Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2019 April 19.

Published in final edited form as:

Cell. 2018 April 19; 173(3): 749–761.e38. doi:10.1016/j.cell.2018.03.007.

Evolutionary Convergence of Pathway-specific Enzyme Expression Stoichiometry

Jean-Benoît Lalanne[1,2], **James C. Taggart**[1], **Monica S. Guo**[1], **Lydia Herzel**[1], **Ariel Schieler**[1], and **Gene-Wei Li**[1,3,*]

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

SUMMARY

Coexpression of proteins in response to pathway-inducing signals is the founding paradigm of gene regulation. Yet, it remains unexplored whether the relative abundance of coregulated proteins requires precise tuning. Here we present large-scale analyses of protein stoichiometry and corresponding regulatory strategies for 21 pathways and 67–224 operons in divergent bacteria separated by 0.6–2 billion years. Using end-enriched RNA-sequencing (Rend-seq) with single-nucleotide resolution, we found that many bacterial gene clusters encoding conserved pathways have undergone massive divergence in transcript abundance and architectures via remodeling of internal promoters and terminators. Remarkably, these evolutionary changes are compensated post-transcriptionally to maintain preferred stoichiometry of protein synthesis rates. Even more strikingly, in eukaryotic budding yeast, functionally analogous proteins that arose independently from bacterial counterparts also evolved to convergent in-pathway expression. The broad requirement for exact protein stoichiometries despite regulatory divergence provides an unexpected principle for building biological pathways both in nature and for synthetic activities.

INTRODUCTION

A proteome is composed of regulatory modules, each with predefined yet disparate stoichiometry of coregulated proteins that participate in related biological pathways. Although aberrant expression stoichiometry caused by variations in gene copy number and regulatory elements is a common driver for both cellular dysfunction and evolutionary innovation (Harper and Bennett, 2016; Ohno, 1970), we lack a general framework for evaluating the impact of such perturbations: Is there a critical subset of proteins whose exact levels determine the activity of a pathway? Or does the overall stoichiometry comprising every protein need to be tuned precisely? Evolutionarily, the set-points of protein stoichiometry amidst all possible abundances could be contingent on the particular history of

*Correspondence to: gwli@mit.edu.
[3]Lead Contact

Author contributions JBL and GWL designed experiments and analysis; JBL collected data and performed analysis; JCT analyzed the yeast ribosome profiling data and collected ribosome profiling data for *V. natriegens*; LH provided method for 5' RACE; MSG and AS performed initial Rend-seq experiments; JBL and GWL wrote the manuscript.

Declaration of interests The authors declare no competing interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the host, or instead represent a universally optimized configuration of each pathway. Understanding the *in vivo* construction of biological pathways will provide foundational guiding principles for the interpretation of large-scale gene expression data and the engineering of biosynthetic processes (Albert and Kruglyak, 2015; Khosla and Keasling, 2003).

Of co-regulated proteins, it is generally thought that cells are particularly sensitive to an imbalance in the components of protein complexes (Oromendia et al., 2012; Papp et al., 2003). Consistent with the rationale that an excess of binding partners could have deleterious consequences, such as aggregation and off-target interactions, it was found that protein complex subunits are more likely to exhibit gene-dosage sensitivity (Papp et al., 2003). In aneuploid cells, a massive imbalance of binding partners is the likely driver for their general proteotoxic stress (Oromendia et al., 2012). These lines of evidence suggest that the production of protein complexes are normally set in proportion to their structural stoichiometry, which we have globally demonstrated using genome-wide quantitation of protein synthesis (Li et al., 2014). Such widespread proportional synthesis of multi-protein complexes highlights a clear rationale for precisely tuned expression among co-regulated proteins.

By contrast, protein stoichiometry of typical enzymatic pathways is viewed to afford greater variabilities. The exact abundance of each enzyme may have limited impact on the overall flux through a multi-step pathway (Fell, 1997; Kacser and Burns, 1981). Consistent with this model, enzymes are overrepresented in haplosufficient genes (Kondrashov and Koonin, 2004), and small down-regulation has no detectable effect on fitness under many experimental conditions (Keren et al., 2016; Peters et al., 2016). Furthermore, a recent systems-level analysis of metabolic flux suggested that cells generally exhibit excess enzyme capacities (Hackett et al., 2016). These lines of evidence lead to a common perception that enzymes are generally overproduced with minimal selective pressure on the surpluses. This perception, if true, further suggests that gene copy number variations and small regulatory changes for most enzymes should have negligible effects for the organism.

To provide a comprehensive view of the constraints on in-pathway protein stoichiometry, we set out to examine the conservation of expression in ancient biological pathways across evolutionarily distant bacterial species. Independent evolutionary trajectories over two billion years offer a stringent test over a wide range of conditions that cannot be accessed by experimental perturbations. Using ribosome profiling to quantify rates of protein synthesis, our analyses on both functional modules and operons identified quantitatively conserved pathway-specific expression stoichiometry: Functionally related or operon-associated proteins are synthesized at distinct rates that span orders of magnitude, but the difference between homologous proteins is typically much less than twofold in distant species. Interestingly, despite the conservation at the levels of genetic organization and protein stoichiometry, many co-regulated genes showed discordant mRNA levels. To dissect the regulatory mechanisms dictating differential expression within operons, we developed an end-enriched RNA-seq (Rend-seq) method to simultaneously (1) resolve the 3’ and 5’ boundaries of overlapping mRNA isoforms in operons with single-nucleotide resolution and (2) precisely quantify their relative abundance. Comparison of conserved bacterial gene

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

clusters revealed widespread remodeling of internal promoters and programmed transcription terminators that drives divergence of transcript architectures. Nevertheless, regardless of evolutionary paths, the stoichiometry of protein synthesis rates is precisely maintained through the mutually compensated strength of transcriptional and translational control elements.

Further corroborating this emerging view that biological pathways generally require exact protein composition, we found that the pathway-specific stoichiometry of proteins is also conserved in the eukaryotic budding yeast, which has dramatically different physiology, regulatory mechanisms, and protein properties compared to bacteria. Consequently, all molecular events of the central dogma, from transcription to translation to mRNA decay, must be tuned collectively to achieve preferred ratios of synthesis rates across orders of magnitude. Together, these results illustrate how fundamental principles of biological processes can be derived from a powerful combination of precise genome-wide quantitation with mechanistically driven analysis.

RESULTS

Comprehensive analysis of pathway-specific stoichiometry of protein synthesis rates

Using ribosome profiling to quantify the rates of protein synthesis (Li et al., 2014), we first compared the expression stoichiometry for ancient biological pathways that have evolved independently for >2 billion years in four divergent bacterial species. Ribosome profiling provides an accurate measurement of protein production during steady-state growth, as evidenced by stoichiometric synthesis among subunits of protein complexes that is otherwise obscured by noise in most other types of proteomic measurements (Li et al., 2014). The synthesis rates (Table S1) are directly proportional to steady-state abundances for proteins whose half-lives far exceed the cell doubling time and are thus diluted at the same rate by cell division (Alon, 2007; Li et al., 2014). In rapidly growing bacterial and yeast cells, this comprises the overwhelming majority of proteins (~99% for *E. coli* and 85% for yeast) (Christiano et al., 2014; Larrabee et al., 1980).

We carried out systematic pathway analysis for the Gram-positive and -negative model organisms *B. subtilis* and *E. coli*, respectively. Because many pathways are interconnected and differentially regulated as a consequence of divergent physiology, we focus on pathways that are largely self-contained, such as protein synthesis, DNA replication, and parts of metabolic pathways whose intermediates are neither precursors nor products of other pathways (STAR Methods). We further used curated organism databases to systematically identify homologous proteins that have become functionally divergent, which are subsequently removed from comparative analyses (Keseler et al., 2016; Michna et al., 2016). In total, we compared 21 pathways comprised of 302 homologous pairs (or groups if paralogs exist), with expression level ranging from 30 to 500,000 copies per cell (Li et al., 2014) and collectively constituting >45% of the proteomes by mass (in exponential growth phase with ~20 min doubling time for both *E. coli* and *B. subtilis*).

We found exquisitely conserved expression stoichiometry for nearly all of these ancient pathways. As an example, the synthesis rates among the 136 protein factors involved in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

mRNA translation vary by three orders of magnitude, but the majority are expressed at the same levels between *B. subtilis* and *E. coli* (<twofold difference for 87% of pairs, or 82% excluding ribosomal proteins, Fig. 1A). Not only are translation factors for elongation, initiation, and release synthesized at constant ratios relative to ribosomal proteins, other factors that do not directly interact with active ribosomes, such as aminoacyl-tRNA synthetases and RNA modification enzymes, also have conserved stoichiometry following distinct evolutionary history. On the other hand, expression is not conserved for homologous proteins that either are functionally divergent or have differential activity requirement, such as several ribosomal proteins in *E. coli* whose *B. subtilis* counterparts are not stably associated with the ribosome (STAR Methods). More broadly, conserved expression stoichiometry was also observed for 3 pathways involved in DNA maintenance (Fig. 1B) and 15 self-contained metabolic pathways (Fig. 1C, Fig. S1A). Only purine biosynthesis showed non-conserved expression stoichiometry (see Table S2 and STAR Methods for statistical testing for the significance of stoichiometry conservation).

The conservation of expression stoichiometry applies to the non-model, halophilic bacterium *Vibrio natriegens*. For this largely uncharacterized species, we did not systematically exclude homologous but functionally divergent proteins. Its comparison to *B. subtilis* and *E. coli* therefore potentially underestimates the fraction of functionally conserved proteins that have conserved stoichiometry. Nevertheless, most proteins (>86% compared to *E. coli*, >79% compared to *B. subtilis*) have <twofold deviation in expression relative to their respective pathway (Fig S1D-F).

The in-pathway stoichiometry also remains unchanged under different growth conditions, in spite of differential expression of pathways (Fig. S1B, C). For example, in *E. coli*, a three-fold difference in growth rate is accompanied by a three-fold change in the expression of every translation factor across a wide dynamic range of synthesis rates (Fig. S1B). The same trend is also observed for the slow-growing bacterium *Caulobacter crescentus* at various growth rates (96 and 146 min doubling time), although these ribosome profiling samples were collected under sub-optimal conditions (Schrader et al., 2014) (Fig. S1C-E, G, STAR Methods). The quantitative co-regulation highlights a strong demand for the cis-regulatory elements of all target genes to respond proportionally to a shared pathway-inducing signal. Such proportional expression changes amongst a large group of genes may be achieved by frequency-modulated activities of transcription factors, such as sigmas (Cai et al., 2008; Locke et al., 2011), or in some cases by co-transcription in polycistronic operons.

Conserved bacterial gene clusters produce conserved stoichiometry of proteins

We further take advantage of the fact that many functionally related proteins are co-expressed in operons to extend the pathway analysis beyond existing functional annotations. Among 80 conserved gene clusters encompassing 225 genes in *B. subtilis* and *E. coli* (including 118 additional genes not in the 302 manually curated homologs in the 21 pathways; STAR Methods), we found that 86% maintain conserved hierarchical expression: Proteins from neighboring genes are differentially synthesized with preferred stoichiometry ranging from 1- to >100-fold (Fig. 2AB, see Table S2 and STAR Methods for statistical testing). More closely related species, such as the Gammaproteobacteria *E. coli* and *V.*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

natriegens, have many more syntenic genes (224 clusters encompassing 706 genes) and, consequently, many more clusters with conserved expression (176 clusters encompassing 531 genes) (Fig. 2C). Together, 666 proteins in *E. coli* have conserved in-cluster stoichiometry in at least one other species considered, accounting for 56% of its proteome by mass (in exponential growth phase with ~20 min doubling time).

We found that exceptions to this trend, i.e. clusters with non-conserved stoichiometry, are also likely under positive selection (STAR Methods). For example, the ribosomal protein S12 cluster includes genes that encode the abundant translation factors EF-Tu (*tufA*) and EF-G (*fusA*), which often have additional copies of paralogous genes outside the cluster (Fig. 2D). Since different species have different copy numbers and expression of the additional paralogs, the copies in the S12 cluster are expressed differently so as to maintain similar levels of total EF-Tu and EF-G relative to ribosomal proteins (Fig. 2E). Overall, 8 out of the 11 clusters that have divergent stoichiometry between *B. subtilis* and *E. coli* can be attributed to similar genetic or other functional differences (STAR Methods).

Intriguingly, many homologous gene clusters with conserved expression have undergone structural changes during their respective evolutionary history, including insertion/deletion of non-conserved genes and potential regulatory sequences (Fig. 2A). Such structural remodeling raises the question of whether the cis elements responsible for conserved differential expression are still conserved.

Rend-seq allows precise mapping and quantitation of cluster-derived mRNA isoforms

To examine the regulatory mechanisms that differentiate expression within gene clusters, we first sought to resolve the precise units of transcription, which are often overlapping and difficult to quantify unit-by-unit (Cho et al., 2009; Conway et al., 2014; Nicolas et al., 2012). Here, we report an end-enriched RNA sequencing (Rend-seq) to mark both the 5’ and 3’ ends of mRNAs with single-nucleotide resolution, and also provide the expression levels across the body of mRNAs. We obtain end-enrichment by introducing sparse and random cleavage to RNAs (Stern-Ginossar et al., 2012): For each molecule of RNA subject to fragmentation with a low probability of cleavage per base ($p \ll 1$), the original 5’ end and 3’ end will always become a terminal nucleotide of one resulting fragment, whereas an internal position can only become a terminal nucleotide if cleavage occurs at that particular position. As a result, the original mRNA ends are overrepresented among the terminal nucleotides of fragmented RNAs, with an enrichment factor of $1/p$ compared to positions within the transcript body (Fig. 3A and STAR Methods).

We generate Rend-seq libraries by briefly subjecting purified RNA to zinc-mediated cleavage at 95°C. We then select short RNA fragments (15–45 nt), which permits quantitative conversion to cDNAs (Ingolia et al., 2012) (STAR Methods). The compact length of the cDNA library also allows for the use of short-read high-throughput sequencing to determine both 5’ and 3’ ends of each RNA fragment. The read counts for 5’-mapped and 3’-mapped fragments are then plotted separately (schematic in Fig. 3A). The resultant data for a simple, non-overlapping transcription unit display a single-nucleotide peak at each end of the mRNA, with a largely uniform coverage across the transcript body (Fig. 3B, S2B). Peak height relative to transcript-body coverage, i.e., end-enrichment, is inversely related to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

fragmentation time, and is followed by proportional increases/decreases in read density downstream (Fig. 3B and Fig. S2C).

We individually validated the abilities of Rend-seq in measuring RNA levels and in capturing the precise position of transcript ends. Using the extensive *B. subtilis* and *E. coli* literatures, we found that the locations of terminal nucleotides predicted by Rend-seq are consistent with >500 previously published 5' ends, mapped by primer extension, and >1,000 3' ends determined by genome-wide 3'-mapping strategies (STAR Methods and Mendeley Data) (Dar et al., 2016; Keseler et al., 2016; Mondal et al., 2016; Sierro et al., 2008). Novel 5' and 3' ends identified in this study are also confirmed by independent methods (STAR Methods and Mendeley Data, Fig. S3). Meanwhile, RNA levels estimated by Rend-seq are consistent with other gene expression datasets (Nicolas et al., 2012), as well as Northern blotting results in this work (STAR Methods and Mendeley Data). Therefore, Rend-seq not only shares similar advantages with several recent high-throughput end-mapping methods (Dar et al., 2016; DiChiara et al., 2016; Irnov et al., 2010; Mendoza-Vargas et al., 2009; Sharma et al., 2010), but also has a unique advantage in its abilities to determine both 5' and 3' ends, and to quantitate RNA levels in a single experiment.

Rend-seq further enables quantitative profiling of complex transcript architecture for bacteria gene clusters. With >50-fold end-enrichment (Fig. 3B, Fig. S2C) and limited variations in internal read coverage, minor mRNA isoforms nested in major ones give rise to detectable peaks in 5'- and 3'-mapped reads. This is illustrated by the *hbs* locus in *B. subtilis* which has been shown to have multiple 5' isoforms (Daou-Chabo et al., 2009) (Fig. 3C, Mendeley Data). Most of these 5' ends are difficult to detect by conventional RNA-seq or high-density microarrays due to large variations in internal signal and lack of end-enrichment, e.g., (Brinsmade et al., 2014; Nicolas et al., 2012).

To systematically resolve complex mRNA isoforms, we developed an automated pipeline to identify isoform boundaries based on both peaks and step-wise changes in Rend-seq signals (STAR Methods, Fig. S2A, D-F). The obligatory "peak shadows," which arise from fragmented RNAs that share an aligned 5' (or 3') end at the peak and have narrowly distributed 3' (or 5') ends, are computationally removed for data visualization (Fig. 3A, Fig. S2G-I, and STAR Methods). We further developed a mathematical framework for reconstructing mRNA isoforms and their abundances after the boundaries are identified (Fig. S2J and STAR Methods). Northern blot analysis confirmed the reconstructed transcript architectures for gene clusters (Fig. S3 B, F).

Widespread transcription terminator read-through differentiates expression of operonic genes

Systematic analysis of mRNA isoforms (Fig. S2) revealed that partial transcription termination is a major driver for tuning differential expression among neighboring genes: Many gene clusters are punctuated by 3'-mapped peaks followed by an incomplete decrease in read density downstream (e.g., Fig. 4A, C-D). These 3'-mapped peaks are often associated with the characteristic upstream sequence for factor-independent, i.e. intrinsic, termination (STAR Methods). Perturbations that disrupt terminator sequences also abolish

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the shorter isoforms, supporting that these intra-cluster 3'-mapped peaks are generated by intrinsic transcription termination (Fig. 4B and Fig. S3 B, F and J).

With little or no *de novo* promoter activity dedicated to downstream genes, differential expression (both mRNA and proteins) between these adjacent genes is quantitatively tuned by the read-through fraction of transcription terminators (Fig. S3H). In *B. subtilis*, we identified 167 intergenic "tuned" terminators that singly or in combination set the expression of 276 genes, including 33 essential genes and several previously characterized cases (Commichau et al., 2009; Mondal et al., 2016; Shunsuke et al., 1983) (excluding riboswitches or other known attenuators, Fig. 4E). Similar prevalence of tuned terminators is found for the other bacterial species included in this study (STAR Methods, Table S3).

The decrease in mRNA read density following tuned terminators ranges from <twofold to 100-fold. The difference in isoform abundance is primarily driven by terminator read-through, with minor contributions from differential RNA stability for a small number of cases (Fig. S3K-L, S, STAR Methods). Across all intrinsic terminators, the spectrum of read-through fraction is distinguished by the length of the U-tract upstream of the 3' end, whereas the stability of the stem-loop structure—the other defining feature of intrinsic terminator—is weakly correlated with read-through (Fig. 4E, Fig. S3M-S). The same trend also applies to the other bacterial species (Fig. S3M-O, R). Overall, our data suggest that transcription terminators are not simply all-or-none switches for regulating or insulating genes, but are also commonly programmed as a fine dial for differentiating the levels of operonic genes.

Extensive compensation between transcriptional and post-transcriptional activities

Comparison of Rend-seq data between *B. subtilis*, *E. coli, V. natriegens, and C. crescentus* showed extensive remodeling of transcript architecture in conserved gene clusters in spite of the similar protein expression stoichiometry. This is illustrated for a translation-related operon which is expressed as a contiguous four-gene mRNA (*rpsP-rimM-trmD-rplS*) in *E. coli*, whereas in *B. subtilis*, the cluster is differentially transcribed using both a tuned terminator (after *rpsP*) and an internal promoter (before *rplS*) (Fig. 5A). The two middle genes (*rimM* and *trmD*) encode for an rRNA-maturation factor and a tRNA-modification enzyme, respectively. These two proteins are naturally required—and produced—at much lower levels compared to the other two ribosomal proteins (*rpsP* and *rplS*) (Fig. 5B). As indicated by the Rend-seq and ribosome profiling data, the differential expression is mainly achieved at the transcriptional level in *B. subtilis*, in contrast to the translational control in *E. coli* (Fig. 5C). Notably, the mRNA secondary structures that are known to sequester the ribosome binding sites for *rimM* and *trmD* in *E. coli* (Burkhardt et al., 2017; Wikström et al., 1992) are absent in *B. subtilis*, consistent with the dramatic differences in translation in compensation for transcriptional changes (STAR Methods).

