
26. Kol, S., Ley, D., Wulff, T., Decker, M., Arnsdorf, J., Schoffelen, S., Hansen, A.H., Jensen, T.L., Gutierrez,
34. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics
28. Jan, C.H., Williams, C.C., Weissman, J.S.: Principles of ER cotranslational translocation revealed by
32. Kim, D., Langmead, B., Salzberg, S.L.: HISAT: a fast spliced aligner with low memory requirements. Nat.
30. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal
35. Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K. Jr, Hannick, L.I., Maiti, R., Ronning,
16. Costa, E.A., Subramanian, K., Nunnari, J., Weissman, J.S.: Defining the physiological role of SRP in
15. Keenan, R.J., Freymann, D.M., Stroud, R.M., Walter, P.: The signal recognition particle. Annu. Rev. Biochem.
14. Aviram, N., Schuldiner, M.: Targeting and translocation of proteins to the endoplasmic reticulum at a glance.
13. Ast, T., Cohen, G., Schuldiner, M.: A network of cytosolic factors targets SRP-independent proteins to the
endoplasmic reticulum. Cell 152(5), 1134–1145 (2013)
12. Deschran, J.W., Hunt, K.C.L., Werner-Washburne, M., Craig, E.A., Scheiman, R.: A subfamily of stress proteins
facilitates translocation of secretory and mitochondrial precursor polypeptides. Nature 332(6167), 800–805 (1988)
11. Delic, M., Valli, M., Graf, A.B., Pfeffer, M., Mattanovich, D., Gasser, B.: The secretory pathway: exploring
yeast diversity. FEMS Microbiol. Rev. 37(6), 872–914 (2013)
10. Zahrl, R.J., Mattanovich, D., Gasser, B.: The impact of ERAD on recombinant protein secretion in pichia
pastoris (syn komagatetella spp). Microbiology 164(4), 453–463 (2018)
9. Love, K.R., Politano, T.J., Panagiotou, V., Jiang, B., Stadheim, T.A., Christopher Love, J.: Systematic
Single-Cell analysis of pichia pastoris reveals secretory capacity limits productivity. PLoS One 7(6), 37915 (2012)
8. Zhou, Y., Raju, R., Alves, C., Gilbert, A.: Debottlenecking protein secretion and reducing protein aggregation
in the cellular host. Curr. Opin. Biotechnol. 53, 151–157 (2018)
7. Jiang, H., Horwitz, A.A., Wright, C., Tai, A., Znameroski, E.A., Tsegaye, Y., Warbington, H., Bower, B.S.,
Alves, C., Co, C., Jonnalagadda, K., Platt, D., Walter, J.M., Natarajan, V., Ubersax, J.A., Cherry, J.R., Love,
J.C.: Challenging the workhorse: Comparative analysis of eukaryotic microorganisms for expressing monoclonal
antibodies. Biotechnol. Bioeng. (2019)
6. Jiang, H., Horwitz, A.A., Wright, C., Tai, A., Znameroski, E.A., Tsegaye, Y., Warbington, H., Bower, B.S.,
Alves, C., Co, C., Jonnalagadda, K., Platt, D., Walter, J.M., Natarajan, V., Ubersax, J.A., Cherry, J.R., Love,
J.C.: Challenging the workhorse: Comparative analysis of eukaryotic microorganisms for expressing monoclonal
antibodies. Biotechnol. Bioeng. (2019)
5. Crowell, L.E., Lu, A.E., Love, K.R., Stockdale, A., Timnwick, S.M., Wu, D., Wang, Y.A., Doherty, W.,
Bonnynman, A., Vecchiolo, N., Godwine, C., Bradbury, L., Brady, J.R., Clark, J.J., Colant, N.A., Cvetkovic,
A., Dalvie, N.C., Liu, D., Liu, Y., Mascarenhas, C.A., Matthews, C.B., Mzdzierz, N.J., Shah, K.A., Wu, S.-L.,
Hancock, W.S., Braatz, R.D., Cramer, S.M., Love, J.C.: On-demand manufacturing of clinical-quality
biopharmaceuticals. Nat. Biotechnol. (2018)
4. Zhou, Y., Raju, R., Alves, C., Gilbert, A.: Debottlenecking protein secretion and reducing protein aggregation
in the cellular host. Curr. Opin. Biotechnol. 53, 151–157 (2018)
3. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18), 3094–3100 (2018)
2. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18), 3094–3100 (2018)
1. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal
17(1), 10–12 (2011)
1. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal
17(1), 10–12 (2011)
1. Love, K.R., Shah, K.A., Whittaker, C.A., Wu, J., Bartlett, M.C., Ma, D., Leeson, R.L., Priest, M., Borowsky,
J., Young, S.K., Love, J.C.: Comparative genomics and transcriptomics of pichia pastoris. BMC Genomics 17, 550 (2016)
1. Kim, D., Langmead, B., Salzberg, S.L.: HISAT: a fast spliced aligner with low memory requirements. Nat.
Methods 12(4), 357–360 (2015)
1. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18), 3094–3100 (2018)
1. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L.: StringTie enables
improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33(3), 290–295 (2015)
1. Haas, B.J., Delcher, A.L., Mount, S.W., Worthington, J.R., Smith, K.R. Jr, Hannick, L.I., Maiti, R., Running,
C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., White, O.: Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31(19), 5654–5666 (2003)