Another representative example for the gain and loss of cis regulatory elements is illustrated for the cluster containing the ribosome binding factor RbfA (Fig. 5D). A tuned terminator is found upstream of the gene *rbfA* in *E. coli*, but downstream in *B. subtilis*, with a net result of a discordant mRNA level for the *rbfA* gene. The difference is compensated translationally to produce the same ratio of proteins (Fig. 5E, F).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Across all pairs of species considered, we found pervasive remodeling of transcript architectures, including half of gene clusters that have conserved expression stoichiometry (Fig. 5G, Data S1). Even between the more closely related Gammaproteobacteria *E. coli* and *V. natriegens*, the majority of 177 conserved clusters—including those containing *rimM* and *rbfA*—have gained or lost promoters and terminators (Fig. S4 and Data S1). In all these scenarios, the strength of molecular events is precisely tuned to reach the convergent stoichiometry of protein production.

In addition to the conserved gene clusters that experienced divergence of cis-regulatory elements, many pathways analyzed here have their operon linkage completely altered between *E. coli* and *B. subtilis* (subset shown in Fig. 6). The redistribution of genes along the chromosome not only requires new promoters and ribosome binding sites, but also leads to drastic changes of gene dosage due to multi-fork replication (Fig. S5). This diverse set of events—gain and loss of regulatory elements, dissolution and formation of operons, changes in gene copy numbers—likely led to temporary imbalance of expression stoichiometry for one or multiple proteins (Fig. S4C), which were quantitatively reverted through compensatory evolutionary changes.

Convergence of pathway-specific expression stoichiometry without sequence similarity

Given that differences in regulatory mechanism do not impede conservation of protein expression patterns, we next explored whether the same expression stoichiometry extends to the eukaryotic budding yeast (Weinberg et al., 2016). Between *E. coli* and yeast, mRNA translation and glycolysis are two major pathways that remain largely conserved at the molecular level. We found that the expression stoichiometry for each pathway is also quantitatively conserved: the rates of synthesis maintain a linear relationship over three orders of magnitude (Fig. 7A, B), with 96% of protein pairs differing by less than twofold despite extensive differences in gene dosage (Fig. 7C, see Fig. S6 for comparison across all species). Notably, bacteria and yeast share several factors that are functional analogs without sequence, or even structural, homology (e.g., elongation factor eEF1B/EF-Ts (Andersen et al., 2000) and release factor eRF1/RF1, RF2 (Kisselev, 2002)). Their consistent synthesis rates suggest that convergent evolution has occurred not only for their biochemical functions, but also for the expression levels relative to the respective pathways.

DISCUSSION

Enzymatic pathways and their regulation have been thoroughly characterized over many decades. Yet remarkably little has been considered about the quantitative composition of their protein effectors *in vivo*. Here we used pathway-centric analyses to demonstrate the strong preference for these proteins to be produced at a defined stoichiometry, irrespective of how the expression is achieved at the mechanistic level.

A narrow zone of preferred stoichiometry challenges the notion that enzyme levels need not to be set precisely as long as they are in excess to the flux requirement. In fact, the precise proportion of each enzyme within a pathway shares striking similarity with obligate multi-protein complexes. However, unlike protein complexes with well-defined structural arrangements, the rationales for preferred enzyme stoichiometry are obscure. Because

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

divergent species have evolved the same expression, what constrains the pathway-specific stoichiometry is likely to be independent of most cellular properties that vary across the tree of life, such as metabolite concentration, optimal growth temperature, and subcellular compartmentalization. To our knowledge, the determinant of optimal enzyme levels is understood in only a few cases even for well-characterized *E. coli* (Dekel and Alon, 2005; Eames and Kortemme, 2012; Klumpp et al., 2013; Li et al., 2014). Our study suggests that a cost-benefit tradeoff for protein production is widespread across entire pathways, providing an important design principle for metabolic engineering and highlighting the need to better understand cellular economy.

From the experimental perspective, the observed evolutionary convergence suggests that perturbations to gene expression—both over and under the endogenous levels—should lead to discernable phenotypes. Recent developments in CRISPRi and array-based promoter synthesis made it possible to finely manipulate the expression of single genes in systematic ways (Gilbert et al., 2013; Keren et al., 2016; Peters et al., 2016). These studies revealed that the growth rate or fitness of a cell is often insensitive to small perturbations to the expression of a gene, and that the effects are only present in specialized environmental conditions. Because each pathway is co-regulated as a whole in changing environments, we propose that the preferred protein stoichiometry is constrained by the most sensitive conditions, which makes it difficult to identify in limited experimental settings. Furthermore, the effects of single-gene perturbation can be masked by compensatory changes in other genes that are paralogous or partially overlapping in function (DeLuna et al., 2010; Ihmels et al., 2007; Kafri et al., 2006). Consistently, we observed that the conservation of stoichiometry operates at the level of homologous protein groups and not on individual genes (Fig. 1, 2E, and 7). Taken together, the comparative analysis at the protein level allows us to access a much broader space of selective conditions and complements perturbation studies at the single-gene level.

From the evolutionary perspective, our observation of divergent regulatory strategies underlying conserved protein stoichiometry dramatically extends recent findings based on human individuals and other closely related eukaryotic species ($<10^6$ years of separation) (Artieri and Fraser, 2014; Battle et al., 2014; Khan et al., 2013; McManus et al., 2014). Unlike the highly conserved protein sequences (>99% identity) among the species of those studies, proteins analyzed here not only share limited sequence similarity (median amino acid identity of 42% between *B. subtilis* and *E. coli* for proteins compared in Fig. 1), but also have different gene copy numbers. The divergence of sequence and gene dosage should be accompanied by ample opportunity for independent evolution in their biochemical properties and structures—both of which could change preferred expression levels. Conservation of pathway-specific protein stoichiometry could imply that the biochemistry and structure of these distant homologs have either converged to or remained at optima despite changes in sequence. In particular, for proteins that carry out the same activities but have different evolutionary origins, both functional properties and expression levels are subject to convergent evolution.

Because of the widespread requirement on protein production rates, the strength of each underlying regulatory element in a given species must be tuned at a quantitative level. In

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

bacteria, it is well established that differential expression within gene clusters can be set by differences in the initiation rates of transcription and translation. Here, our high-resolution end mapping and isoform quantitation demonstrated that RNA polymerase read-through at transcription terminators is also a common mechanism to tune the expression between neighboring genes. Although the molecular processes involved in these regulatory components have been characterized in depth, our knowledge remains at a qualitative level in contrast to the quantitative precision required in the cell. Rend-seq offers a simple path towards identifying and characterizing these evolutionarily tuned elements at the genomic scale, providing an orthogonal approach to current high-throughput efforts using mutagenized reporters to assess the impact of each residue (Cambray et al., 2013; Chen et al., 2013).

Comparative analysis of genes and genomes has proven extremely powerful in revealing key features dictating protein functions and regulation. Our pathway-centric comparison of mRNAs and proteins expands the scope of evolutionary analysis and provides a new framework for probing the construction principles of biological activities. The broadly preferred protein stoichiometry suggests that well-placed protein levels are integral design considerations for each enzymatic pathway, whereas differences in regulatory strategies may merely reflect the distinct evolutionary history. More generally, our results signal a severely underdeveloped field of pathway optimization under cellular constraints, which will require a holistic view of the cell (Karr et al., 2012; Scott et al., 2010) and precise measurements of biosynthetic activities.

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and request for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Gene-Wei Li (gwli@mit.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Strain construction

E. coli—The *rnb* and *pnp* deletions were transferred to K-12 MG1655 from the Keio collection strains (Baba et al., 2006) following standard P1 phage transduction.

B. subtilis—Strains were constructed using standard genetic cloning protocols for *B. subtilis* (Harwood and Cutting, 1990). Perturbation to putative intrinsic terminators (removal of stem plus U-tract, denoted by *Δ*sU, and removal of U-tract only, denoted by *Δ*U) at the native locus without otherwise perturbing the endogenous operon structure were generated by constructing synthetic DNA by *in vitro* isothermal assembly (NEBuilder HiFi DNA Assembly Master Mix, New England Biolabs) followed by allelic exchange via natural transformation. Genetic perturbations were confirmed by Sanger sequencing of PCR products straddling the altered terminators.

Strains and growth conditions

For ribosome profiling experiments, *Bacillus subtilis* subsp. subtilis str. 168 and *Δ*sU *ylqC/ylqD* terminator mutants (Fig. S3G) were grown in LB. *Vibrio natriegens* (strain NBRC

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

15636) was grown in MOPS complete medium (Neidhardt et al., 1974) (Teknova) supplemented with 3% NaCl. In all cases, an overnight liquid culture (started from a single colony from a fresh plate) was diluted to an approximate OD_{590} of 3×10^{-4} (about 10 000-fold) into fresh media (250 mL for *B. subtilis*, 100 mL for *V. natriegens*). The cultures were kept in a 2.8 L flask at 37°C with aeration (200 rpm) until OD_{590} reached 0.3.

Protein synthesis rates in *Escherichia coli* derived from ribosome profiling were obtained from (Li et al., 2014). Protein synthesis rates in *Caulobacter crescentus* and *Saccharomyces cerevisiae* were derived in the current work from the published raw ribosome profiling datasets from (Schrader et al., 2014; Weinberg et al., 2016), see section “Data conversion to protein synthesis rates" below.

For Rend-seq experiments, *E. coli* (K-12, MG1655 wild-type as well as *pnp* and *rnb* knockouts) was grown in MOPS complete medium (Neidhardt et al., 1974) (Teknova). *B. subtilis* subsp. subtilis str. 168 (and various mutants (Koo et al., 2017) as specified) was grown in LB. *Vibrio natriegens* (strain NBRC 15636) was grown in MOPS complete medium (Neidhardt et al., 1974) (Teknova) supplemented with 3% NaCl. *C. crescentus* strain ML76 (Evinger and Agabian, 1977) was grown in PYE medium. Overnight liquid cultures were diluted to an approximate OD_{590} of 3×10^{-4}, or about 10 000-fold (deletions for *rph, yhaM* and *rnr* in *B. subtilis* were diluted to a starting OD_{590} of 10^{-3}) into 20 ml fresh media in 125 mL flasks at 37°C (*E. coli, B. subtilis* and *V. natriegens*) or 30°C (*C. crescentus*) with aeration (200 rpm) until OD_{590} reached 0.3.

METHOD DETAILS

Pathway-specific expression stoichiometry

Ribosome profiling—For ribosome profiling experiments in *B. subtilis*, we follow the protocol from (Li et al., 2014) with slight modifications. Briefly, 250 mL of cell culture (OD_{590} = 0.3) was rapidly filtered at 37°C by passing through a nitrocellulose filter with 200 nm pore size (Supor Membrane Disc Filters, Sigma Aldrich). Cell pellets were rapidly collected using a prewarmed metal table crumber, flash frozen in liquid nitrogen, and combined with 650 μL of frozen droplets of lysis buffer (10 mM $MgCl_2$, 100 mM NH_4Cl, 20 mM Tris pH 8.0, 0.1% NP-40, 0.4% Triton X-100, 100 U/μL DNase I (Sigma-Aldrich), 1 mM chloramphenicol). Cells and lysis buffer were pulverized in 10 ml canisters (10 mL grinding jars, QIAGEN) prechilled in liquid nitrogen using TissueLyser II (QIAGEN) for 5 cycles of 3 min at 15 Hz. Pulverized lysate was thawed on ice and clarified by centrifugation at 20,000 rcf for 10 min at 4°C. 5 mM $CaCl_2$ was added to 0.5 mg of RNA from the clarified lysate containing, which was then digested with 750 U of micrococcal nuclease (Roche) at 25°C for 1 hr. The reaction was quenched by adding EGTA to 6 mM and moved on ice.

The monosome fraction following nuclease digestion was collected using sucrose gradient and the RNA extracted by hot-phenol extraction. Ribosome-protected mRNA fragments were isolated by size excision on a denaturing polyacrylamide gel (15%, TBE-Urea, Thermo Fisher Scientific). Fragments with size ranging from 15 to 45 nucleotides were excised from the gel. The 3’ end of footprints was dephosphorylated using 20 units of T4 polynucleotide kinase (New England Biolabs) at 37°C for one hour. Three picomoles of footprints were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ligated to 100 pmole of 5’ adenylated and 3’-end blocked DNA oligo (linker1, Table S4) using truncated T4 RNA ligase 2 K277Q at 37°C for 2.5 hr (25% PEG 8000). The ligated product was purified by size excision on a 10% TBE-Urea polyacrylamide gel (Thermo Fisher Scientific). cDNA was generated by reverse transcription using Superscript III (Thermo Fisher Scientific) at 50°C for 45 min with primer ocj485 (Table S4), and isolated by size excision on a 10% TBE-Urea polyacrylamide gel (Thermo Fisher Scientific).

Single-stranded cDNA was circularized using 100 U of CircLigase (Epicenter) at 60°C for 2 hr (additional 100 U added after the first hour). Ribosomal RNA fragments were removed using biotin-linked DNA oligos (Table S4) and MyOne Streptavidin C1 Dynabeads (Thermo Fisher Scientific). After being purified using isopropanol precipitation, the remaining cDNA was amplified using Phusion DNA polymerase (New England Biolabs) with o231 primer and indexing primers (Table S4). After 6–10 rounds of PCR amplification, the product was selected by size excision on a 8% TB polyacrylamide gel (Thermo Fisher Scientific).

Sequencing for the *B. subtilis* ribosome profiling experiment was performed on an Illumina HiSeq 2000 or NextSeq500. 3’ linker sequences were stripped. Bowtie (Langmead et al., 2009) v. 1.0.1 (options -v 1 -k 1) was used for sequence alignment to the reference genome NC_000964.3 obtained from NCBI Reference Sequence Bank. To deal with non-template addition during reverse transcription, reads with a mismatch at their 5’ end had their 5’ end re-assigned to the immediate next downstream position. The footprint reads with size between 20 to 42 nucleotides in length were mapped to the genome using the center-weighted approach. In the end, the number of ribosome footprints mapped to coding sequences was 16.8 M, with 2382 genes with more than 128 footprint reads mapped.

For ribosome profiling in *V. natriegens*, 100 ml of cell culture (OD_{590} = 0.3) was rapidly filtered at 37°C by passing through a nitrocellulose filter with 450 nm pore size (Supor Membrane Disc Filters, Sigma Aldrich), cell pellets were rapidly collected and flash frozen in liquid nitrogen. Cells pellets from two 100 mL cultures were combined for future steps. Lysis, ribosome footprint purification, size selection and dephosphorylation then proceeded exactly as described above (Li et al., 2014). Dephosphorylated ribosome footprints were then converted to cDNAs using the SMARTer smRNA-seq Kit (Clontech) following the manufacturer’s instructions. No rRNA removal was performed for the *V. natriegens* ribosome profiling experiment.

Sequencing for the *V. natriegens* ribosome profiling experiment was performed on an Illumina HiSeq 2000. 3’ poly A sequences were stripped. Bowtie v. 1.0.1 (options -v 2 -k 1) was used to align the resulting sequences to the reference genome (on the two chromosomes, CP009977.1 and CP009978.1) obtained from NCBI Reference Sequence Bank. The footprint reads with size between 20 to 42 nucleotides in length were mapped to the genome. For each footprint read, the 5’ end was given a weight of 1 and shifted 11 nt downstream. A different weighting approach was used because the above cDNA library preparation does not preserve 3’ end information. In the end, the number of ribosome footprints mapped to coding sequences was 2.8 M. Given the lower depth of the *V. natriegens* data, we used throughout a threshold of 16 footprint reads (post-Winsorziation) per gene (25% relative error from counting noise). 2494 genes had more than 16 footprint reads mapped.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Data conversion to protein synthesis rates—Protein synthesis rates for *E. coli* (K-12, MG1655) grown in MOPS complete and minimal medium were obtained from (Li et al., 2014). Ribosome profiling data for *C. crescentus* (strain NA1000 grown in PYE and M2G media) and *S. cerevisiae* (strain BY4741 grown in YPD) were obtained respectively from (Schrader et al., 2014) and (Weinberg et al., 2016), and analyzed as described below.

Protein synthesis rates in bacteria were determined as described in (Li et al., 2014), with slight modifications. Briefly, the protein synthesis rate is proportional to the mean ribosome footprint density across a gene as determined from ribosome profiling. Two main assumptions are required for this proportionality to be valid: (1) the average translation elongation rate must be the same across different mRNAs, and (2) the fraction of prematurely terminated ribosomes (ribosome drop-off) must be small. The validity using of ribosome profiling as a precise measure of differential translation is thoroughly documented (Li et al., 2014; Weinberg et al., 2016). The mean ribosome footprint read density across a gene was calculated by excluding the first and last five codons (to avoid biases from increased ribosome footprint densities arising from initiation and termination).

Because liquid culture of *C. crescentus* cannot be rapidly filtered, the cells were harvested differently from the other species (Schrader et al., 2014). Chloramphenicol was added to the culture before pelleting (Schrader et al., 2014). A potential consequence of this approach is that translation and the rates of protein synthesis might be perturbed during the drug treatment. For example, one difference of the *C. crescentus* dataset with other ribosome profiling datasets analyzed is a sharp peak in ribosome footprint density near the start codon of genes (for the first 20 codons), likely arising from the chloramphenicol treatment prior to harvesting. To avoid biases from this 5' peak, the footprint reads mapping to the first 20 codons of genes were excluded from the averaging window for synthesis rate measurement. Other biases, harder to identify or to correct for, could also be present.

To correct for ribosome pausing at Shine-Dalgarno like sequences (Li et al., 2012), we followed the same approach as in (Li et al., 2014) on the *B. subtilis, V. natriegens* and *C. crescentus* ribosome profiling datasets. Specifically, the average ribosome occupancy downstream of each hexanucleotide sequence was determined. A line was fit through these ribosome occupancies versus the affinity of the respective hexanucleotide sequence for the anti-Shine-Dalgarno sequence. This line was then used to adjust the ribosome occupancy at each position in each gene: at each position, the measured occupancy is divided by the expected pause duration (in relative units) based on the strongest hexanucleotide sequence at the 6–11 bases upstream. The adjusted ribosome occupancy is no longer correlated with the anti-Shine-Dalgarno affinity. Residual variation not accounted for was removed with 90% Winsorization (bottom and top 5% ribosome occupancies replaced by the 5 and 95 percentile values respectively) in the *B. subtilis* and *C. crescentus* datasets, and by 98% Winsorization in the *V. natriegens* dataset (different Winsorization in *V. natriegens* since we are not using the center-weighted approach to map the ribosome footprints). Given the lower depth of the *V. natriegens* and *C. crescentus* datasets, we used a threshold of 16 (relative error of 25% from counting noise) footprint reads (post-Winsorization) mapping to a gene as our expression cutoff for these species (2494 and 3001 genes above threshold in each species respectively).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We saw no evidence of a 5' downward ramp (increased density at the 5' end of coding sequences) in our *B. subtilis* and *V. natriegens* datasets. The mild 5' ramp (to be distinguished from the sharp 5' peak discussed above) in *C. crescentus* was corrected for as follows (Li et al., 2014): the ribosome occupancy profiles for genes longer than 175 nt with density above 0.5 read/nt were smoothed by a travelling average (window size 100 nt) and normalized to the average of the first 100 nt (after the first excluded 20 codons at the 5' end, see above). The median of these normalized profiles at each position (forming a meta-gene profile) was used to fit an exponential decaying function of the form $f(x) = A + (1 - A)\, e^{-x/D}$ where x is the position from start of the profile, D is the decay length and A is the offset level. The fit parameters obtained for the *C. crescentus* data were $A = 0.66 \pm 0.01$, $D = 520 \pm 20$ nt in PYE medium and $A = 0.74 \pm 0.02$, $D = 1200 \pm 100$ nt in M2G, where the range is the 95% confidence bound for the fit. Not much variability in the fit parameters was observed in a twofold density threshold and window size around the chosen values. The decaying function $f(x)$ was then used to correct for the increased ribosome density at the 5' end of coding sequences. Specifically, the ribosome occupancy at position x along a gene of total length L (excluding regions not considered at the gene's 5' and 3' ends in our analysis) was weighted by a factor $f(L)/f(x)$.

The corrected ribosome occupancy was then used to compute the mean density, which was taken as directly proportional to the protein synthesis rate. Overall, the above corrections (combined from ramp and Shine-Dalgarno like sequences) were small, with the 14th and 86th percentile (for all genes) of correction factors respectively of 0.76 and 1.12 for *B. subtilis*, 0.83 and 1.12 for *V. natriegens*, 0.82 and 1.12 for *C. crescentus* in PYE and 0.83 and 1.10 for *C. crescentus* in M2G. Our conclusions are unaffected if uncorrected synthesis rate (using directly the winsorized ribosome footprint density) are used. Corrected mean ribosome densities across genes can be found in Table S1.

For *S. cerevisae*, raw ribosome profiling reads were downloaded from (Weinberg et al., 2016) and aligned to the yeast genome (the genome sequence for strain S288C, of which strain BY4741 is a derivative, was obtained from (Cherry et al., 2012)). To account for reads mapping across intron splice junctions, reads not aligning to the yeast genome were aligned to the nucleotide sequences of open reading frames (including 50 nt flanking regions) and added to previously aligned reads. As a result, a ribosome footprint read profile for each gene was generated. In order to avoid the increased footprint read density due to translation initiation and termination, we excluded the first and last 30 nt of open reading frames from our analysis. To address the unexplained density at the 5' end of genes (5' ramp), we followed the correction protocol of (Weinberg et al., 2016), with slight modifications. Briefly, footprint reads profiles were smoothed by averaging read counts within each codon and then normalized by the mean read density within the gene. For each position x relative to the beginning of the coding sequence, the median of the normalized footprint read counts at x across all genes profiles (only including genes of length more than x) was obtained and the resulting profile smoothed by a 50 nt median sliding window (henceforth: the metagene profile). The metagene profile reveals a downward 5' ramp (near-exponential) in footprint read density, with a decay length scale of about 300 nt and initial amplitude of about 1.5× the long-distance plateau value. To correct for this additional footprint density at the 5' end of genes, each gene footprint read counts profile was divided position-by-position by the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

metagene profile. For positions beyond 1600 nt, the average of the metagene profile between 1500 to 1600 nt was used for the correction (to avoid variability in the metagene profile coming from the increasingly low number of genes longer than 1600 nt). Synthesis rates were then calculated as the mean of corrected footprint read densities across genes, which can be found in Table S1.

We can estimate an upper bound on the uncertainty of our protein synthesis rate measurement using values obtained for ribosomal proteins (RPs). Because all but one RPs are present with equimolar stoichiometry in mature ribosomes, the distribution of synthesis rates across these RPs should reflect both biological deviation from strict proportional synthesis and measurement uncertainty. Because RPs are generally shorter than typical proteins, the measurement uncertainty is also more sensitive to variations in elongation rates and potential cloning biases. The range of expression (normalized to the median) we observed (10^{th} to 90^{th} percentile) for ribosomal proteins was: 0.9 to 1.27 for *E. coli*, 0.62 to 1.25 for *B. subtilis*, 0.68 to 1.62 for *V. natriegens*, 0.81 to 1.28 for *C. crescentus* and 0.86 to 1.21 for *S. cerevisiae*. This suggests that our measurement uncertainty for protein synthesis rates is substantially lower than twofold for highly expressed genes (where counting noise is small).

For the quantification of closely related paralogs (the different copies of EF-Tu, c.f., Fig. 2E), see dedicated section "Synthesis rate for the two EF-Tu copies".

The translation efficiency (TE) (see Mendeley Data) was computed as the final synthesis rate divided by the mRNA abundance determined from Rend-seq and normalized to the median TE across all genes of the species.

Curation of conserved pathways in bacteria—Functional annotation for *E. coli* and *B. subtilis* was used as a starting point to curate a list of self-contained pathways. We focused on two major sets of enzymatic pathways: those involved in the central dogma, and parts of the metabolic network whose intermediates are not shared with others pathways. For central dogma related enzymes, proteins involved in transcription (not shown in figures, see Table S2), translation, and DNA maintenance were identified using the EcoCyc (Keseler et al., 2016) and SubtiWiki (Michna et al., 2016) databases. DNA maintenance includes three sub-pathways: replication, condensation/segregation, and repair/recombination. For metabolic pathways, our analysis was guided by a recent reconstruction of metabolic network for *E. coli*, as detailed below.

To match proteins for comparison between *E. coli* and *B. subtilis*, we relied on a combination of functional annotations and sequence homology. To assess sequence homology, amino acid sequences of genes for *E. coli* (K-12 MG1655, NC_000913.3) and *B. subtilis* (subsp. subtilis str. 168, NC_000964.3) were downloaded from GenBank. Duplicate gene names were renamed with by adding an index (e.g., *insC* to *insC1, insC2*, etc.). BLASTP databases for each genome were created with command “makeblastdb" (using the BLAST+ suite (Agarwala et al., 2016)). Gene pairwise BLASTP scores were obtained with the command “blastp", with a significance threshold (option -evalue) of 10^{-7}, selecting output format (option -outfmt) 6. A homology matrix storing the pairwise blast scores

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

between all proteins in *E. coli* and *B. subtilis* was created. The homology matrix was then used to obtain a connectivity matrix for a bipartite graph by a binary thresholding (pairwise BLAST score above 45 as a permissive cut). The connected components of the graph correspond to cliques of homologous genes used for further analysis.

To match/group proteins in homology cliques related to central dogma processes, we manually curated functional annotations downloaded from EcoCyc (Keseler et al., 2016) and SubtiWiki (Michna et al., 2016). List of proteins involved in DNA maintenance, transcription and translation were obtained by parsing the list of genes and their functions with exhaustive search terms (for example for translation, sub-categories ribosome maturation: 'ribosome assembly', 'ribosomal small subunit assembly', 'ribosomal large subunit assembly', 'ribosomal small subunit biogenesis', 'ribosomal large subunit biogenesis', 'ribosome biogenesis', 'rRNA processing'). Homology cliques (see previous paragraph) containing at least one gene for the functional category of interest were obtained, leading to 141 translation related cliques (104 of which were one-to-one) and 87 DNA maintenance related cliques (37 of which were one-to-one). For the transcription class, we restricted our attention to components of the RNA polymerase and conserved elongation/termination factors. Guided by the sequence homology, sub-grouping from within each clique was manually performed by verifying the detailed functional characterizations. Ambiguous groups of paralogs, left out of our comparison, are detailed in Table S2. Genes with divergent functions were also removed: three ribosomal proteins (*E. coli*'s genes *rplY, rplI, rpsA*) were left out of our comparison as they were not found in the cryo-EM structure of the *B. subtilis*' ribosome (Sohmen et al., 2015). Incidentally, their expression levels in *B. subtilis* were over an order of magnitude lower than other ribosomal proteins. *B. subtilis'* ribosome assembly factor *engB* and its *E. coli* homolog *yihA* were also left out, as *yihA* is suggested to function as a cell division protein. Homologs of the DNA maintenance class shown to have divergent functions and thus excluded from our comparisons were: *E. coli*'s *lrp* (transcription factor) and *B. subtilis' lrpC* (DNA binding/bending involved in DNA repair), *E. coli*'s *ftsX* (cell division membrane protein) and *B. subtilis' ftsX* (sporulation initiation), *E. coli*'s *smf* (no known role in DNA repair/recombination) and *B. subtilis' dprA* (involved in *recA* function), *E. coli*'s *recT* (part of the rac prophage) and *B. subtilis' yqaK* (no known function), *E. coli*'s *rnhA* (RNase HI) and *B. subtilis' ypdQ* (known to have no RNase H like activity). Final groups used for comparison in Fig. 1A, B from the main text can be found, with synthesis rates, in Table S2.

In order to compare the synthesis rates of proteins in metabolic pathways despite changes in physiology and local reorganization of the metabolic networks over evolution, we restricted our attention to regions of the metabolic network for which intermediates are not shared with other pathways (see below for minor exceptions). Specifically, the stoichiometry matrix of a recent metabolic network reconstruction for *E. coli* K-12 MG1655 was obtained from (Orth et al., 2011). "Simple" metabolites, defined as metabolites participating in two reactions (with both in and out fluxes), were identified from the stoichiometry matrix. 947 out of 1805 metabolites classified as simple metabolites. Clusters of simple metabolites were formed by connecting two simple metabolites participating in the same reaction. Clusters involving a transport reaction were excluded. The resulting 92 clusters and respective enzymes were curated for conservation of both protein sequence (see below), enzyme function, and local

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

network topology in *B. subtilis* using annotations from EcoCyc (Keseler et al., 2016), Subtiwiki (Michna et al., 2016) and KEGG (Kanehisa et al., 2016). Many of the clusters were specific to Gram-negative bacteria (e.g., lipopolysaccharide biosynthesis). In the curation process, we both extended clusters (e.g., in the case where a substrate was not identified as simple because of the same reaction catalyzed by two independent enzymes) and corrected mistakes in the flux balance stoichiometry matrix based on recent functional characterizations (e.g., missing reaction EC 4.1.99.22 in molybdopterin biosynthesis). We also restricted our attention to pathways expressed in both species in our conditions (e.g., excluded histidine biosynthesis and *de novo* pyrimidine synthesis). Core glycolysis was included despite possible internal fluxes to other pathways, given that in rich medium the largest flux is expected to be towards pyruvate (for ATP production), and not towards anabolic pathways. Indeed, flux measurements indicate that only a small fraction of the flux is diverted to side pathways even in minimal medium (Emmerling et al., 2002; Blank et al., 2005).

We note the following additional exceptions for the selection of self-contained set of reactions. First, for peptidoglycan biosynthesis, peptidoglycan recycling in *E. coli* was not considered, as a strain with the enzyme responsible for the flux (Mpl) knocked out has no growth or morphological defect as well as no decrease in cell peptidoglycan content. Second, for fatty acid biosynthesis, the flux out of malonyl-CoA (not present in *B. subtilis*) towards biotin biosynthesis (EC 2.1.1.197) was not considered, as the enzyme (BioC) catalyzing that step was over 25× lower than that leading to fatty acid biosynthesis (FabD). Third, for CoA biosynthesis, the possible flux out of dephospho-CoA by CitG (EC 2.4.2.52) was neglected because CitG is not measurably expressed in our condition. The weak MocA homolog in *B. subtilis*, PucB (molydbopterin biosynthesis, EC 2.7.7.76) was expressed below our read depth threshold and so was not included in the comparison (MocA and PucB expression were nevertheless within 2×). Finally, diaminopimelic acid biosynthesis differs at three intermediate steps in *E. coli* and *B. subtilis*. The chemical step catalyzed by homologous enzymes DapD (*E. coli*, EC 2.3.1.117) and DapH (*B. subtilis*, EC 2.3.1.89) are different but functionally related, and so this comparison is included (the level of the two non-homologous intervening enzyme pairs between DapD/DapH and DapF are also within 2×, not shown). Reaction EC 2.7.1.71 (chorismate pathway) is catalyzed by two enzymes AroL and AroK in *E. coli*. AroK has however been shown to have much lower activity than AroL (100×) in *E. coli* and is thus left out of the comparison. In addition, *E. coli*'s *mog* (EC 2.7.7.75) has no known counterparts in *B. subtilis*. Finally, reaction EC 3.1.3.104 (riboflavin pathway) is known to be catalyzed by three enzymes in *B. subtilis*, one of which has about ten times higher specific activity than the other two (*ycsE*). Thus, only *ycsE* is included in the comparison. See Table S2 for details.

Reactions where not all enzymes responsible for catalysis have sequence homology above our 45 BLASTP threshold were excluded from Fig. 1C of the main text, but included in Table S2 to broaden the comparison (e.g., *fbaA* and *fbaB* in *E. coli* catalyze reaction EC 4.1.2.13, but *fbaB* is not homologous to *B. subtilis' fbaA*).

The distribution of sequence identity between our final list of homologs between *E. coli* and *B. subtilis* was obtained by directly taking the sequence identity (as obtained from the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

BLASTP alignment) for one-to-one pairs and as the average identity of all pairs in the case of homology cliques. In the rare cases where multiple independent alignments were found for a single protein (e.g., for RpoB and RpoC), the maximum identity among the alignments was taken as the identity of the two proteins (providing an upper bound). The median identity was 42% (10th percentile: 30% identity, 90th percentile: 58% identity).

The functional annotations, protein characterizations, and metabolic network reconstructions for *C. crescentus* and especially *V. natriegens* are not as complete as those of *E. coli* and *B. subtilis*. To group proteins for comparison of synthesis rates in *V. natriegens* and *C. crescentus*, we thus entirely relied on sequence homology to members of our curated pathways in *B. subtilis* and *E. coli* (see above), with an approach similar to that described above, by identifying clusters of homologous proteins (connected components of the connectivity matrix obtained from thresholding to higher than 45 BLASTP score and including at least one member from each of the four bacterial species considered). Clusters of homologous proteins containing proteins in our curated list from the *E. coli/B. subtilis* comparison were retained. To deal with groups with multiple homologs in *V. natriegens* and *C. crescentus*, we retained all proteins within 50 BLASTP score of the maximum homology score pairing. If a protein was present in more than one group, it was removed from the groups where it appeared not as a unique member (e.g., for close homologs such as ParE and GyrB).

The final result was manually inspected for inconsistencies (e.g., strong homologs missed because of the 50 BLASTP score difference threshold). The following manual corrections were made: in the translation class, *asnC* was removed from the comparison given its absence in *C. crescentus, lysS* in *C. crescentus* was added to CCNA_00757 for the lysine tRNA synthetase group and the three homologs for *fusA* (the two proteins named *fusA* and PN96_01780) were included in the *fusA* (EF-G) group for *V. natriegens* (renamed for Fig. 2 and S4). In the DNA maintenance class, the two genes named *dnaE* in *C. crescentus* were included in the *dnaE* group. Integration host factors *ihfA* and *ihfB* from *C. crescentus* were excluded from the *hupA* and *hupB* homology group because of known different functions. PN96_02165 in *V. natriegens* was added to the *recQ* group as a result of its high homology (like *B. subtilis, V. natriegens* seems to have two similar copies of this DNA helicase). Finally, both the *ada* and *alkA* homology groups include two proteins in *E. coli* and *B. subtilis*, a constitutive and an inducible copy (Morohoshi et al., 1993), with sequence homology lacking between the constitutive and inducible copies the *alkA* group in *E. coli*. Because of these complications and the lack of fine characterization for these proteins in *C. crescentus* and *V. natriegens*, these two groups were omitted from our comparison.

In the end, we had 134, 50, 8 and 86 groups of homologs to compare in the four bacterial species for translation, DNA maintenance, transcription, and the metabolic pathways considered, respectively. Final groups for comparisons with synthesis rates can be found in Supplementary Data 2. These final groupings were also used to compare synthesis rates in rich and minimal media in *E. coli* and *C. crescentus*. Resulting comparison of protein synthesis rates can be seen in Figure S1.

Final groupings and synthesis rates can be found in Table S2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Comparison between yeast and bacteria—To match proteins for comparison between *E. coli* and *S. cerevisiae*, we followed a similar approach as the *E. coli/B. subtilis* comparison. Amino acid sequences of genes for *E. coli* (K-12 MG1655, NC_000913.3) were downloaded from GenBank and from (Cherry et al., 2012) for *S. cerevisiae* (for strain S288C, the parent strain of BY4741). As before, we first relied on sequence homology (BLASTP) to obtain homology cliques. Homology cliques containing at least one gene from the translation functional category (form the EcoCyc annotation) were obtained, leading to 69 translation related cliques. Guided by the sequence homology, sub-grouping from within each clique was manually performed by verifying the detailed functional annotations (Cherry et al., 2012). Conserved ribosomal proteins between yeast and bacteria (based on structural analyses) were taken from (Ban et al., 2014). Ambiguous groups of paralogs, left out of our comparison, are detailed in Table S2. Genes with characterized divergent functions were also removed (see below). Given the extensive characterization of *E. coli* and *S. cerevisiae* translation systems, factors with known conserved functions but lacking sequence homology were also added (see below). In all, we could make 73 comparisons of synthesis rates for proteins involved in translation. For core glycolysis, enzymes involved were obtained from KEGG and manually confirmed. Sequence homology between involved proteins was assessed by BLASTP. We note that in some reactions were catalyzed by more than one enzyme in *E. coli*, with one of the enzyme not being homologous to its *S. cerevisiae* counterpart (specifically: PfkB, FbaB and GpmM). These are included in Fig. 7B. Final groups used for comparison in Fig. 7 from the main text can be found, with synthesis rates, in Table S2.

The following homologs with divergent functions were excluded from our comparison (convention in what follows: *S. cerevisiae* protein/*E. coli* protein): Rnt1p/Rnc are RNase III and responsible for rRNA processing in both species, but Rnt1p uses snoRNPs for processing and has exonuclease activity (in contrast to Rnc). Both nuclease also have additional non-conserved system's level functions in RNA decay in the two species. Dis3p/Rnr: multi-functional ribonucleases with no strict conserved function. Dis3p has both endo and exonuclease activity (Rnr only has exonuclease activity) and is part of the exosome (Rnr is not part of the degradosome). Ncs6p/TtcA: tRNA modification. Ncs6p thiolates U34 of tRNAs whereas TtcA thiolates position C32. Ncl1p, Nop2p/Rsmb, RsmF: Ncl1p methylates tRNAs whereas Nop2p/RsmB, RsmF methylate rRNA, but have different targets in *S. cerevisiae* and *E. coli*.

In the current work, we define functional analogs (stars in Fig. 7A) as proteins with a specific and conserved function, yet with less than 45 pairwise BLASTP score. This includes proteins with truly no sequence homology, but also proteins with some, but very limited, sequence homology. Figure 7A includes the following functional analogs. Five of the aminoacyl tRNA synthetases had very limited sequence homology (for leucine, glycine, tryptophan, tyrosine and one of the heterodimer subunit for the phenylalanine synthetase, see below), while having a clearly conserved function. Comparison of translation initiation and termination factors is confounded by differences in the two pathways in prokaryotes and eukaryotes. Still, given the conserved function of peptide release, RF1, RF2 were compared to eRF1 (Kisselev, 2002). We note that RF3 and eRF3 were not included as they have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

notably different function (e.g., eRF3 tightly interacts with eRF1 in contrast to RF3, eRF3 is essential in contrast to RF3). For initiation, we included the two pairs of homologs (eIF1A/IF1 and eIF5B/IF2). We note that part of the function (binding to initiator tRNA) of IF2 is performed by the non-sequence homolog eIF2. We only included eIF5B in the comparison (we note however that the aggregated synthesis rate of eIF2 and eIF5B normalized to the median ribosomal protein are within 10% of the synthesis rate of that of IF2). We also included eIF1/IF3, as these are recognized to be functional analogs and can complement each other in heterologous assays (Lomakin et al., 2006). Guanine nucleotide exchange factor for elongation factor Tu (EF-Ts in bacteria, eEF1B in eukaryotes) has conserved functions but no conserved sequences or structure between prokaryotes and eukaryotes (Andersen et al., 2000). eIF5A/EF-P have limited sequence homology, but are structural homologs and are believed to play a similar role in translation elongation (Saini et al., 2009).

Finally, we corrected for the synthesis rates of proteins with different oligomeric states in *E. coli* and *S. cerevisiae*. The glycyl-tRNA synthetase has an α_2 structure in *S. cerevisiae* and a $\alpha_2\beta_2$ structure in *E. coli*. The synthesis rate of the *E. coli* synthetase (sum of synthesis rates of the two subunits) was divided by two. eIF5A is known to form homodimers, in contrast to *E. coli*'s EF-P, so that its synthesis rate was divided in two. TruA forms a homodimer in *E. coli*, in contrast to *S. cerevisiae*'s homolog Deg1p. TruA's synthesis rate was thus divided by two. See Table S2 for details.

Final groupings and synthesis rates for proteins shown Fig. 7A, B can be found in Table S2. Comparison with other bacteria (based on matching to *E. coli*) is also included and shown in Fig. S6.

To generate Fig. 7C, the proteins compared in Fig. 7A, B were grouped based on the number of proteins in each species, and the distribution of the ratios of their synthesis rates (normalized by the overall pathway relative expression factor) displayed. For comparison, the expression ratio for all one-to-one sequence homologs between *E. coli* and *S. cerevisiae* (cutoff BLASTP score of 45, or at least 100 score difference to nearest homolog), with mitochondrial genes excluded (as assessed by manual curation using descriptions from (Cherry et al., 2012) are shown in Fig. 7C. 177 homologs satisfying the above criteria were expressed in both species (>128 reads mapping to the coding sequence), see Table S2 for the list with synthesis rates.

Expression in gene clusters: species pairs—To investigate more systematically conservation in expression stoichiometry, we compared expression in conserved gene clusters in all bacterial species pairs considered.

The identification of conserved gene clusters was performed in two stages. First, we operationally assigned homologous genes from bidirectional best BLASTP (Agarwala et al., 2016) hits (and requiring minimum BLASTP score of 45). Second, these homologs were clustered based on their co-conserved chromosomal locations (and conserved strand orientation). Specifically, the intergenic distance $d_{i,j}$, defined as the minimum start to stop distance between the two genes, was used. Intergenic distance instead of gene midpoints

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

distance was used to avoid the confounding effect of variable gene sizes for clustering threshold selection (see below).