36. Haas, B., Papanicolaou, A., et al.: Transdecoder (find coding regions within transcripts). Github, nd https://github.com/TransDecoder/TransDecoder (accessed May 17, 2018) (2015)

37. Majoros, W.H., Pertea, M., Salzberg, S.L.: TigScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20(16), 2578–2579 (2004)

38. Testa, A.C., Hane, J.K., Elwood, S.R., Oliver, R.P.: CodingQuary: highly accurate hidden markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 16, 170 (2015)

39. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., Wortman, J.R.: Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 9(1), 7 (2008)

40. Palmer, J., Stojich, J.: nextgenusfs/funannotate: funannotate v1.5.3 (2019)

41. Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L.: Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37(8), 907–915 (2019)

42. 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 25(16), 2078–2079 (2009)

43. Lawrence, M., Huber, W., Pagés, H., Abyouyn, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J.: Software for computing and annotating genomic ranges. PLoS Comput. Biol. 9(8), 1003118 (2013)

44. Popa, A., Le, E., Estelle, K., Paquet, A., Notte, N., Robbe-Sermesant, K., Waldmann, R., Barby, P.: RiboProfiling: a biocomputer package for standard ribo-seq pipeline processing. F1000Res. 5, 1309 (2016)

45. Mohammad, F., Green, R., Buskirk, A.R.: A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. Elife 8 (2019)

46. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410 (1990)

47. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H.,eller, D., Walter, M.C., Rattie, T., Mende, D.R., Sunagawa, S., Kuhn, M., Jensen, L.J., 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. 44(D1), 286–93 (2016)

48. Liebermeister, W., Noor, E., Flamholz, A., Davidi, D., Bernhardt, J., Milo, R.: Visual account of protein investment in cellular functions. Proc. Natl. Acad. Sci. U. S. A. 111(23), 8488–8493 (2014)

49. Almagro Armenteros, J.J., Sønderby, C.K., Sønderby, S.K., Nielsen, H., Winther, O.: DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33(21), 3387–3395 (2017)

50. Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., Nielsen, H.: SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. 37(4), 420–423 (2019)

51. Tsirigos, K.D., Petersen, C., Shu, N., Käll, L., Elofsson, A.: The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 43(W1), 401–7 (2015)

52. Pierleoni, A., Martelli, P.L., Casadio, R.: PredGPI: a GPI-anchor predictor. BMC Bioinformatics 9, 392 (2008)

53. Taggart, J.C., Li, G.-W.: Production of Protein-Complex components is stoichiometric and lacks general feedback regulation in eukaryotes. Cell Syst 7(6), 580–5894 (2018)

54. Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C., Hitz, B.C., Karra, A., Nash, R.S., Weng, S., Wong, E.D., Lloyd, P., Krzypeki, M.S., Miyasato, S.R., Simson, M., Cherry, J.M.: The reference genome sequence of saccharomyces cerevisiae: then and now. G3 4(3), 389–398 (2014)

55. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., Weissman, J.S.: Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 347(6227), 741–746 (2015)

56. Lareau, L.F., Hite, D.H., Hogan, G.J., Brown, P.O.: Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNAs. Elife 3, e01257 (2014)

57. Ingolia, N.T.: Ribosome footprint profiling of translation throughout the genome. Cell 165(1), 22–33 (2016)

58. Valli, M., Tattoo, N.E., Peymann, A., Gruber, C., Landes, N., Ekker, H., Thallinger, G.G., Mattanovich, D., Gauer, R., Graf, A.B.: Curation of the genome annotation of pichia pastoris (komagataella phaffii) CBS7435 from gene level to protein function. FEMS Yeast Res. 16(6) (2016)

59. 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. 27(6), 561–566 (2009)

60. Blevins, W.R., Tavella, T., Moro, S.G., Blasco-Moreno, B., Closa-Mosquera, A., Diez, J., Carey, L.B., Mar Albà, M.: Extensive post-transcriptional buffering of gene expression in the response to oxidative stress in baker’s yeast (2019)

61. Dillies, M.-A., Rau, A., Aubert, J., Henniquet-Antier, C., Jeannougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guenec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaffter, B., Le Cram, S., Guedj, M., Jaffrézic, F., French StatOmique Consortium: A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. 14(6), 671–683 (2013)

62. Wagner, G.P., Kin, K., Lynch, V.J.: Measurement of mRNA abundance using RNA-seq data: RPMK measure is inconsistent among samples. Theory Biosci. 131(4), 281–285 (2012)

63. Anders, S., Pyl, P.T., Huber, W.: HTSeq—a python framework to work with high-throughput sequencing data. Bioinformatics 31(2), 166–169 (2015)

64. Baudin-Bailleu, A., Legendre, R., Kuchly, C., Hatin, I., Demais, S., Mestdagh, C., Gauthier, D., Namy, O.: Genome-wide translational changes induced by the prion [PSI+]. Cell Rep. 8(2), 439–448 (2014)

65. Guo, H., Ingolia, N.T., Weissman, J.S., Bartel, D.P.: Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466(7308), 835–840 (2010)

66. Xiao, Z., Zou, Q., Liu, Y., Yang, X., Genome-wide assessment of differential translations with ribosome profiling data. Nat. Commun. 7, 11194 (2016)

67. Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., Futch, B.: Measurement of average decoding rates
of the 61 sense codons in vivo. Elife 3 (2014)

68. Trapnell, C., Pachter, L., Salzberg, S.L.: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9), 1105–1111 (2009)

69. Scannell, D.R., Butler, G., Wolfe, K.H.: Yeast genome evolution—the origin of the species. Yeast 24(11), 929–942 (2007)

70. Pfeiffer, T., Marley, A.: An evolutionary perspective on the crabtree effect. Front Mol Biosci 1, 17 (2014)

71. Mori, A., Haras, 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. 120(5), 518–525 (2015)

72. Sumi, A., Okuyama, K., Kobayashi, K., Ohtani, W., Ohmura, T., Yokoyama, K.: Purification of recombinant human serum albumin efficient purification using STREAMLINE. Bioseparation 8(1-5), 195–200 (1999)