For each species pair (e.g., *B. subtilis* and *E. coli*), we constructed a graph where each node corresponded to a homolog (from the bidirectional best BLASTP hit). Homologs i and j in this graph were connected according to the connectivity matrix $C_{i,j}^{Bsub-Ecol}$, where $C_{i,j}^{Bsub-Ecol}$ equals 1 (connected) if $d_{i,j}^{Bsub} < \delta$ and $d_{i,j}^{Ecol} < \delta$ ($d_{i,j}$ was taken as infinite for homologs i and j on different strands, the current analysis discarding gene inversions), and 0 otherwise. δ is the distance threshold for spatial clustering. Connected components of this graph were the conserved gene clusters in our analysis.

We use co-localization on the chromosome as a proxy for functional relatedness. Genes for species pairs with larger evolutionary distances will have had more opportunities to become separated. To account for this, we used a different distance threshold δ for the different species pairs. To rationally set δ, we performed the gene clustering and expression categorization (see below) for varying δ (between 0.1 and 20 kb) for all species pairs. For all species pairs, we observed a characteristic spatial distance δ^* beyond which spurious clusters are formed, with a sharp decrease in the fraction of clusters with conserved expression stoichiometry. δ^* then represents a plausible length-scale beyond which synteny could not be taken as an indicator for functional relatedness. We thus used δ^* as a distance threshold for the rest of our analysis ($\delta^*_{Ecol-Vnat} = 400$ nt, $\delta^*_{Ecol-Bsub} = 1200$ nt, $\delta^*_{Ecol-Caulo} = 600$ nt, $\delta^*_{Vnat-Bsub} = 1200$ nt, $\delta^*_{Vnat-Caulo} = 300$ nt, $\delta^*_{Bsub-Caulo} = 1200$ nt.). Note that the δ^*'s are on the order of the mean gene size, making it critical to use intergenic, and not midpoint, gene distance for this analysis.

We classified the resulting conserved gene clusters in species pairs based on the synthesis rates of their genes as described below. We denote the synthesis rate for gene j in cluster i by $k_{i,j}^{Ecol}$ and $k_{i,j}^{Bsub}$ for *E. coli* and *B. subtilis*, respectively.

First, we discarded from further analysis clusters in which some synthesis rates were below our depth threshold. Table S2 provides a complete list of resulting clusters with measured synthesis rates (as well as classifications, see below). For retained clusters a line of slope 1 was fitted (least square) through the logarithm of the synthesis rates (same procedure as Fig. 1 from the main text). This allowed us to account for an overall differential expression for genes in each cluster between species pairs (corresponding to an offset from the main diagonal on a log-log plot). Denote this overall factor for cluster i in *E. coli* and *B. subtilis* (notation for this species pair throughout) by $\alpha_i^{Ecol-Bsub}$. The resulting maximum deviation factor, $D_i^{Ecol-Bsub}$, from stoichiometric synthesis in cluster i is then:

$$D_i^{Ecol-Bsub} := 2^{a_i^{Ecol-Bsub}}, \text{ with } a_i^{Ecol-Bsub} := \max_j \left| \log_2 \left(\frac{k_{i,j}^{Ecol}}{\alpha_i^{Ecol-Bsub} k_{i,j}^{Bsub}} \right) \right|.$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For example, imagine hypothetical genes A, B and C that are expressed at rates 20, 10 and 5 (AU) in *E. coli* and rates 8, 4 and 2 in *B. subtilis*. The stoichiometry of expression would then be perfect ($D = 1$), but with an overall multiplicative factor of $\alpha = 2.5$.

The range in expression for genes in cluster i is defined as the maximum over the minimum synthesis rate, or

$$r_i^{Ecol} := \frac{\max_j(k_{i,j}^{Ecol})}{\min_j(k_{i,j}^{Ecol})}, \quad r_i^{Bsub} := \frac{\max_j(k_{i,j}^{Bsub})}{\min_j(k_{i,j}^{Bsub})}$$

for the two species respectively. The synthesis range for the cluster is taken as the geometric mean of that in the two species:

$$r_i^{Ecol-Bsub} = \sqrt{r_i^{Ecol}\ r_i^{Bsub}}.$$

Using the above quantities, gene cluster i was categorized using the quantities defined above as follows:

- Highly conserved stoichiometry (~1:1 production): if $r_i^{Ecol} < 2$ and $r_i^{Bsub} < 2$ (no more than twofold difference in synthesis for all genes in the cluster).
- Highly conserved stoichiometry (unequal stoichiometry): if $r_i^{Ecol-Bsub} \geq 2$ and $D_i^{Ecol-Bsub} < 1.5$ (≥twofold range of synthesis and within 50% of same expression stoichiometry).
- Partially conserved: $D_i^{Ecol-Bsub} < 2$ (expression stoichiometry within 2 ×).
- Divergent: if none of the above applies.

Categorization for *E. coli* and *B. subtilis* is shown in Fig. 2B and results for all pairs detailed in Fig. 2C. Enrichment for conserved expression stoichiometry is highly significant ($p < 0.005$ across all pairs, see details in section “Expression conservation and synteny"). For example, within this categorization, 87% of the genes (86% of clusters) show some level of conservation of protein expression stoichiometry in the *E. coli* to *B. subtilis* comparison.

An approach avoiding the categorization described above is to consider the distribution of deviations from stoichiometric production in each cluster, i.e., the distribution of

$$D_{i,j}^{Ecol-Bsub} := \frac{k_{i,j}^{Ecol}}{\alpha_i^{Ecol-Bsub} k_{i,j}^{Bsub}}.$$

Across species, between 79% (*V. natriegens* vs. *C. crescentus*) and 91% (*B. subtilis* vs. *C. crescentus*) of genes in considered clusters had $0.5 < D_{i,j} < 2$ ($p < 0.002$ across all pairs, see section “Expression conservation and synteny").

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Interestingly, most gene clusters with divergent production could either be attributed to characterized functional or structural differences, or to presence of paralogous proteins outside of the conserved clusters in our *B. subtilis* and *E. coli* comparison. We now list these rationalized expression differences.

First, the ribose binding protein from the ribose ABC transporter has a much higher expression stoichiometry in *E. coli* compared to *B. subtilis*. In Gram-negative *E. coli*, substrate binding proteins for ABC transporters reside in the periplasm, whereas the Gram-positive *B. subtilis* has the substrate binding proteins on the plasma membrane. Divergence in production is then not surprising, given that these complexes reside in different cellular compartments (i.e., *B. subtilis* has no periplasm) and that the geometry of the diffusion problem for substrate binding is different. (Notably, we observe this expression divergence in other ABC transporters not showing up in our gene cluster analysis due to broken linkage in *B. subtilis*.) The glycerol facilitator *glpF*, which resides in the inner membrane in *E. coli*, also has a very different level relative to the glycerol kinase *glpK* compared to what we observe in *B. subtilis*.

As other examples: *hfq* and small RNA regulation are different between the two species (much higher level observed in *E. coli*). Also, chemotaxis signaling gene *cheW* has a paralog (*cheV*) outside the conserved cluster in *B. subtilis*. Ribosomal protein S1 from *E. coli* (*rpsA*) is not part of the core ribosome in *B. subtilis* (*ypfD*) (Sohmen et al., 2015), consistent with the over 15-fold difference in expression between them. Similar rationales are detailed in Table S2 for other enzymes.

Overall, extensive functional characterization of proteins in *E. coli* and *B. subtilis* suggests that genuine biological differences underlie many of the expression divergence observed in conserved clusters. Further accounting for paralogous copies, 8/11 clusters categorized as divergent can be reasonably rationalized.

Despite the caveats arising from systematic analysis based on sequence homology (exemplified above), the comparison of expression in conserved gene clusters, independent from functional annotation, strengthens the observations made for curated pathways.

Definition of four species gene clusters—To systematically compile information about transcript architecture remodeling across the four bacterial species considered, we identified conserved gene clusters (looser definition, see below) across these species.

First, we generated the set of homologs used for spatial clustering. To do so, a graph where each protein from each species corresponded to a node was generated. Nodes (proteins) in this graph were connected if they were pairwise best BLASTP hits (with BLASTP score > 45) in their respective species pairs (proteins from the same species were not connected). Connected components of this graph with only four members, with one member from each species, served as our operational definition of “one-to-one homologs" for our four species comparison.

Manifest omissions/incorrect assignments from the above analysis were corrected manually (denoted as stars in Fig. S4 and Data S1): the main copy of *rpsN* was taken in place of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

rpsNB in *B. subtilis*, the *rbfA* homolog in *C. crescentus* (CCNA_00036) was included, the *rnpA* homolog in *C. crescentus* (CCNA_00807) was included, *atpE* was included, subunits B and B' of the ATP synthase F_0 complex from *C. crescentus* (CCNA_00370 and CCNA_00371) were included, and ribosomal protein L35 from *C. crescentus* (CCNA_01098) was included. We note that the S12 operon (including EF-G and EF-Tu) displayed in Fig. 2E and S4F–H was assembled outside of the current described framework due to the multiple paralogous copies involved.

To spatially cluster the above homologs, we applied less stringent clustering criteria than for the pairwise analysis detailed above, as our main purpose to compile information (Data S1). We constructed a graph where each node i was a homolog in our analysis, and defined the following connectivity matrix $C_{i,j}$ for the graph: $C_{i,j}$ is equal to 1 if $d_{i,j} < \delta$ (with $\delta = 10$ kb) in three out of four species, and 0 otherwise. $d_{i,j}$ now corresponds to the distance between gene midpoints, the difference between intergenic and midpoint distance being unimportant given the distance threshold chosen. The subset of conserved clusters with divergent operon architectures are displayed in Fig. S4 and Data S1.

Rend-seq: method and quality control

Rend-seq library generation—For RNA extraction, 5 mL of cell culture (OD_{590} = 0.3) was added to 5 mL of cold (−30°C) methanol, mixed by inversion and spun down at 3000 rcf for 10 min at 4°C. The supernatant was decanted and the cell pellet frozen at −80°C. RNA was extracted using the RNAeasy kit (QIAGEN) with on-column DNase treatment. Ribosomal RNA was depleted using the MICROBExpress kit (Thermo Fisher Scientific). The resulting RNA was purified by isopropanol precipitation and resuspended in 40 μL of 10 mM Tris 7.0.

To assess the molecular nature of the 5' ends observed in Rend-seq, treatment by a 5'-monophosphate sensitive exonuclease was performed prior to fragmentation (no rRNA removal was performed for these samples) for some libraries (5'-exo treated) (DiChiara et al., 2016). Briefly, 7.5 μg of total extracted RNA was resuspended in 17 μL of water and mixed with 2 μL of 10× buffer A and 1 μL of exonuclease (Terminator 5'-Phosphate-Dependent Exonuclease, Epicentre). The reaction was incubated at 30°C for 60 min, followed by isopropanol precipitation. The exonuclease treated RNA was then fragmented as for other Rend-seq libraries (below).

We used RNA fragmentation reagents (Thermo Fisher Scientific) to fragment the RNA. Specifically, RNA was incubated at 95°C for 2 min in a PCR thermocycler and placed back on ice for at least 1 min. 4.4 μL of 10× fragmentation buffer was added to the RNA on ice. The RNA was then fragmented by heating the solution to 95°C (on a pre-heated PCR thermocycler) for 25 s unless otherwise specified (for the experiment with variable fragmentation time, the RNA was left at 95°C for 25 s, 50 s, 100 s, and 200 s; for deletions of *rnr, rph* and *yhaM* in *B. subtilis*, the RNA was fragmented for 105 s). 5 μL of stop buffer was then quickly added, and the resulting solution mixed by pipetting and placed back on ice. The fragmented RNA was then purified by isopropanol precipitation. Fragments in the 15–45 nt size range were obtained by excision on a 15% TBE-Urea polyacrylamide gel (Thermo Fisher Scientific). The remainder of the cDNA library preparation (3'

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

dephosphorylation, 3’ adapter ligation, reverse transcription, circularization and polymerase chain reaction) proceeded identically to that of ribosome profiling as described in the ribosome profiling protocol above (Li et al., 2014), except for the rRNA removal, which was performed prior to fragmentation for Rend-seq (rRNA fragments are removed post circularization in ribosome profiling).

Sequencing was performed on an Illumina HiSeq 2000 or NextSeq500. 3’ linker sequences were stripped. Bowtie v. 1.0.1 (options -v 1 -k 1 unless otherwise specified) was used for sequence alignment to the reference genome NC_000964.3 (*B. subtilis*), NC_000913.2 (*E. coli*), CP009977.1 and CP009978.1 (the two chromosomes of *V. natriegens*, and CP001340.1 for *C. crescentus* obtained from NCBI Reference Sequence Bank. To deal with non-template addition during reverse transcription, reads with a mismatch at their 5’ end had their 5’ end re-assigned to the immediate next downstream position. The 5’ and 3’ ends of mapped reads between 15 and 45 nt in sizes were added separately at genomic positions. For additional treatment of the data, see the section on peak shadow removal below. Given the high GC content of *C. crescentus*, which leads to higher prevalence of multiple possible mapping positions for 15 nt long reads, we mapped reads in the range 17 to 45 nt for that species.

For the purpose of quantification of mRNA for paralogous copies of EF-Tu and gene cluster comparisons (Fig. S4, and Data S1), we mapped the reads using Bowtie options -v 1 -m 2 -k 2 (no more than one mismatch and no more than two alignments). The reads (within the same size ranges as above) with a single reported alignment were mapped as before. Reads with two reported alignments were treated as follows. If only one of the reported alignment had no mismatch, the perfect alignment was kept. If both reads had a single mismatch but one had a mismatch corresponding to non-template addition at the 5’ end, the alignment with non-template addition mismatch was kept. Otherwise (two perfect alignments, two alignments each with non-template addition mismatches, or two alignments with mismatches not corresponding to non-template addition), reads were discarded from this analysis.

To quantify mRNA levels (rpkm), we computed the mean read density of the 1% winsorized 3’-mapped reads across genes (leaving a 45 nt gap from both start and stop codons) (see Mendeley Data). Note that this quantification does not take into account possible internal 5’ or 3’ ends internal to the gene of interest.

Peak finding strategy—Given the end-enrichment provided by Rend-seq (see later section for mathematical description of the end enrichment), ends of transcription units can be identified by finding sharp peaks in the data. Let *n(x)* denote the read counts of either 5’ or 3’-mapped reads as a function of genomic position *x*. Consider the following modified z score (referred to as the peak z score), which exclude the central position for the calculation of the mean and standard deviation of the signal (essential for peak identification):

$$z_{g,\ell}(x) := \frac{n(x) - \langle n(x) \rangle_{g,\ell}}{\sigma_{g,\ell}(x)},$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where

$$\langle n(x)\rangle_{g,\ell} := \frac{1}{N_{g,\ell}} \sum_{g < \left|x-y\right| \le \frac{\ell}{2}} n(y), \quad \sigma_{g,\ell}(x) := \sqrt{\left\langle n(x)^2 \right\rangle_{g,\ell} - \langle n(x)\rangle^2_{g,\ell}}.$$

g above is the central half gap and ℓ is the averaging length. $N_{g,\ell}$ is simply the number of positions satisfying the constraint $g < \left|x-y\right| \le \frac{\ell}{2}$.

The modified z score can be rapidly computed for all genomic positions by fast Fourier transforms. As the z score tracks deviation from the mean signal, peaks at ends of transcription units are marked by a large z score. As an illustration, the resulting large tail of the z score distribution across all genomic positions is shown in Fig. S2A for different fragmentation time with g = 2 nt (chosen based on observed peak width, see below and Fig. S2B) and ℓ = 100 nt. The distribution in modified peak z score shows two populations, as evidenced by the sudden change in slope near $z \approx 7$, which corresponds to positions that are peaks versus non-peaks. The distinction becomes more evident with higher end-enrichment, i.e. lower fragmentation times. These distinct populations make it possible to achieve a low false-positive rate in automated peak identification by thresholding based on the peak z scores. We used a z score of 12 (unless otherwise noted). To remove false-positives due to counting noise, only peak positions with neighboring average read density above a certain threshold ($\langle n(x)\rangle_{g,\ell}$ > 0.25 read/nt) are retained. Finally, if there were multiple positions within a 3 nt window with high z-score (e.g., for a peak of >2 nt wide), only the position with maximum z score within the immediate neighborhood was retained.

Quantification of end enrichment—We first identified 5’ and 3’ peaks (positions with modified peak z score >12, see previous section). Peaks with an additional modified z score above 7 within 150 nt were discarded from the end-enrichment analysis to avoid complexity arising from multiple nearby density steps. Peaks in rRNA and tRNA genes, as well as those in small RNAs, were also discarded from the current analysis. The height h of the peak at position x was computed as the sum of reads counts at positions with signal above the half maximal value of the peak in the $x \pm 5$ nt neighborhood of the peak (see next section for the distribution of peak width). The step size was computed as the difference between the mean read density upstream u (between $x - \ell - g$ and $x - g$) and downstream d (between $x + g$ and $x + g + \ell$), with gap g = 5 nt. Less than 1% of peaks have steps of the opposite sign as expected (e.g., negative for 5’ mapped reads or positive for 3’-mapped reads on the forward strand) probably arising from nearby missed peaks. For those anomalies, the width of the averaging regions are extended to be as long as possible to increase precision (manually) or discarded if there were clearly missed nearby peaks. Ultimately, 976 3’ peaks and 1003 5’ peaks pass the various cuts for the *B. subtilis* data at 25 s fragmentation time. Fig. S2C displays h as a function of $(u-d)$ (5’) and $(d-u)$ (3’), confirming that peaks in the signal are accompanied by density steps in Rend-seq (up for 5’ peaks and down for 3’ peaks). The end enrichment was then computed as $h/(u-d)$ for 5’ peaks and $h/(d-u)$ for 3’ peaks. The median from such an analysis for both 5’ and 3’ peaks for *B. subtilis*’ libraries generated at with different fragmentation times are displayed in Fig. 3B in the main text.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Resolution of RNA ends—The ends of transcripts are identified by finding peaks in Rend-seq data. The precision with which these ends are determined then depends on the width of the peaks in Rend-seq. For each Rend-seq peak identified in our automated analysis, we calculated the peak width as full-width at half maximum (FWHM), i.e. the number of positions in the ±5 nt neighbourhood of the peak with greater than half the read counts at the peak. The distributions of peak widths for *B. subtilis* and *E. coli* Rend-seq with 25 s fragmentation time are shown in the form of pie charts in Fig. S2B. Most peaks are either a single nucleotide wide (for 5' ends) and one or two nucleotides wide (for 3' end), consistent with the fact that intrinsic termination can happen at more than one neighboring positions. The sharpness of these peaks confirms that Rend-seq provides transcript boundary information with single-nucleotide precision (see validation section for assessment of accuracy).

Removing peak shadows—Inspecting Rend-seq data, regions of high read density downstream of 5' peaks in the 3'-mapped channel (and upstream of 3' peaks in the 5'-mapped channel) can be seen. These regions of high read density next to peaks, which we refer to as peak "shadows", arise from the high end-enrichment and narrow read length distribution used in our library preparation strategy (Fig. 3A, S2G). Take shadows in the 3'-mapped channel for example, the signal comes from the abundant species of RNA fragments that have the original 5' ends and variable 3' ends distributed between 15 and 45 nt downstream (see later section for the mathematical derivation of the signal, with analytical solution shown in Fig. S2G). Given their well-understood origin, peak shadows were removed to improve clarity when displaying the Rend-seq data.

To do so, we took an iterative approach in two steps. First, the positions with peaks in the 5' and 3' channels were identified (modified peak z score >12) from the raw Rend-seq data. The end enrichment for each of these positions was computed as the ratio of the peak height h to the downstream (5') or upstream (3') read density ρ. The end enrichment factor ($h\rho^{-1}$), position, strand, and type of peak (5' or 3') were then stored for each identified peak. Second, the 5' and 3'-mapped reads were summed again at all genomic positions, but this time with a corrected weight factor for reads whose ends correspond to peaks. For example, if a read of length ℓ has its 5' end within 2 nt of an identified peak position x (as identified in the first step), it was still counted as one read for the 5'-mapped channel. However, the same read is down-weighted by the end-enrichment factor of the corresponding peak (i.e. 1/end-enrichment) in the 3'-mapped channel. This efficiently gets rid of peak shadows, c.f., Fig. S2I vs. H) (with the caveat that if a peak is missed by the z-score thresholding approach, its shadow will not be removed). This approach also gets rid of spurious over-amplification of specific reads by PCR (i.e., "jackpot" amplification) as such events lead to closely paired 5' and 3' peaks in Rend-seq. The corresponding over-amplified reads thus get weighted inversely to their over-amplification in the shadow removal stage.

The pile-up files (read counts at each positions, .wig format) with peak shadows removed as described above can be found (together with the raw data) in the Gene Expression Omnibus series for the main Rend-seq datasets of the current publication. The Rend-seq data presented in Fig. 3–5 of the main text as well as supplementary figures has shadows removed.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Detection sensitivity—The high (over 50-fold) end-enrichment (e.g., Fig. 3B and Fig. S2C) increases Rend-seq's sensitivity towards detecting rare transcription units in complex operon architectures. In particular, end-enrichment provides a definitive advantage for nested transcripts, transcription units of low abundance which have at least one end internal to a major mRNA isoform. The section below presents a detailed analysis as to why nested transcripts are challenging to identify by searching for shifts in mean read density alone. By systematically characterizing the variability in the Rend-seq data, we also find that nested transcripts with an abundance 10% of the major isoform in which they are embedded can reliably be detected from Rend-seq data with less than 5% false positive probability per gene (this sensitivity is partly limited by sequencing depth for 3' ends).

Fig. S2K illustrates the difficulty of identifying rare nested transcripts without end-enrichment for a constructed hypothetical example: a mean shift in read density is added to the Rend-seq signal for gene *rplA* from *B. subtilis* (as a representative noisy signal) to simulate a minor nested transcript. Without end enrichment, nested transcription units can be identified by shifts in the mean read density. To find shift in the mean read density, one can average the signal over a length L_{av}. Given noise in read counts of σ, a minor isoform of abundance m (in units of read counts per unit position) will be detectable provided $m \gg (L_{av}/l_{corr})^{-1/2}\ \sigma$, where l_{corr} is the correlation length of the Rend-seq signal (length scale over which fluctuations in signal occur). This relationship comes from the fact that averaging decreases the noise as one over the square root of independent measurements. Averaging a signal with correlation length l_{corr} over a length L_{av} provides L_{av}/l_{corr} independent measurements. The difficulty in identifying isoforms purely from shifts in mean signal arises from the square root and the fact that $l_{corr} \approx 4$ nt > 1 nt for Rend-seq (auto-correlation function not shown). In contrast, with end-enrichment (Fig. S2L), the minor isoform is detectable provided m (e.e.) $\gg \sigma$, where now (e.e.) stands for the end-enrichment factor. With 50 × end-enrichment, one would need $L_{av} = (e.e.)^2\ l_{corr} \approx 10$ kb to achieve a similar sensitivity without end-enrichment, much longer than most transcripts. Indeed, it is difficult to identify the minor isoform from even at an abundance of 50% of the major one without end-enrichment (Fig. S2K), but quite feasible at an abundance of 10% with 50× end-enrichment (Fig. S2L).

In order to quantify sensitivity, we systematically characterized Rend-seq's variability in the read count signal across the body of genes. To that end, the read count distribution for both 5' and 3' ends of mapped reads were collected for all genes with mean read density above a selected threshold (varied, see below) and no identified internal end. To check for internal ends, peaks were identified (modified z score above 7 to be conservative). If a change in density was seen across the peak (for 5' end: upstream density smaller than downstream density with 95% confidence from bootstrap in ± 100 nt region including a 5 nt gap around the peak, vice versa for 3' end), the peak was deemed real and the gene discarded from analysis. For remaining genes, the read counts were normalized to the gene's mean read count. The distribution of normalized read counts per position (read counts from each gene normalized by the mean read count for the gene) across all genes above the density threshold was then computed. In order to compare the observed variability to counting noise, the distribution of normalized read counts for Poisson statistics with mean equal to the read

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

density threshold was also computed. Most genes with read density above a threshold will be near that threshold. Hence, the comparison to Poisson statistics at the threshold is a worst case scenario in terms of counting noise for a given density threshold.

Fig. S2M and N display the fraction of positions above (equivalent to one minus the cumulative distribution) different normalized read counts for both 5’ and 3’ ends of mapped reads and different density thresholds (in both *B. subtilis* and *E. coli*), together with the expectation coming from counting noise alone (dashed colored lines). The 5’ end distribution has a longer tail (more variability) than the 3’ end distribution, consistent with increased end bias arising from the 5’ end ligation (circLigase) compared to the 3’ end ligation in the cDNA library preparation (this can already be seen qualitatively in the Rend-seq traces shown throughout). More variability is also seen in *E. coli*. Interestingly, counting noise accounts for over half of the variability for the 3’-mapped reads signal even at high read densities.

The results from Fig. S2M and N suggest that the sensitivity of Rend-seq towards rare isoforms depends on the sequencing depth. The discussion below applies to a mean read density of 10 reads/nt. In order to reliably detect an internal 5’ or 3’ end amidst the noise in Rend-seq with a false detection rate of less than 5% per gene requires (assuming an average gene length of 1 kb) a false detection rate of 5×10^{-5} per nt is needed (horizontal dashed black lines in Fig. S2M and N). Looking at the distribution from Fig. S2M and N (specifically, the point where the distributions cross the dashed black lines), we thus need a read count of ≈6.2 (4.9 for *B. subtilis*, 7.4 for *E. coli*) times the mean read density for 5’ ends, and ≈4.2 (4.0 for *B. subtilis*, 4.3 for *E. coli*) times the mean read density for 3’ ends. Using that the end-enrichment is of about 58× (see Fig. S2C) with 25 s fragmentation, this means that rare overlapping transcripts of abundance ≈9.0% (5’) and ≈5.5 % (3’) of the main transcript can be reliably detected in Rend-seq. In addition to the read density (i.e., sequencing depth), the sensitivity clearly depends on the type of ends (3’ vs. 5’), but the above analysis suggests that rare nested isoforms present at an abundance of 5 to 10% of the major isoform should be detectable. An important assumption of the above is that the isoform’s 5’ and 3’ boundaries need to be defined to one nucleotide for the above sensitivity to hold. For instance, if a nested isoform has two equally important close-by 5’ ends, an abundance twice as large would be detectable (since the peak z score is computed one nucleotide at a time).

A corollary to the above discussion is that in spite of the end-enrichment, Rend-seq still cannot identify promoters for downstream genes inside the body of a very highly expressed upstream gene. One example is the *rpmE/tdk* region depicted in Fig. 4C of the main text. Since the RNA level for *rpmE* is over 50× that of *tdk*, it is difficult from Rend-seq alone to rule out a *tdk* transcription start site inside *rpmE*. Nevertheless, given the mechanism of intrinsic transcription termination, if the 5’ end occurs prior to the intrinsic terminator hairpin, termination should still occur. This thus leaves a short region of about 30 nt (typical size of an intrinsic terminator) where current Rend-seq datasets cannot detect minor 5’ ends inside highly expressed genes. Such 5’ ends would also affect our measurement of terminator read-through fraction. We can in fact identify from our data such promoters overlapping with, and upstream of, intrinsic terminators (e.g., in *B. subtilis* between *ywtF/*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ywtE, ydcI/ydcK, ydbI/ydbJ, gudB/ypdA, ykzT/patA, yrvD/recJ, ycbK/ycbL, ybfG/ybfF and *cydD/yxkO*, identified by searching for 5’ peaks in a 30 nt window upstream of 3’ peaks). Another example is a likely promoter just upstream of the terminator between *frr* and *uppS* in *B. subtilis* (black arrow in Fig. 4A). In these cases, the overlapping transcription units are detectable because they have similar abundances to the upstream transcription units (in contrast to the *rpmE/tdk* example). Though rare, these instances illustrate that such overlapping transcription units are possible.

To extend Rend-seq’s ability to identify TSS inside the body of upstream genes, one could size select longer RNA pieces post fragmentation. For example, if RNA fragments of sizes between 100 and 130 nt were selected (and mapped with paired-end sequencing), then there should be no 5’-mapped reads upstream of a 3’ end for 100 nt upstream of that 3’ end (the smallest sizes in the range selected) unless there is a promoter in that region. This lack of 5’-mapped reads from the abundant isoform of the upstream gene for 100 nt would then have the potential to reveal 5’-mapped reads not originating from the 3’ end (e.g., a promoter overlapping with the upstream transcript) in this region.

Another potential limitation is that the high end-enrichment from Rend-seq could amplify ends arising from, for example, semi-stable mRNA degradation intermediates, complicating the interpretation of the data. Without further treatment of the extracted RNA, the origin of ends in Rend-seq remains unspecified (sequence context can provide strong clues). Rend-seq data from *E. coli* does appear to be more “spiky" (see in particular Fig. S2N) than that from *B. subtilis* (although there is still a nice separate population of high peak z score positions as in Fig. S2A) and it remains unclear what these ends correspond to (degradation intermediates, weak/spurious promoters, etc.) and whether they arise as partial degradation products occurring in the RNA purification step. We have addressed this limitation (and to distinguish TSS from processed 5’ ends) by treating extracted RNA in some libraries with a 5’ monophosphate sensitive exonuclease, as has been previously done for genome-wide TSS mapping (Sharma et al., 2010; DiChiara et al., 2016).

Rend-seq: validation

Literature-based confirmation of 5’ ends—For both *B. subtilis* and *E. coli*, there is a large body of works using primer extension to identify the 5’ nucleotide of individual mRNAs. These hundreds of experimentally mapped sites provide independent validations for Rend-seq, which we detail below.

For 5’ ends that correspond to transcription start sites (TSSs), there are several curated databases established by literature mining. We started with the databases DBTBS for *B. subtilis* (Sierro et al., 2008) and Ecocyc for *E. coli* (Keseler et al., 2016). TSSs that have been verified using primer extension were selected for further analysis. Within this set, we identified 375 primer extension experiments that showed the exact location of TSSs as 5’ ends predicted in Rend-seq (114 in *B. subtilis* and 261 in *E. coli*). Another 204 TSSs have been mapped to be only ±1 nt away from the 5’ ends predicted by Rend-seq (69 in *B. subtilis* and 135 in *E. coli*), see Mendeley Data. The positions of these validated TSSs are listed in Mendeley Data. Therefore, Rend-seq results are consistent with at least 579 primer extension assays performed in different laboratories over several decades.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In addition to TSSs, 5’ ends can be produced by RNA processing. Although there are no curated databases for RNA processing to our knowledge, we validated the nucleotide-resolution of Rend-seq for these events on a gene-by-gene basis. As the first example, the 5’ UTR of *hbs* in B. subtilis is matured by RNase Y and RNase J1 (Daou-Chabo et al., 2009). The position of the mature 5’ end, as well as two TSSs, have been mapped, and it is reproduced by Rend-seq (Mendeley Data). We further confirmed that this mature 5’ end carries 5’ monophosphate, instead of 5’ triphosphate as expected for TSS. To do so, we pre-treated purified RNAs with 5’ monophosphate specific exonuclease. The resultant Rend-seq data shows the disappearance of the mature product (Mendeley Data).

Another well-established RNA processing site is in the sigma operon in *E. coli*. Rend-seq not only recapitulates the mRNA architecture of this entire operon, but also confirms the RNA processing site located immediately downstream of the stop codon of *dnaG* (Mendeley Data). This processing event creates a 5’ end for the downstream gene *rpoD*, whereas the upstream 3’ end is presumably rapidly degraded and thus not visible, consistent with previous reports. See Mendely Data for details and references.

Similarly, the atp operon in E. coli is processed by RNase E at multiple sites within and downstream of the atpB gene. These cleavage events also lead to downstream 5’ ends that are more stable than the upstream products. Rend-seq showed good agreement with the previously mapped sites (see Mendeley Data for details and references).

Finally, several mRNAs are processed by RNase III to produce mature 5’ ends. The transcript of *pnp* in *E. coli* is cleaved from its upstream *rpsO* gene. RNase III cleavage at two sites of a structured region between *rpsO* and pnp creates a 3’ end for the *rpsO* transcript and a 5’ end for the *pnp* transcript. The positions of these ends in Rend-seq agree perfectly with previously published results. Similar to *pnp*, the leader region of *sucA* is also processed by RNase III. The resulting 5’ end of the *sucA* mRNA is detected in Rend-seq at the same position. See Mendeley Data for details and references.

Validation of novel 5’ ends—To further validate novel 5’ ends that were identified by Rend-seq, we used 5’ rapid amplification of cDNA ends (5’ RACE) to PCR amplify specific mRNAs. First-strand cDNAs were synthesized using random hexamer primers, and tailed with either poly-dC or poly-dA at the 3’ end. Second-strand cDNAs were synthesized using either poly-dG or poly-dT primers with a 5’ adapter, and specific cDNA ends were PCR amplified using the 5’ adapter sequence paired with gene-specific primers. See section “5’ RACE" for additional details. Sanger sequencing for the *pyrH* gene-specific primer in *B. subtilis* confirmed a weak internal promoter identified upstream of *pyrH* (Fig. 4A, Mendeley Data). We attribute this 5’ end to a transcription start site because it is resistant to treatment by 5’ monophosphate specific exoribonuclease and has a putative -10 and -35 sequence (Mendeley Data).

For the *rpsP* operon in *B. subtilis* (Fig. 5A), 5’ RACE data also confirmed two 5’ ends upstream of *rpsP* (Mendeley Data). Rend-seq data with 5’ monophosphate exoribonuclease show that the first 5’ end is a product of RNA processing, whereas the second 5’ end is a TSS with a putative -10 and -35 sequence (Mendeley Data). These gene-specific validations

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

constitute a very small number in comparison to the hundreds of literature-based comparisons, but nevertheless demonstrate the predictive power of our approach.

Literature-based confirmation of 3’ ends—There exist two major types of evidence for RNA 3’ ends in the literature: those determined by S1 nuclease mapping and those determined by genome-wide methods based on high-throughput sequencing. Below we describe the agreement between Rend-seq and these existing results.

For 3’ ends determined by S1 nuclease mapping, we started with a database of putative *B. subtilis* transcription terminators curated by de Hoon et al (De Hoon et al., 2005). The majority of putative terminators in this database are proposed based on either purely sequence features or very low-resolution measurements such as Northern blotting, and hence they are unsuitable for validation purposes. Of the remaining 3’ ends with evidence from S1 nuclease mapping, we identified 13 cases whose experimental data together with the sequence were published in primary literatures. Rend-seq data confirmed 11 of the 13 cases, whereas the remaining two mRNAs were not expressed under our growth condition (Mendeley Data). Although the approximate positions of these results agree, the resolution of S1 mapping is limited due to the well-documented “nibbling" effect.

In addition to the database by de Hoon et al., we also identified several 3’ ends of mRNAs mapped with high resolution including the intrinsic terminators in the *rplL-rpoB intergenic* region in *E. coli*, the *rpsU-dnaG* intergenic region in *E. coli*, and downstream of *slrA* in *B. subtilis*. These results are again confirmed by Rend-seq (see Mendeley Data for details and references).

In addition to gene-specific 3’ end mapping, two groups have reported different high-throughput sequencing-based methods for genome-wide 3’ end mapping (Dar et al., 2016; Mondal et al., 2016). For example, Dar et al. developed Term-seq and identified 1,430 3’ ends in *B. subtilis*. We compared this list with the 1,803 3’ ends identified by Rend-seq (3’-mapped peak z score >12), which were obtained under the same growth condition as Term-seq.

In total, 1,202 3’ ends were identified in both methods (≤1 nt apart), i.e. 84% of Term-seq 3’ ends are identified by Rend-seq. Term-seq and Rend-seq agreed on the exact end position in 86% of these cases (Mendeley Data). Of the remaining 601 3’ ends called in Rend-seq but not in Term-seq, 83% has their locations outside the window used in the Term-seq analysis pipeline (i.e., less than 150 nt downstream of the closest upstream gene, and no more than 10 nt inside downstream gene). These 3’ ends after long 3’ UTRs or inside coding regions likely represent genuine RNA isoforms, given the additional signature of step-wise decrease in Rend-seq read density that accompanies each peak. There remain 96 ends called in Rend-seq (5% of 1,803) but not in Term-seq, even though their locations satisfy the Term-seq analysis pipeline. Comparison of peak height to step sizes surrounding the peaks not listed as Term-seq 3’ ends suggests that these are valid ends of transcription units (Mendeley Data).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Conversely, there are 191 ends called in Term-seq but are missed in our analysis. Over 30% of these have moderately high z scores in Rend-seq (above 7 and below 12, which is the cutoff for automated peak-calling; see Figure S3A for overall z score distribution). Most of these Rend-seq peaks have other peaks nearby, which may reduce the z scores to be below the cutoff. Hence, the 3' ends called by Term-seq provide us with an estimate of the false negative rate of our automated analysis (with a modified peak z score threshold of 12), at about 4%. Of the remaining 129 3' ends missed in Rend-seq, 83 have low surrounding read density (<0.25 read/nt) in our dataset (and are thus missed by our automated analysis), presumably coming from lowly expressed genes.

Similar comparison for the common 3' ends from Rend-seq and those reported by (Mondal et al., 2016) also reveals good agreement at the single nucleotide level (92% of ends within 1 nt, Mendeley Data). Most of the 209 ends identified in Rend-seq but absent from (Mondal et al., 2016) also show the linear relationship between peak height and step size (Mendeley Data).

Validation of novel 3' ends—To further validate novel 3' ends that are identified by Rend-seq, we first used Northern blotting to confirm the approximate relative sizes of mRNA isoforms was same as that predicted by Rend-seq. We then used site-directed mutagenesis towards terminator sequences near the 3' ends and confirmed that such mutations lead to 3' extensions.

We targeted novel 3' ends whose upstream sequences resemble the canonical intrinsic transcription terminator in *B. subtilis*. These include 3' ends downstream of *ylqC* (Fig. 5A), *rpsB* (Fig. 4A), *tsf* (Fig. 4A), and *rpmA* (Fig. S3I–J). For each terminator, we removed the sequence containing both the U-tract only (*Δ*U) and the U-tract plus the second half of the hairpin stem (*Δ*sU). The deletions were introduced at the native loci on the *B. subtilis* chromosome. If the mRNA species seen on the Northern blot was indeed associated with the 3' end generated by transcription termination, disruption of terminator activity would lead to both the disappearance of this species, as well as increased abundance of longer RNA isoforms that corresponds to termination downstream.

For every case tested, Northern blotting confirmed the expected loss of shorter mRNA isoforms upon the disruption of their 3' terminator sequence (Fig. S3). Furthermore, the expression of longer isoforms increased, which again is consistent with the notion that these novel 3' ends are generated by partial transcription termination. For example, Rend-seq data suggest that *rpsB* in *B. subtilis* is expressed in at least three different isoforms with the same 5' ends and different 3' ends that correspond to intrinsic terminator sequences downstream of *rpsB, tsf*, and *frr*. Northern blotting against the *rpsB* gene in wildtype cells confirmed the sizes of these three major isoforms (Fig. S3B). Upon removal of the terminator downstream of *rpsB*, the shorter isoform disappears, whereas the longer isoforms become more abundant. The latter is also confirmed by Northern blotting against the *tsf* gene and the *pyrH* gene.

Northern blotting results for the four novel 3' ends tested are shown in Fig. S3B, F and J.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Validation of novel isoforms—The overlapping RNA isoforms predicted by Rend-seq are consistent with many previously reported cases, such as the sigma operon (*E. coli*), the *atp* operon (*E. coli*), the *rpsO-pnp* operon (*E. coli*), the *sucA* operon (*E. coli*), the *rplL* operon (*E. coli*), the *rpsU* operon (*E. coli*), the *hbs* gene (*B. subtilis*), and the *slrA* operon (*B. subtilis*) (see discussions in “Literature-based confirmation of 5’ ends" and “Literature-based confirmation of 3’ ends"). In addition to these known examples, we further confirmed the novel isoforms that are identified by Rend-seq using Northern blotting.

For the *ylxM-ffh-rpsP-ylqC-ylqD-rimM-trmD-rplS* gene cluster in *B. subtilis* (Fig. 5A), Rend-seq predicts seven different RNA isoforms (Fig. S3 E). Two isoforms share the same 5’ end located upstream of *ylxM* but have different 3’ ends, either downstream of *ylqC* or downstream of *rplS*. Two 5’ ends upstream of *rpsP*, in combination with the two 3’ ends (*ylqC* and *rplS*), give rise to four other isoforms. Finally, there is an abundant *rplS*-only isoform with its own 5’ end. To validate these isoforms, we used ^{32}P-labeled ssDNA probes against *ffh, ylqC, trmD*, and *rplS*. The *ffh* probe confirmed the first two isoforms (*ylxM-ffh-rpsP-ylqC* and *ylxM-ffh-rpsP-ylqC-ylqD-rimM-trmD-rplS*). The *ylqC* probe showed four major bands consistent with *rpsP-ylqC, ylxM-ffh-rpsP-ylqC, rpsP-ylqC-ylqD-rimM-trmD-rplS*, and *ylxM-ffh-rpsP-ylqC-ylqD-rimM-trmD-rplS*. The pairs of isoforms that correspond to two 5’ ends upstream of *rpsP* cannot be resolved on Northern blots (independently confirmed by 5’ RACE, see section “Validation of novel 5’ ends" and Fig. S3E–F), presumably because their sizes are too close. The *trmD* probe showed two major bands, corresponding to the full 8-gene transcript and the transcripts starting upstream of *rpsP*. Finally, the *rplS* probe confirmed the four major bands predicted by Rend-seq.