73. Potgieter, T.I., Cukan, M., Drummond, J.E., Houston-Cummings, N.R., Jiang, Y., Li, F., Lynamgh, H., Mallem, M., McKeelvey, T.W., Mitchell, T., Nylen, A., Rittenhour, A., Studheim, T.A., Zha, D., d'Anjou, M.: Production of monoclonal antibodies by glycoengineered Pichia pastoris. J. Biotechnol. 139(4), 318–325 (2009)

74. Li, G.-W., Burkhardt, D., Gross, C., Weissman, J.S.: Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157(3), 624–635 (2014)

75. Duttler, S., Pechmann, S., Frydman, J.: Principles of cotranslational ubiquitination and quality control at the ribosome. Mol. Cell 50(3), 379–393 (2013)

76. Matlack, K.E., Misseiwitz, B., Plath, K., Rapoport, T.A.: BiP acts as a molecular ratchet during post-translational transport of prepro-alpha factor across the ER membrane. Cell 97(5), 553–564 (1999)

77. Wu, X., Cabanos, C., Rapoport, T.A.: Structure of the post-translational protein translocation machinery of the ER membrane. Nature, 1 (2018)

78. Itskanov, S., Park, E.: Structure of the posttranslational sec protein-translocation channel complex from yeast. Science 363(6422), 84–87 (2019)

79. del Alamo, M., Hogan, D.J., Pechmann, S., Albanese, V., Brown, P.O., Frydman, J.: Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with Ribosome-Nascent chain complexes. PLoS Biol. 9(7), 1001100 (2011)

80. Diehn, M., Eisen, M.B., Botstein, D., Brown, P.O.: Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat. Genet. 25(1), 58–62 (2000)

81. Mayor, S., Riezman, H.: Sorting GPI-anchored proteins. Nat. Rev. Mol. Cell Biol. 5(2), 110–120 (2004)

82. Semenza, J.C., Hardwick, K.G., Dean, N., Pelham, H.R.: ERD2, a yeast gene required for the receptor-mediated retrieval of luminal ER proteins from the secretory pathway. Cell 61(7), 1349–1357 (1990)

83. Geva, Y., Schuldiner, M.: The back and forth of cargo exit from the endoplasmic reticulum. Curr. Biol. 24(3), 130–6 (2014)
Figures

**Figure 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.

**Figure 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.

**Figure 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 2-fold membrane enrichment. This includes mitochondrial and ER destined proteins. c. Nascent chains from genes showing 2-fold membrane enrichment that are not predicted to be mitochondrial. d. Nascent chains from genes showing less than 2-fold membrane enrichment but with a predicted ER signal sequence.

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

**Figure 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 2-fold membrane enrichment, which are considered posttranslationally targeted.

**Figure 6** Correlation of membrane enrichment scores between species. Scores are determined 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.

**Figure 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.
Figure 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.
Supplemental Figures

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

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 (bottom register) reveal both the open reading frame and the untranslated regions (UTRs). c. Flow-chart of the annotation pipeline.

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.

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.
Supplemental Tables

Table S1  Biosynthetic demands for proteins with unknown functions by predicted subcellular localization

| Genes (n) | Mean length (aa) | Nascent chains\(^a\) (%) | Ribosomes\(^b\) (%) |
|-----------|------------------|--------------------------|-------------------|
| **Cotranslationally Translocated\(^c\)** | | | |
| Endoplasmic reticulum | 113 | 446 | 7.0% | 19.0% |
| Cell membrane | 56 | 494 | 6.0% | 15.0% |
| Lysosome/Vacuole | 30 | 482 | 2.0% | 7.0% |
| **Posttranslationally Translocated\(^d\)** | | | |
| Extracellular | 13 | 246 | 79.0% | 44.0% |
| Cell membrane | 9 | 267 | 2.0% | 3.0% |
| Endoplasmic reticulum | 7 | 453 | 0.0% | 1.0% |

\(^a\)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 2-fold membrane enrichment and not predicted to be mitochondrial.
\(^d\)Proteins with less than 2-fold membrane enrichment and not predicted to be mitochondrial and contained a predicted signal sequence.