Similarly for the *rpsB-tsf-pyrH-frr* gene cluster in *B. subtilis*, Northern blotting against *rpsB* confirmed that it is present in three isoforms (Fig. S3A, B). A probe against *tsf* confirmed that is it present in two major isoforms, with another minor short isoform that is not predicted by Rend-seq. This band might be an artifact due to off-target interaction of the *tsf* probe in Northern blot, generated by a weak promoter inside the *rpsB* gene that is not detected by Rend-seq, or could represent a degradation intermediate with boundaries not defined at single resolution (making it harder to identify from Rend-seq). Finally, a probe against *pyrH* confirmed that it is present in the two major isoforms predicted by Rend-seq.

For the *rplU-ysxB-rpmA-spo0B-obgE-pheB-pheA* gene cluster in *B. subtilis*, Northern blotting against *rpmA* and *ysxB* confirmed four main isoforms predicted by Rend-seq (Fig. S3 I,J). Additional bands on the Northern were presumably due to other low abundance isoforms (with 5’ ends upstream of spo0B) detected in Rend-seq, but unambiguous band assignment was not possible. A probe against *spo0B* and *pheA* confirmed the two main isoforms due 3’ extensions (respectively partial termination and likely 3’ to 5’ chewback event, as the 3’ end between *obgE* and *pheB* disappears in our *pnpA* knockout data).

Literature-based confirmation of abundance—Rend-seq not only provides information about the positions of RNA 5’ and 3’ ends, it also estimates RNA abundance based on read density internal to the body of transcripts. To provide independent validation for its ability to quantify transcript abundance, we compared the gene-level read density to microarray-based expression profiling (Nicolas et al., 2012) (shown in Mendeley Data is the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

median of the three LB exponential replicates in (Nicolas et al., 2012), results are unaffected if compared to each replicate separately) for *B. subtilis* under the same growth condition (LB exponential growth, with the minor caveat that cells were harvested at OD= 0.5, in contrast to OD=0.3 in the current work). Overall, there is a strong correlation (R^2=0.80 for the linear fit of log-transformed abundances) between the two data sets (Mendeley Data). Within individual gene clusters, differential expression among neighboring genes is also consistent between the two data sets (see Mendeley Data for the comparisons for the gene clusters highlighted in Fig. 4 and Fig. 5).

Validation of isoform abundance—To further confirm the observed variations in isoform abundance, we quantified the Northern blot results for the *ylxM-ffh-rpsP-ylqC-ylqD-rimM-trmD-rplS* and *rpsB-tsf-pyrH-frr* gene clusters in *B. subtilis* (quantitation for the *rpmA* operon was not possible as two probes to *rpmA* and *ysxB* were simultaneously used for one of hybridization). These clusters contain tuned transcription terminators that generate isoforms with and without 3' extensions. Therefore, quantitation using Northern blots also provides independent measurements for the read-through propensity estimated by Rend-seq.

The quantitation based on band intensity is summarized in Fig. S3B, F and J. Isoform abundance from Rend-seq is estimated as described in section "Estimating abundance of overlapping mRNA isoforms", with the exception of the *pyrH-frr* isoform. Estimating the abundance of this isoform based on the read density in the short region (36 nt, leaving the usual gap way from the peaks) between the *tsf* 3' end and the *pyrH* 5' end lead to a lower value compared to the Northern blot. Substantial variability in Rend-seq signal read counts over such short range is not unlikely, presumably due to cloning biases (c.f., short range variability in Rend-seq signal in traces shown in main figures). We thus relied on the height of the 5' peak (abundance given by peak height divided by median end-enrichment, see section "Quantification of end-enrichment") to quantify the abundance of the *pyrH-frr* isoform. Increasing the size of selected fragments in Rend-seq would be a simple way to alleviate the variability coming from averaging over short windows (see discussion in section "Detection sensitivity").

Overall, there is a good correspondence between results from Rend-seq and Northern blotting. For example, in the *rpsB-tsf-pyrH-frr* gene cluster, Rend-seq predicts two consecutive transcription terminators with 54% and 12% read-through after *rpsB* and *tsf*, respectively. As a result, the *rpsB* gene is expected to be present in three major isoforms with relative abundance of 100:101:12 from the shortest to the longest (*rpsB:rpsB-tsf:rpsB-tsf-pyrH-frr*). This ratio is confirmed by Northern blotting, which measured 100:93:10. Northern blotting against *tsf* and *pyrH* in this gene cluster also provide independent confirmation of the relative abundance (Fig. S3B).

Northern blotting against different regions of the gene clusters also provides many independent measurements for the read-through propensity estimated by Rend-seq. For example, the band intensity of various isoforms of *rpsB* suggests 51% and 10% read-through at the *rpsB* and *tsf* terminators, respectively. The band intensity of the two major isoforms of *tsf* similarly suggests 14% read-through at the *tsf* terminator (Fig. S3C and D).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For the tuned terminator downstream of *ylqC* in the *ylxM-ffh-rpsP-ylqC-ylqD-rimM-trmD-rplS* gene cluster, its read-through propensity (4.9% by Rend-seq) is confirmed by Northern blot probes against *ffh* (6.3%) and *ylqC* (5.7% for transcripts starting before *ylxM* and 4.0% for transcripts starting before *rpsP*), see inset in Fig. S3E.

Characterization of internal 3’ ends

Global analysis of internal 3’ ends—We first sought a global quantification of 3’ ends in *B. subtilis* (we did not carry this analysis in other species due to the increased prevalence of rho-dependent termination, which does not always lead to a well-defined 3’ end and thus complicates the analysis), and in particular, characterization of the fraction of operons with internal 3’ end. To do so, peaks in 3’-mapped reads were identified (1776 total) as our starting set of 3’ ends. The read density upstream and downstream of these 3’ ends was then obtained by averaging over a 50 nt window (restricted to a shorter region if additional ends are identified in the 50 nt upstream and downstream windows). We operationally defined internal 3’ ends as those with more than 5% relative (to upstream) downstream read density. The remaining set of 3’ ends (termed “primary 3’ ends") were considered to mark boundaries of contiguous operons. These criteria identify 708 primary 3’ ends and 1068 internal 3’ ends. To determine the number of internal 3’ ends per operon, we counted the number of internal 3’ ends between primary 3’ ends (in a strand-specific manner). The mean number of internal 3’ ends per operon using this criterion was equal to 1.5 (median of 1), suggesting that 3’ extensions were prevalent.

Assignment of intrinsic terminators—To further characterize the nature of identified 3’ ends (and internal 3’ ends), with emphasis towards intrinsic termination, we sought more stringent identification and categorization criteria.

To robustly identify 3’ ends of transcription units for the purpose of intrinsic terminator identification, we thus thresholded on two statistics (in place of only the peak z score for other analyses in the current work). In particular, in addition to thresholding on the peak z score (P below) of the 3’-mapped Rend-seq signal, we also thresholded on the step z score (S below). The additional step z score statistic was included to decrease the number of false negatives. The step z score is defined as the difference between upstream and downstream read densities (averaged over 100 nt excluding the central 3 nt) divided by the standard deviation of read counts ($\sigma \coloneqq \sqrt{\sigma_u^2 + \sigma_d^2}$, where σ_u and σ_d are the standard deviation of read counts for the upstream and downstream regions, respectively). We note that S should be positive across steps around 3’ ends. We used an empirical threshold in the plane of these two statistics as a first cut through the data (for $S < 0.2$, require $P \geq 15$, for $0.2 \leq S \leq 1$, require $P \geq -3.125 \times (S - 0.2) + 10$, for $S > 1$, require $P \geq 7.5$). Neighbouring positions (closer than 2 nt) passing the threshold were grouped and only the position in the neighbourood with the largest peak z score was kept. 1856 and 1544 positions passed this first selection in *B. subtilis* and *E. coli* respectively.

We then searched for potential intrinsic terminators among the identified 3’ ends above by looking for the canonical hairpin and U-tract features upstream of the identified 3’ ends (previous paragraph). To do so, the RNA sequences immediately upstream of the 3’ ends (in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a 55 nt window) were folded using RNAfold (Lorenz et al., 2011). In order to break the ambiguity in the possible RNA secondary structures upstream of the 3' ends (multiple alternative hairpin structures with similar folding energies are often possible in a 55 nt window), sequences progressively longer extending from the 3' end (extended by 1 nt from the 3' end) were generated and folded while constraining the first 6 nt upstream of the 3' end to remain unfolded. The minimum free energy structure was stored for each sub-sequence. If a single hairpin structure of 5 base pairs or more was not replaced by a more stable structure as the folded sequence was extended further by 9 nt, the structure was stored as the hairpin for the 3' end. Otherwise, the structure of the full 55 nt region was selected. The hairpin parameters were then extracted: $\Delta G_{hairpin}$) (RNA free energy of folding), ℓ_{loop} (number of bases in the hairpin loop), N_{bp} (number of bases paired in the hairpin stem, which does not include the pairing between a possible upstream A-tract and the U-tract, because of the folding constraint above) and L_U (length of the U-tract length, defined as the maximal number of consecutive U's in the 8 nt upstream of the identified 3' end). Threshold were then set on these various characteristics: $\Delta G_{hairpin} < -7$ kcal/mol (moderately strong hairpins), $3 \leq \ell_{loop} \leq 10$ nt (no excessive loops), $4 \leq N_{bp} \leq 17$ (no excessively short or long hairpins) and $L_U \geq 2$ (presence of a minimal U-tract). 85% (1583/1856) of the identified ends passed these combined cuts in *B. subtilis* and 58% of the ends (889/1544) passed these cuts in *E. coli* (probably reflecting the increased prevalence of Rho-dependent termination in *E. coli*). Sequences randomly chosen from each genome passed the hairpin and U-tract cuts above at about 7%. Given that we were stringently selecting for peaks and steps via appropriate statistics from our data, as well as rejecting ends of transcripts arising from different molecular mechanisms (see below), our false positive rate was likely substantially lower.

Similar analyses in *V. natriegens* and *C. crescentus* (no exclusion of possible decay products or rho-dependent termination for these species, see below) respectively yielded 1257 (87% of 3' ends passing the terminator cuts above) and 374 (42% of 3' ends passing the terminator cuts, with a requirement of $\Delta G_{hairpin} < -15$ kcal/mol in place of -7 kcal/mol to avoid too many false positives given *C. crescentus*' high GC content) high confidence terminators. Identification of putative terminators was also performed in the various mutants for the purpose of read-through fraction characterization (see section below).

Finally, for *E. coli* and *B. subtilis*, we further discarded the 3' ends that likely result from processing from other downstream 3' ends or rho-dependent termination, but nevertheless have features that resembles intrinsic terminators. These transcripts may be terminated by Rho or other factors, and therefore lack stem-loops at the termination site to prevent processive degradation by 3'-to-5' exonucleases (the major exonuclease being PnpA in *B. subtilis*) until reaching stable hairpins. Furthermore, comparing the positions of 3' ends in both wild-type and all our mutants

To do so for *B. subtilis*, we carried out Rend-seq for *Δrho* and *ΔpnpA* strains and discarded 3' peaks that were absent (maximum peak 3' z score in a ±5 nt neighborhood less than 7) in either of these mutants. Considering regions with sufficient read depth for characterization, we found that, with the above criterion, 98% and 96% of 3' ends were retained in *Δrho* and *ΔpnpA* strains respectively. We did not exclude 3' ends for which the depth in the mutants

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

was insufficient for characterization. After removal of 3’ ends lacking in the mutants, we were left with 1486 terminators.

A similar analysis was performed in *E. coli* with data from *Δpnp* and *Δrnb* Rend-seq data (two 3’ to 5’ exonuclease known to act on mRNAs in *E. coli*, (Donovan and Kushner, 1986)). There, we found that 90% and 80% of 3’ ends were retained in *Δrnb* and *Δpnp* strains respectively. After removal of 3’ ends not present in the mutants, we were left with 722 putative terminators. To further remove possible Rho-dependent terminators in *E. coli*, remaining putative intrinsic terminators overlapping with BST ("bicyclomycin significant transcript") from (Peters et al., 2012) were removed, leaving final list of 630 high confidence intrinsic terminators.

The subset of these terminators for which read-through fraction could be quantified (see below) can be found in Table S3.

Validation of intrinsic terminators—To test whether these terminator-like 3’ ends are indeed generated via intrinsic transcription termination, we created mutant strains of *B. subtilis* for which parts of the canonical terminator features are removed. We tested the 3’ ends that we refer to as tuned terminators, as highlighted in Fig. 4 (two 3’ ends in *rpsB-tsf-pyrH-frr* cluster), Fig. 5 (*ylxM-ffh-rpsP-ylqC-ylqD-rimM-trmD-rplS*), and another case (*rplU-ysxB-rpmA-spo0B-obgE-pheB-pheA*). For each of the four terminator sequences, we created two mutant strains: one with the U-tract deleted (*Δ*U) and another with both the U-tract and the second half of the hairpin stem deleted (*Δ*sU). The deletions were introduced at the native loci on the *B. subtilis* chromosome.

Both *Δ*U and *Δ*sU are expected to disrupt termination and therefore increase the expression of downstream genes. A major difference between these mutants, however, is that the *Δ*U strain preserves the RNA hairpin structure that is known to stabilize the upstream portion of transcripts by inhibiting 3’-to-5’ exonuclease activity, independent of the termination activity. We therefore expect that there might still exist a shorter isoform in the *Δ*U strain, albeit with lower abundance, if some transcripts are degraded from the region downstream of the tuned terminator.

Northern blotting, shown in Fig. S3B, F and J, confirmed these expected results of terminator mutations. For the tuned terminator downstream of *rpsB*, both *Δ*sU and *Δ*U increase the abundance of longer isoforms, as measured by probes against either *rpsB, tsf*, or *pyrH*. The shorter, terminated isoform disappears. For the tuned terminator downstream of *tsf*, both *Δ*sU and *Δ*U increase the abundance of the longer isoform, as measured by probes against either *tsf* or *pyrH*. The shorter isoform consisting of *rpsB-tsf* disappears in *Δ*sU and remains present at a lower level (2× decreased) in the *Δ*U strain. This potential decay product, if also present in the wild type, might contribute to our estimate of read-through fraction of this particular terminator. However, the abundance of this product in the *Δ*U strain suggests that its contribution is minor.

For the tuned terminator located downstream of *ylqC* in the *ylxM-ffh-rpsP-ylqC-ylqD-rimM-trmD-rplS* gene cluster, both *Δ*sU and *Δ*U dramatically increase the abundance of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

longer isoforms, as measured by probes against either *ffh, ylqC, trmD*, or *rplS*. Concurrently, the shorter terminated isoforms disappear.

For the tuned terminator located downstream of *rpmA* in the *rplU-ysxB-rpmA-spo0B-obgE-pheB-pheA* gene cluster, both *Δ*sU and *Δ*U dramatically increase the abundance of longer isoforms, as measured by probes against either *ysxB+rpmA, spo0B*, or *pheA*. These mutations lead to disappearance of the shorter isoforms.

Contribution of tuned terminators—From our list of intrinsic terminators (see above), we further identified intergenic terminators that are responsible for setting the transcription level of downstream genes.

We first discarded the terminators that lead to nearly complete termination. For all the intrinsic terminators identified in the previous analysis (listed in Table S3), the positions of the stop codon of the immediate upstream gene and of the start codon of the first downstream, co-directional, gene were determined. If the average read density dropped below 0.15 read/nt anywhere over the intergenic range (moving average over 40 nt windows), the terminator was regarded as complete and hence excluded for this analysis. Most terminators that are followed by genes in the opposite direction were thereby excluded using this criterion.

For the remaining terminators, we calculated the contribution of the read-through activity to the expression of downstream genes. Because some of these terminators are followed by one or more promoters, we quantified the level of transcription from read-through using the Rend-seq read density between the terminator and either the first downstream promoter or the downstream start codon if no promoters were detected. We then calculated the “fraction of mRNA level from read-through” for the downstream gene by taking the ratio of the read-through isoform and the mean Rend-seq read density across the body of the downstream gene. The list of terminators was further manually curated to remove false positives, known riboswitches, and other putative leader sequences that do not encompass the entirety of the upstream gene. In the rare case of double intergenic terminators, the combined read-through fraction was considered. The fraction was set to 1 in cases where it was slightly above 1 due to noise in our estimates of read densities (especially given the small sizes of some intergenic regions).

The resulting distribution of fractions from read-through is plotted in Fig. 4E for terminators in *B. subtilis*. 167 terminators were responsible for the majority (>50%) of transcription of the downstream, which we termed “tuned” terminators (Table S3). As compared to the number of riboswitches or other 5’ cis-regulators (82 in *B. subtilis* (Dar et al., 2016)), the above quantitation illustrates that partial terminator read-through plays an important role in tuning gene expression in *B. subtilis*. The number of genes whose expression is set by these terminators (many genes can be downstream of a terminator on a poly-cistronic transcript) was determined manually, by going through the list of terminators and counting the number of genes downstream of each terminator (without an intervening 5’ end). We obtained 276 genes in *B. subtilis* for which over 50% of their mRNA level arises from read-through at tuned terminators.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The same analysis for *E. coli, V. natriegens* and *C. crescentus* yielded respectively 88, 140 and 47 tuned terminators (Table S3), suggesting that tuned termination is a common mechanism to differentiate the expression of operonic genes in bacteria.

We again emphasize that we could not rule of exonucleolytic chewback events or rho-dependent terminators for *V. natriegens* and *C. crescentus*. In addition, as detailed in section “Detection sensitivity", current Rend-seq datasets (with size selection range 15 to 45 nt) cannot always rule out the presence of promoters upstream of terminators in the body of the upstream transcript, especially for high differential abundance of upstream and downstream regions. As such, some of the reported tuned terminators could be a combination of very strong terminator and missed promoter in the upstream transcript. Given the over 50-fold end-enrichment for Rend-seq, we consider this to be an unlikely possibility for tuned terminators with read-through fraction exceeding 10%.

Quantification of read-through fraction—The fraction of RNA polymerases continuing through an intrinsic terminator (hereafter read-through fraction) identified in the previous section can be estimated from Rend-seq data (see (Cambray et al., 2013; Chen et al., 2013) different approaches for the characterization of read-through at intrinsic terminators) by comparing the read density upstream and downstream of the terminator:

$$\text{Read-through fraction} := \frac{\text{Read density downstream of terminator}}{\text{Read density upstream of terminator}}.$$

This definition assumes that the short and long transcripts, i.e. terminated and read-through mRNAs respectively, are degraded at the same rate. Because the canonical mRNA decay starts near the 5’ end, we indeed expect the same half-life for mRNA isoforms that share identical 5’ sequence but have different 3’ extension. We further controlled for the possibility of 3’ specific influence of mRNA stability by comparing the wild type with mutants of 3’-to-5’ exoribonucleases (see below).

To calculate read-through fraction, we quantify read density over the 200 nt regions upstream and downstream of each terminator (leaving a ±6 nt gap around the peak). If a nearby promoter, terminator, or processing site was found in these regions using Rend-seq, the averaging window was restricted to include only segments between the terminator and the additional end. If the averaging window was too small for accurate quantification (<20 nt) or zero reads were identified in averaging region (impossible to quantify read-through), the terminator was discarded for further analysis. In rare occasions, estimated read-through fractions were over 1, which was likely due to a cryptic promoter that is missed in our automated analysis. We discarded these instances. In all, we could quantify read-through for 1414, 599, 1154 and 338 terminators in wild-type *B. subtilis, E. coli, V. natriegens* and *C. crescentus* respectively. For each terminator identified in wild-type, the same upstream and downstream averaging range was used in *B. subtilis* and *E. coli* mutants to quantify read-through (see below). The final list of intrinsic terminators for the four species considered can be found in Table S3, with measured read-through in wild-types and in the various mutants (depth permitting), as well as and terminator properties (section below).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To assess the validity of using steady-state mRNA level to estimate terminator read-through, we compared the read-through fractions from wild-type to mutants of 3'-to-5' exoribonucleases (main one PnpA in *B. subtilis* (Oussenko et al., 2005), with other 3' to 5' exonuclease involved in stable RNA processing also profiled: YhaM, Rnr, and Rph in *B. subtilis*; Pnp and RNase II are the two major ones in *E. coli* (Donovan and Kushner, 1986)), as well as read-through in Rho mutant *B. subtilis*. The comparisons are shown in Fig. S3K, L for terminators with sufficient read depth in the mutants. The agreement between the wild type and mutants support our assumption that the short and long isoforms have similar half-lives. For the analysis on the determinant of read-through propensity in *B. subtilis*, we used the measurement made for the mutant of PnpA, which is the major 3'-to-5' exoribonuclease (Oussenko et al., 2005), to avoid slight differences in isoform stability. In order to maintain a high number of terminators for analysis of determinants of read-through, we used datasets from wild-type terminators in other species (next section).

Contribution of differential mRNA stability—To further evaluate the contribution of differential isoform stability to apparent read-through fraction, we compared their respective decay rates using data from (Moffitt et al., 2016) in *E. coli*.

Briefly, following (Moffitt et al., 2016), we normalized read density across ORF in the different rifampicin cut-off time points to the read density mapping to the tmRNA (a stable RNA). For all tuned terminators identified in *E. coli* (see above section), the normalized RNA level post rifampicin treatment (denoted by $m(t)$) for genes upstream and downstream (termed gene 1 and 2 respectively for the current analysis) of the tuned terminator was fit by non-linear least square to a decaying exponential with delay ($m(t) = N_i + N_f$ for $t \leq \Delta t$ and $m(t) = N_i + N_f e^{-(t-\Delta t)/\tau}$ for $t > \Delta t$, where fit parameters are N_i, N_f, Δt and τ). If the best-fit lifetime τ resulting from this procedure was shorter than 0.2 min, $m(t)$ was fit to a simpler exponential decay $m(t) = N_i + N_f e^{-t/\tau}$). This procedure was performed for data from two biological replicates from (Moffitt et al., 2016). Genes with manifest non-monotonic decay of mRNA level post rifampicin treatment, widely different time dependence across replicates, or with no reads mapped to the genes of interest, were discarded. Overall, we could quantify the lifetimes τ_1 and τ_2 for genes upstream and downstream of tuned terminators in 83 cases.

Obtaining the steady-state abundances m_1 and m_2 for these transcripts from Rend-seq (1% winsorized read density across the ORF, leaving 45 nt gap at the start and stop codons for the averaging window) then allowed us to determine whether differential lifetimes were dominant in establishing differential transcript levels. The steady-state mRNA level for a transcript is equal to its production rate k multiplied by its lifetime τ, i.e., $m = k\tau$. In our current analysis tuned termination pertains to a different production rate k, where differential stability corresponds to differences in lifetime τ. We compared the ratios m_1/m_2 and τ_1/τ_2 for genes upstream and downstream of tuned terminators in *E. coli*, Fig. S3S. Equal contribution from differential production and differential stability towards establishing differential abundance corresponds to $\tau_1/\tau_2 = k_1/k_2 = \sqrt{m_1/m_2}$. Therefore, gene pairs straddling tuned terminators for which $\tau_1/\tau_2 < \sqrt{m_1/m_2}$ correspond to examples where differential stability plays a minor role in establishing differential level (red points in Fig.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

S3S). We observe that nearly 90% genes straddling terminators have their abundance set by differential production, and not differential mRNA stability ($\tau_1/\tau_2 < \sqrt{m_1/m_2}$).

Determinants of read-through fraction—Our list of high confidence intrinsic terminators with corresponding read-through fraction measurements for multiple species (and mutants in *B. subtilis* and *E. coli*) provides an opportunity to identify sequence/structure characteristics of terminators influencing their read-through properties (analogously to recent works (Cambray et al., 2013; Chen et al., 2013), but with an orthogonal way to measure read-through and in the endogenous chromosomal context of the terminators).

Various structural and sequence features of the terminators were investigated for correlations to read-through fraction. We determined the following feature for each terminator: U-tract length L_U (number of consecutive U's in the 8 nt upstream of the identified ends, see below), stability of the U-tract's RNA/DNA hybrid ΔG_U (for the 8 nt immediately upstream of the 3' end, computed with a nearest-neighbor model (Sugimoto et al., 1995)), number of U's in the U-tract n_U (total number of U's in the 8 nt upstream of the 3' end), hairpin free energy (see section "Assignment of intrinsic terminators" for details), hairpin free energy over stem length, loop size, stem length, the fraction of bases paired in the stem, the strength of competing upstream RNA secondary structure (defined as the minimum free energy of folding of the 50 nt upstream of the center of the loop of the terminator) and the strength of the upstream A-tract (defined as the difference in free energy of folding for the 50 nt upstream of the identified 3' end with and without constraining the U-tract to be unpaired). The distribution of a subset of these quantities are shown in Fig. S3M–O.

More specifically, with respect to the definition of the U-tract length, Rend-seq data in mutants allowed us to assess whether the position of 3' ends identified in wild-type depended on 3' to 5' exonucleolytic activity. To do so, for each 3' end identified in WT, the position of the corresponding 3' end (depth permitting) in mutants was identified. Fig. S3P provides the distance distribution (where $\Delta x < 0$ corresponds to the end being downstream in the mutant) between wild-type and our *B. subtilis* mutants. We did not observe systematic shifts in *Δrho* or *Δrnr* mutants but did see that about 15% of ends where 1 nt downstream of the wild-type 3' ends in *ΔpnpA, Δrph* and *ΔyhaM* mutants. For the purpose of the analysis here, we recalculated the U-tract length for terminators identified in wild-type based on the position of the observed 3' end in *ΔpnpA* (depth permitting). A similar analysis in *E. coli* revealed a shift of about 15% of ends in the *Δpnp* strain (Fig. S3Q), consistent with the *B. subtilis ΔpnpA* data. In contrast, we observed prevalent short 3' extensions (median extension of 2 nt) in the *Δrnb* dataset, suggesting that the exact location of most 3' ends *in vivo* are nibbled by RNase II. For the purpose of U-tract length for our terminators in *E. coli*, we thus used the 3' extended ends identified in *Δrnb* (depth permitting).

Linear regressions on the logarithm of the read-through fractions versus the above listed features were performed, the results are tabulated in Fig. S3R. Across all species, the U-tract quality (all three features L_U, n_U and ΔG_U) emerged as the most highly correlated variable for the read-through fraction ($R^2 \geq 0.15$), consistently with previous reports based on measurements with fluorescence reporters (Cambray et al., 2013; Chen et al., 2013). We

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

note that the correlation between read-through and L_U was substantially weaker in wild-type *B. subtilis* compared to measurements in *ΔpnpA* (R^2 = 0.11 vs. R^2 = 0.18). Hairpin stability over the stem length was also modestly correlated across species ($0.03 \leq R^2 \leq 0.09$). Beyond factors possibly confounding our read-through measurements (missed promoters inside the body of upstream transcripts), the search for additional biophysical determinants of read-through by focusing only on the terminator properties is further complicated by the fact that additional factors, e.g., NusA (Mondal et al., 2016), have been shown influence read-through at intrinsic transcription terminators.

Characterization of internal 5’ ends

Global analysis of internal 5’ ends—Similar to the analysis for internal 3’ ends, we sought to globally determine the fraction of operons with internal 5’ ends. Under our growth conditions, we identified 1,984 5’ ends in *B. subtilis* and 2,288 in *E. coli* (using peak z score > 12). We quantified the number of 5’ ends that are internal to other transcription units by looking for those that have substantial RNA levels immediately upstream (50-nt window, restricting the averaging window if additional peaks are found in proximity). We operationally defined “internal 5’ ends" as those whose upstream RNA level is > 5% of the downstream level. Internal 5’ ends account for 54% of total 5’ ends in *B. subtilis* and 69% in *E. coli*. The remaining 5’ ends are referred to as “primary 5’ ends." Primary 5’ ends can be viewed as the first 5’ end of contiguous operons, and the number of primary 5’ ends marks the number of contiguous operons that are expressed under our growth conditions. By counting the number of internal 5’ ends between consecutive primary 5’ ends on the same strand, we estimate that 49% of contiguous operons have internal 5’ ends in *B. subtilis* and 65% in *E. coli*. The mean number of internal ends per operon is 1.2 and 2.2 in *B. subtilis* and *E. coli* respectively. Note that these 5’ ends consist of both TSSs and RNA processing sites of mature mRNAs.

Assignment of processed 5’ ends—To further characterize the nature of the identified internal 5’ ends (see section above), we performed Rend-seq with prior treatment of the RNA with 5’ monophosphate sensitive exonuclease (Epicentre), which should degrade processed 5’ ends (the first nucleotide of a transcribed mRNA (TSS) should have 5’ triphosphate and therefore be resistant to 5’ monophosphate sensitive exonuclease treatment).

We considered an internal 5’ end to be the result of a processing event if the maximum 5’-mapped peak z score (depth permitting) for the 5’-exo treated sample in the ±2 nt neighborhood of the position identified in the 5’-exo untreated sample was below 7 (fraction largely independent of the chosen peak z score threshold provided it is above 5). Using this criterion, we estimate that 17% and 45% of internal 5’ ends are the result of processing or arise from degradation intermediates in *B. subtilis* and *E. coli* respectively.

Sequence context of putative TSSs—Rend-seq with 5’-exo treated RNA also provides an additional way to validate the high resolution of our 5’ end mapping. 5’ peaks in the 5’-exo treated datasets should correspond to transcription start sites. In rich medium, the majority of transcription start site are directed by the house keeping sigma factor which has

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the characteristic -10 and -35 sequence motifs of TATAAT and TTGACA respectively. We could identify 976 and 815 high confidence putative TSS from our Rend-seq data in *B. subtilis* and *E. coli* respectively (5'-mapped peak z score > 12 in both 5'-exo treated and untreated Rend-seq datasets, not shown). The upstream sequence context of these putative TSSs shows strong enrichment (e.g., > 1.3 bits for 3 position in the -10 region in *B. subtilis*) for the expected housekeeping sigma factor motifs with an alignment tolerance of ±1 nt, further supporting the high resolution of Rend-seq's 5' end mapping.

Estimating abundances of overlapping isoforms

Assumptions—For overlapping transcription units (e.g., Fig. 4, 5), we can use Rend-seq data to provide estimates of the relative abundances of the different mRNA isoforms. Like most RNA-seq approaches, Rend-seq strictly speaking contains only local information. The peaks in the data provides information about the ends of possible transcription units. However, Rend-seq data alone are insufficient to determine the relative abundance of each isoform in cases where there are multiple 5' and 3' ends. In particular, Rend-seq data cannot determine that all the possible isoforms are indeed present in complex cases (two or more 5' and two or more 3' ends). The reason is that *a priori*, the processes generating the ends of transcription units could be coupled. For example, the read-through at an intrinsic terminator could depend on features upstream of the terminator hairpin and thus on the specific 5' end in operons for which multiple upstream transcription start sites exist. Retrieval of the abundance of mRNA isoforms (from a short read RNA-seq approach) thus requires the assumption that processes generating the ends of transcripts be independent, as detailed below.

To make the discussion concrete, consider a schematic example, where a locus with three genes has two promoters and an internal partial intrinsic terminator (with order from 5' to 3': promoter 1, gene 1, promoter 2, gene 2, terminator 1, gene 3, terminator 2). The observables from Rend-seq for such an example are the position of the ends of transcription units (denoted x_1^5, x_2^5 and x_1^3, x_2^3 for 5' and 3' ends respectively) and the mean RNA levels between these ends (denoted by n_1, n_2 and n_3). The corresponding abundance of the possible isoforms, unknown, are denoted by $n_{1,1}$, $n_{1,2}$, $n_{2,1}$ and $n_{2,2}$ (indices corresponding to the identities of 5' and 3' ends). The inverse problem is to infer $n_{1,1}$, $n_{1,2}$, $n_{2,1}$ and $n_{2,2}$ from the observables n_1, n_2 and n_3. The connection between the observables and the desired quantities is:

$$\begin{aligned} n_1 &= n_{1,1} + n_{1,2}, \\ n_2 &= n_{1,1} + n_{1,2} + n_{2,1} + n_{2,2}, \\ n_3 &= n_{1,2} + n_{2,2}. \end{aligned}$$

To solve the inverse problem, we make the assumption that the ratio of steady-state isoform abundances for transcripts across a 3' end is independent of upstream 3' or 5' ends (the independence assumption). For the specific example above, this assumption amounts to (first and second equalities):

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\frac{n_{1,2}}{n_{1,1}+n_{1,2}} = \frac{n_{2,2}}{n_{2,1}+n_{2,2}} = \frac{n_{1,2}+n_{2,2}}{n_{1,1}+n_{1,2}+n_{2,1}+n_{2,2}} = \frac{n_3}{n_2} \coloneqq f_1,$$

where the defined parameter f_1 is associated with the 3' end at x_1^3 (in word, the fraction of read density remaining past the 3' end at x_1^3.

The total abundance of all transcripts with the same 5' end is equal (in units of read density) to the difference in total read densities (i.e., observables) after and before the 5' end. For the specific example above, this mean:

$$n_{1,1}+n_{1,2} = n_1 \coloneqq \rho_1 \text{ and } n_{2,1}+n_{2,2} = n_2 - n_1 \coloneqq \rho_2,$$

where we have defined ρ_I as the summed abundance of all transcripts starting at the 5' end at x_i^5. With these definitions, the solution of the inverse problem, in terms of accessible observables, is:

$$\begin{aligned} n_{1,1} &= \rho_1(1-f_1), \\ n_{1,2} &= \rho_1 f_1, \\ n_{2,1} &= \rho_2(1-f_1), \\ n_{2,2} &= \rho_2 f_1 . \end{aligned}$$

Although the above discussion was for a very specific example, the argument can be generalized. We detail below a procedure to reconstruct the isoform abundance from Rend-seq data for arbitrarily complex regions, given the independence assumption explained above.

Derivation—For the purpose of the isoform abundance reconstruction algorithm, we focus our attention on a specific genomic region. We first identify ends of transcription units by thresholding on the peak z score (see earlier section). Suppose we find N ends. Denote the position of the 5' ends by $\{x_1^5, x_2^5, \ldots, x_r^5\}$ (ordered by position, from low to high) and the 3' ends by $\{x_1^3, x_2^3, \ldots, x_s^3\}$ (also ordered by position), where $r + s = N$ (we take the convention that increasing x is from 5' to 3', so that regions on the reverse strand are flipped from left to right before performing this analysis). For the purpose of the reconstruction algorithm, we assume that x_1^5 and x_s^3 are the first and last end respectively (i.e., 0 read density for $x > x_s^3$ and $x < x_1^5$). It thus suffices to focus on the Rend-seq signal coming from the genomic region between x_1^5 and x_s^3.

What are the possible mRNA isoforms corresponding to these ends of transcripts? For each 5' end, there is, by the independence assumption, one isoform for each downstream 3' end. Denote $n_{i,j}$ the abundance of the isoform starting at the 5' end at x_i^5 and ending at the 3' end

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

at x_j^3. Note that not all combinations of i and j are possible (see example below), in which case $n_{i,j} = 0$ for these non-existant combinations.

The mean read density between each end is then determined (leaving appropriate gaps around peaks to ensure the peak signal does not contribute to the measured mean). Denote the average levels between ends by $\{n_1, n_2, \ldots, n_{N-1}\}$ (with the position of regions ordered from 5' to 3'), where n_k is defined as the mean read count in region k.

We wish to solve the inverse problem of determining the $n_{i,j}$ (individual isoform abundance) from the n_i (sum of isoform abundances between ends). Fig. S2J illustrates the notation and quantities involved for the genomic region starting with *dnaA* and ending with *gyrA* in *B. subtilis*. In that example, there is no isoform starting at x_2^5 and ending at x_2^3, simply because $x_2^5 > x_2^3$. Hence, $n_{2,2} = 0$, etc. Note also how each mean read counts between ends of transcription units is the sum of a multitude of isoforms (under our independence assumption). For example, in the *dnaA* region, $n_5 = n_{1,3} + n_{1,4} + n_{2,3} + n_{2,4} + n_{3,3} + n_{3,4}$).

To solve the inverse problem, we focus on one 5' end at a time and derive the abundance of all the transcription units starting at that 5' end. Consider first the 5' end located at x_1^5 (the 5' end upstream of *dnaA* in the example of Fig. S2J). In order to compute required abundances (here $n_{1,1}$, $n_{1,2}$, $n_{1,3}$, $n_{1,4}$), two types of quantities are needed (as introduced in the previously described schematic example). First, we need the total abundance of transcripts starting at the 5' end in question, here $\rho_1 = n_1 = n_{1,1} + n_{1,2} + n_{1,3} + n_{1,4}$). In general, we have, for the total abundance of transcripts with x_i^5 as their 5' ends,

$$\rho_i = n_l - n_{l-1} (i > 1),$$

where x_i^5 is the end between regions $l - 1$ and l. We also assume $\rho_1 = n_1$) (no isoform with a 5' end upstream of the first considered 5' end in the region).

The second type of quantity needed is the fraction of density remaining past 3' ends, denoted f_j for the 3' end at x_j^3, or

$$f_j = \frac{n_k}{n_{k-1}} < 1,$$

where x_j^3 is the end separating regions k and $k - 1$ (note that $n_{k-1} > n_k$), since the mean read density should go down at a 3' end). We assume $f_s = 0$ (no read density past the last 3' end of the considered region).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Returning to the example of $n_{1,1}$, $n_{1,2}$, $n_{1,3}$ and $n_{1,4}$ from Fig. S2J, consider the isoform starting at x_1^5 and ending at x_1^3. By definition of f_1, a fraction $(1 - f_1)$ of all the isoforms starting at x_1^5 end at x_1^3. Indeed:

$$1 - f_1 := 1 - \frac{n_2}{n_1} = \frac{n_{1,1}}{n_{1,1} + n_{1,2} + n_{1,3} + n_{1,4}} = \frac{n_{1,1}}{\rho_1}.$$

Hence, $n_{1,1} = \rho_1(1 - f_1)$ (recall that $\rho_1 = n_1 = n_{1,1} + n_{1,2} + n_{1,3} + n_{1,4}$). For the second isoform, we have $n_{1,2} = \rho_1 f_1(1 - f_2)$, since now a fraction f_1 of transcripts starting at x_1^5 continue across the 3' end at x_1^3, but only a fraction $(1 - f_2)$ of those further stop at the 3' end at x_2^3. Since the decreases in read density across 3' ends are assumed independent, these fractions are multiplied together to obtain the overall abundance. Explicitly, we have:

$$f_1(1 - f_2) := \frac{n_{1,2} + n_{1,3} + n_{1,4}}{n_{1,1} + n_{1,2} + n_{1,3} + n_{1,4}} \left(\frac{n_{1,2}}{n_{1,2} + n_{1,3} + n_{1,4}} \right) = \frac{n_{1,2}}{n_{1,1} + n_{1,2} + n_{1,3} + n_{1,4}} = \frac{n_{1,2}}{\rho_1}.$$

And so indeed, $n_{1,2} = \rho_1 f_1 (1 - f_2)$. Similarly, $n_{1,3} = \rho_1 f_1 f_2(1 - f_3)$. In general, the solution is:

$$n_{i,j} = \rho_i(1 - f_j) \prod_{\substack{k \text{ s.t.} x_k^3 > x_i^5 \\ k < j}} f_k.$$

Although the example with the first 5' end was the simplest, the reasoning also applies for transcripts starting at an internal 5' end, given our independence assumption.

If $\rho_i < 0$ or $f_j > 1$ (rarely, a peak is not accompanied by a concomitant measurable change in read density), the corresponding end is treated as inexistent, and the levels n_i's recomputed with the new, reduced, set of ends.

The abundance reconstruction scheme can be summarized as follows:

- **i.** Identify the positions of the 5' and 3' ends in the region, respectively $\{x_1^5, x_2^5, \ldots, x_r^5\}$ and $\{x_1^3, x_2^3, \ldots, x_s^3\}$.
- **ii.** Determine the mean read density between these ends, $\{n_1, n_2, \ldots, n_N\}$.
- **iii.** For each 5' end, compute the summed abundances ρ_i of the transcripts starting at that end.
- **iv.** For each 3' end, compute the fraction f_j of density remaining past that 3' end.
- **v.** Compute the abundance $n_{i,j}$) of transcript starting at x_i^5 and ending at x_j^3 (provided it exists).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The isoform abundances estimated for the *dnaA* region in *B. subtilis* are shown in Fig. S2J and the obtained values for *rpsB* and *rpsP* operon displayed in Fig. S3B and F (with minor modification, see section “Validation of isoform abundance").

Mathematical derivation of end-enrichment

Heuristic description—In order to understand the origin of end enrichment, we mathematically detail the purely random fragmentation process. Consider a discrete linear chain of $L + 1$ nodes (in our case, ribonucleotides) connected by L links (phosphodiester bonds). We will denote the position of the nodes along the chain by roman indices (e.g., i or j). The fragmentation of the chain is random if each link is broken independently from all the others with the same probability, which we will denote by p. Specifically, this means that for an ensemble of chains undergoing the same fragmentation process, any given link will be broken in a fraction p of the chains in the ensemble.

In such a process, consider the probability for a given chain to have an unfragmented internal fragment of size ℓ (i.e., containing ℓ nodes) starting at internal node i ($i > 1$ for an internal fragment) and ending at node $I + \ell - 1$ (assuming $i + \ell - 1 < L + 1$). For the fragment to begin at position i, the link between node $i - 1$ and i needs to be broken, which occurs with probability p. For the segment to be of length ℓ nodes, the next $\ell - 1$ links need to *not* be broken. Since a link remains intact with probability $1 - p$, and the links are all independent by assumption, the probability of such event is $(1 - p)^{\ell-1}$. Finally, if the segment ends at node $i + \ell$, the link between node $i + \ell - 1$ and $i + \ell$ needs to be broken, which occurs with probability p. Overall, multiplying these three contributions, we thus have a probability $p^2(1 - p)^{\ell-1}$ to have a fragment of size ℓ starting at internal position i along the chain (also ending at an internal position). Note that by independence of the links, there is no need to consider what happens to the rest of the chain.

In contrast, consider the probability of having a fragment of size ℓ starting at one of the original end (e.g., at node $i = 1$ with the above notation). The probability for the fragment to be of size ℓ is as above arising from the probability to not have cleavage $\ell - 1$ consecutive times, or $(1 - p)^{\ell-1}$, and then to have a cleavage at the link between node ℓ and $\ell + 1$, with probability p. Hence, the probability to have a fragment terminating at one of the original end (thereby a terminal fragment) of the chain and of size ℓ is $p(1 - p)^{\ell-1}$.

We thus see that ratio of probabilities of obtaining a terminal versus an internal fragment is p^{-1}. In words: because the difference between an internal and a terminal fragment is the number of cleavages, two and one respectively, internal fragments are p times less common than terminal fragments. This argument holds for any size of fragments. As an extreme instance of the above, having no fragmentation at all would leave only a “terminal" fragment (the original, unfragmented molecule).

The enrichment factor is also strictly restricted to the ends of the chain. Even a fragment starting at node $i = 2$ is internal since the link between node $i = 1$ and $i = 2$ has to be broken in the process. The distinction is binary (either a node is an end, or it is not an end), which is the reason for the high resolution of Rend-seq (i.e., enrichment is only observed at the end of transcription units).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Formal derivation—The heuristic explanation from the previous section is correct, but it is worthwhile to derive the expected (ensemble averaged) read count per position resulting from an idealized Rend-seq procedure, especially to understand the different features in the signal (notably the peak shadows, see below).

As before, we assume perfectly random fragmentation. By independence, it suffices to consider one type of transcript at a time. For the current analysis, we consider an isolated transcription unit (i.e., no additional overlapping mRNA isoforms). The derivation below can readily be generalized to more elaborate cases.

Suppose our transcript is L nucleotides long (nodes in the terminology of the previous section), meaning that $L - 1$ links (labelled by indices $i = 1,\ldots, L - 1$, with link i taken to be between node i and $i + 1$) can be broken during fragmentation. At the end of the fragmentation process, the resulting state of the fragmented chain is uniquely determined by whether each link is broken or not, i.e., by a set of $L - 1$ binary variables denoted by σ_i (with $i = 1,\ldots L - 1$), where $\sigma_i = 1$ if link i is broken in the process, and $\sigma_i = 0$ otherwise. There are thus 2^{L-1}) possible final configurations after the fragmentation of a single chain (e.g., $\sigma_i = 1$ for all i for a completely fragmented chain, or $\sigma_i = 0$ for all i for a completely intact chain). We denote a final fragmentation configuration by $\{\sigma_i\}_{i=1}^{L-1}$ (i.e., a list of these binary variables). Each link has a probability p of being broken (probability $1 - p$ not to be broken), independently of all the others (by assumption). The probability of any final configuration $\{\sigma_i\}_{i=1}^{L-1}$, $P_{conf}(\{\sigma_i\}_{i=1}^{L-1})$, is then given by:

$$P_{conf}(\{\sigma_i\}_{i=1}^{L-1}) = \prod_{s=1}^{L-1} p^{\sigma_s}(1-p)^{1-\sigma_s}.$$

In order to derive the Rend-seq (ensemble averaged) read counts per position, we need to determine the probability to obtain a fragment starting at node i and ending at node j ($j > i$) without any intervening cleavage.

Suppose first that node i is internal ($i > 1$). This is equivalent to computing the probability that after the fragmentation, $\sigma_{i-1} = \sigma_j = 1$, $\sigma_i = \sigma_{i+1} = \ldots = \sigma_{j-1} = 0$, without constraints on other links. In order to obtain the wanted probability, we sum over probabilities of all the final configurations (a final configuration corresponds to choice for the $\{\sigma_i\}_{i=1}^{L-1}$, and each 2^{L-1}) choices are disjoint events in our probability space) constrained to have a non-broken segment from node i to j (the above constraint on the σ's). Explicitly (defining $P_{\text{int}}(i,j)$ the probability to obtain an internal fragment from i to j):

$$P_{int}(i,j) = p^2(1-p)^{j-i} \sum_{\sigma_1=0}^{1} \cdots \sum_{\sigma_{i-2}=0}^{1} \sum_{\sigma_{j+1}=0}^{1} \cdots \sum_{\sigma_{L-1}=0}^{1} \prod_{s \notin \{i-1, i, \ldots, j\}} p^{\sigma_s}(1-p)^{1-\sigma_s}.$$

The initial factor of $p^2(1 - p)^{j-i}$ comes from the constrained variables $\sigma_{i-1},\ldots,\sigma_j$. By rewriting the sum of products as a product of sums, we have:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$P_{\text{int}}(i,j) = p^2(1-p)^{j-i} \prod_{\substack{s=1 \\ s \notin \{i-1, i, \ldots, j\}}}^{L-1} \sum_{\sigma_s=0}^{1} p^{\sigma_s}(1-p)^{1-\sigma_s} = p^2(1-p)^{j-i}.$$

The last equality comes from the fact that since the other variables are unconstrained and independent, each individual sum must be 1 (i.e., $p + (1-p) = 1$). We thus arrive at the same result as for our heuristic (see the previous section). A similar argument leads to $P_{term}(1,j) = p(1-p)^{j-1}$, where $P_{term}(1,j)$ is defined as the probability to obtain a terminal segment starting at node 1 and ending at node j ($j > i$).

These two quantities are sufficient to compute the expected Rend-seq mean read counts per position. Recall that after fragmentation, short fragments with sizes ranging from ℓ_{min} to ℓ_{max} (respectively 15 and 45 nt in our protocol) are selected from the fragmented pool.

Consider the 5' ends of mapped fragments. We will denote the probability to obtain a fragment with 5' end at node i within our size range of ℓ_{min} to ℓ_{max} by $P(i)$. We need to consider three different cases.

First, the probability to have a terminal fragment starting at $i = 1$ ($:= P(i = 1)$) is the sum over the probabilities to have fragments starting at $i = 1$ and ending between $j = \ell_{min}$ and $j = \ell_{max}$ (note, a segment starting at node i and ending at node j is $j - i$ links long), or

$$P(i=1) = \sum_{j=\ell_{min}}^{\ell_{max}} P_{term}(1,j) = p \sum_{s=\ell_{min}}^{\ell_{max}} (1-p)^{s-1} = (1-p)^{\ell_{min}-1} - (1-p)^{\ell_{max}}.$$

Second, the probability to have an internal fragment starting at i, with $2 \le i < L - \ell_{max} + 1$, is again the sum over the respective probabilities, but now from internal fragments:

$$\begin{aligned} P(i, 2 \le i < L - \ell_{max} + 1) &= \sum_{j=i+\ell_{min}-1}^{i+\ell_{max}-1} P_{int}(i,j) = p^2 \sum_{s=\ell_{min}}^{\ell_{max}} (1-p)^{s-1} \\ &= p\{(1-p)^{\ell_{min}-1} - (1-p)^{\ell_{max}}\} = p\,P(i=1). \end{aligned}$$

As already discussed, the signal from at an internal position will be suppressed relative to the signal from the end of the transcript end by a factor p.

Finally, we consider the segments with 5' ends in the range $L - \ell_{max} + 1 \le i \le L - \ell_{min} + 1$ (no fragment can have its 5' end with $i > L - \ell_{min} + 1$ by our size selection assumption). The read counts from fragments in this region come from two sources: (1) from fragments with their 3' ends at the end of the transcription unit and (2) from regular internal fragments. Note that because of the end enrichment (that applies equally at both ends), fragments of type (1) above will be weighted differently (as before). So:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$P(i, L-\ell_{max}+1 \le i \le L-\ell_{min}+1) = P_{term}(i,L) + \sum_{j=i+\ell_{min}-1}^{L-1} P_{int}(i,j)$$
$$= p(1-p)^{L-i} + p^2 \sum_{s=\ell_{min}-1}^{L-i-1} (1-p)^s$$
$$= p(1-p)^{\ell_{min}-1}.$$

The last equality above comes from evaluating the sum explicitly. Hence, if we start with N full length transcripts, the (ensemble averaged) Rend-seq read count (in a context of perfect conversion to a cDNA library and sequencing) for the 5' ends of mapped reads, denoted by $n_5(i)$ ($i = 1$ corresponds to the +1 position of the mRNA) is *NP(i)*, or:

$$n_5(i) = N \begin{cases} (1-p)^{\ell_{min}-1} - (1-p)^{\ell_{max}} & \text{if } i = 1 \quad \text{(peak)} \\ p[(1-p)^{\ell_{min}-1} - (1-p)^{\ell_{max}}] & \text{if } 2 \le i < L-\ell_{max}+1 \quad \text{(uniform internal read density)} \\ p(1-p)^{\ell_{min}-1} & \text{if } L-\ell_{max}+1 \le i \le L-\ell_{min}+1 \quad \text{(peak shadow)} \\ 0 & \text{otherwise.} \end{cases}$$

The read count for the 3' ends of mapped reads is the left-right reflection of the above, i.e., for $1 \le i \le L$, $n_3(i) = n_5(L - i + 1)$, and 0 otherwise. Provided $L > 2\ell_{max}$, the internal mean read density does not depend on L and is proportional to N, the initial number of corresponding molecules in the starting pool of RNA. These two characteristics are pre-requisites for the method to provide quantitative information about mean mRNA levels. We also note that the peak height is directly proportional to N, such that peak height and mean internal read densities provide two different measures of a transcripts' abundance (assuming only full length transcripts are present). The linearity between peak height and internal read density is verified in Fig. S2C. The solution above was verified to be correct by comparison to stochastic simulations of the fragmentation process (not shown). This idealized signal is illustrated for the case of *B. subtilis rpmE* transcript in Fig. S2G. Given that substantial read counts can sometimes be seen at ±1 positions relative to the identified ends (c.f., Fig. S2B), whereas the above model assumes a perfectly defined end, fitting the Rend-seq trace to have the same height at the peak maximum in the data (as done in Fig. S2G) will underestimate the end-enrichment and thus the peak shadow (below).

We finally note that the 5'-mapped read density close to the 3' end of the transcription unit (in the range $L - \ell_{max} + 1 \le i \le L - \ell_{min} + 1$ above) is higher than the mean internal read density. This is the phenomenon of peak shadow (see Fig. 3A and S2G–I). Intuitively, the extra reads arising from the original 3' end must be distributed in a range of size ℓ_{min} to ℓ_{max} upstream for the 5' ends of these mapped reads. As the end-enrichment factor becomes large (p becomes small) there will be an obligate increase in signal by a factor of p^{-1} $(\ell_{max} - \ell_{\min})^{-1}$ in this region (i.e., if the extra signal coming from end enrichment is not "diluted" in the size selection window). See the earlier section on shadow removal for details of how the shadows are systematically removed for displaying data.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Scaling with fragmentation time—Previous sections described the mathematical details of the fragmentation process in terms of the per base cleavage probability, denoted by p. The easiest experimentally controllable parameter for fragmentation is the time t during which the RNA is fragmented. The question is then how to relate the fragmentation probability to the fragmentation time.

Assuming the fragmentation reagents are in large excess and not depleted in the course of the process, fragmentation is expected to occur at a constant rate k_{frag} for each base, in a memoryless process. As a result, the probability for a given base to not be fragmented will be exponentially decreasing with time. The probability to be fragmented is then simply one minus the probability to not be fragmented. We thus have for the per base cleavage probability as a function of time:

$$p(t) = 1 - \mathrm{e}^{-k_{frag}t} \approx k_{frag}t (\text{for } k_{frag}t \ll 1).$$

This, together with the results of the previous section (demonstrating that end-enrichment should be equal to p^{-1}), predicts that end-enrichment is inversely proportional to fragmentation time at short times, which is indeed what we observe (Fig. 3B). In addition, from the end enrichment, we estimate $k_{frag} \approx 7\times10^{-4}\ \mathrm{s}^{-1}$ for our fragmentation conditions. We note that arbitrarily high end-enrichment cannot be achieved in part because of the trade-off with the amount of material remaining post-fragmentation after the size selection. This amount decreases with decreasing fragmentation times at short times.

Divergence of mRNA isoform architecture

Rend-seq data for the gene clusters with conserved differential production identified in our systematic analysis (Fig. 2C, section “Expression in gene clusters: species pairs") were inspected for new 5’ ends (promoter or processed ends) and 3’ ends intervening the conserved genes of interest. Specifically, to be more stringent, changes in transcriptional context outside the conserved gene cluster, were not considered transcriptional remodeling for the purpose of quantification of Fig. 5G. Fig. S4 and Data S1 compiles data for clusters in the four species displaying multiple representative examples of transcript architecture divergences.

Northern blot

Total RNA for Northern blots (same extraction as for Rend-seq library preparation, see “Rend-seq library generation") was run on 1.2% agarose gels containing 20 mM guanidine thiocyanate and transferred to a positively charged nylon membrane (Thermo Fisher Scientific) by downward capillary transfer. Membranes were hybridized (Ultrahyb™-oligo buffer, Thermo Fisher Scientific) with short single stranded DNA probes (Table S4) labelled (T4 PNK, New England Biolabs) with ATP γ–^{32}P (PerkinElmer) and washed following the manufacturer’s protocol. Labelled membranes were exposed to a phosphor storage screen (GE Life Science) between 6 and 35 h and imaged with a laser scanner (Typhoon FLA9500, GE Life Sciences). Membranes were subsequently stripped of transcript specific probes by three washes with boiling 0.1% SDS and hybridized to ^{32}P labelled short single stranded

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

DNA probes to 16S rRNA (Table S4) and imaged as before. Bands intensities were quantified in ImageJ (Schneider et al., 2012). Quantifications from two dilutions of total RNA (3 μg and 1 μg) for each experiment showed complete agreement.

5’ RACE

We used 5’ RACE to validate the location of novel 5’ ends identified from Rend-seq in *B. subtilis*. 1 μg of total RNA from wild-type *B. subtilis* (same extraction protocol as that used for Rend-seq library preparation, see “Rend-seq library generation") was reverse transcribed using SuperScript IV (Thermo Fisher Scientific) following the commercial protocol with random hexamer priming. RNA was hydrolyzed with 50 mM NaOH at 95°C for 5 min, subsequently neutralized with HCl. The resulting cDNA was purified with AMPure beads (Agencourt AMPure, Beckman Coulter). cDNA was tailed using 50 U of terminal deoxynucleotidyl transferase (Thermo Fisher Scientific) at 37°C for 15 min using a 100:1 mixture (final concentration 5 mM) of dATP:ddATP (A-tailing) or dCTP:ddCTP (C-tailing). For each tailing reaction, a negative control without the enzyme was included. Tailing products were purified by isopropanol precipitation. 10 cycles of linear amplification with tail-specific primers with adapter (Table S4) were then performed using the Phusion polymerase (New England Biolabs), followed by column purification (Zymo Research). PCR from linearly amplified tailed-cDNA was then performed with gene specific primers (annealing about 100 nt upstream of the identified end in Rend-seq data). For *rpsP* (which has two close upstream 5’ ends), a single primer upstream of the 5’ end closest to the start of the gene (*rpsP* promoter) was used. PCR products column purified (Oligo Clean and Concentrator, Zymo Research) and ran on 8% TBE polyacrylamide (Thermo Fisher Scientific) gels. Negative controls (no terminal transferase) showed no product. In each case (*pyrH* and *rpsP* specific primers; C and A tailed cDNA), a small (around 120 nt) high intensity product was observed, together with non-specific lower intensity larger products. For the C-tailed *rspP* specific PCR prodcuts, a second larger high intensity product was seen (not for the A-tailed cDNA). The small prominent products were gel extracted and Sanger sequenced. The results are shown in Mendeley Data, confirming with single nucleotide resolution the end position observed in Rend-seq.

Shine-Dalgarno sequence accessibility

Although cis-features of mRNAs causative of translation efficiency (TE) differences are still incompletely characterized, we compared RNA secondary structure near the start of genes for which ribosome profiling and Rend-seq report very low translation efficiency (less than 0.15× the median translation efficiency) in a subset of the species considered. These include the *rpsP* operon (*rimM* and *trmD* low TE genes in *E. coli* and *V. natriegens*, see Fig. S4) and the S10 operon (*secY* low TE gene in *B. subtilis, E. coli*, and *V. natriegens*, see Data S2).

Briefly, we identified the Shine-Dalgarno (SD) sequence upstream of the start codons of genes of interest. The accessibility (one minus the probability for a position to be base-paired) of the identified SD and start codon based on the RNA secondary structure for the surrounding sequence was then determined. As detailed below, we generally find more structured regions near the start codons of translationally repressed genes compared to homologous, non-translationally repressed, genes in the other species.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The putative Shine-Dalgarno (SD) sequence (defining feature: complementarity to the 3' end of the 16S rRNA) for each gene of interest was identified as follows in *B. subtilis, E. coli*, and *V. natriegens* (see below for *C. crescentus*). We searched for the region upstream of the start codon with highest duplex stability ΔG_{SD} (computed with RNAduplex (Lorenz et al., 2011)) to the conserved region near the 3' end of the 16S rRNA (5' TCACCTCCT). We found plausible SD sequences for most genes of interest (quantities reported: minimum ΔG_{SD}, SD sequence, position relative to the start codon): *B. subtilis*' *rimM* (−11.5 kcal/mol, AGAGGTGA, −12 to −5 nt), *B. subtilis*' *trmD* (−6 kcal/mol, GGAAGGG, −18 to −12 nt), *B. subtilis*' *secY* (−11.3 kcal/mol, GAGGTGA, −13 to −7 nt), *E. coli*'s *rimM* (−6.5 kcal/mol, GGTGGTCA, −10 to −3 nt), *E. coli*'s *secY* (−5.9 kcal/mol, GAGGA, −16 to −12 nt), and *V. natriegens*' *secY* (−7.3 kcal/mol, GAGGTA, −9 to −4 nt). No clear SD sequences could be identified for *E. coli*'s *trmD*, and *V. natriegens*'s *trmD* and *rimM* (maximum duplex stability respectively of −3.1 kcal/mol, −3.4 kcal/mol and −0.3 kcal/mol in the region up to 25 nt upstream of the start codon). For these examples and for *secY* in *C. crescentus* (for which most genes do not have identifiable canonical SD sequences (Schrader et al., 2014)), we considered the region −10 to −5 nt (canonical position in *E. coli*) as the *de facto* SD region for our analysis.

From the putative SD sequences identified above, we estimated the accessibility by first computing the ensemble free energy ΔG (using RNAfold with option -p, (Lorenz et al., 2011)) of folding surrounding the region of interest (at the center) was evaluated. Then, for each position x in the SD sequence, the ensemble free energy for the same region was obtained, now with the added constraint that the nucleotide at position x had to remain unpaired (using RNAfold with options -p -C, (Lorenz et al., 2011)). We denote this quantity ΔG_x. The probability of not being paired (accessibility) for the position x, A_x, was then obtained as $A_x = e^{(\Delta G - \Delta G_x)/RT}$, where RT=0.59 kcal/mol sets the temperature scale of the thermodynamic calculation. The accessibility for the Shine-Dalgarno sequence, A_{SD}, was taken as the mean of A_x, $A_{SD} = \langle A_x \rangle_{x \in SD}$). We observed that A_{SD} varied somewhat with the size the flanking regions that were folded. As our measure of accessibility, we thus report here the minimum A_{SD} which was constant over a region 50 nt within the range 25 to 500 nt of total folded region size (for *secY* in *C. crescentus*, the upstream region was restricted to -100 nt given a 5' end we observe from Rend-seq). A similar analysis was also performed for the pairing probability of the start codon A_{AUG} for genes considered.

For the *rpsP* operon, this computational analysis revealed more secondary RNA structure surrounding the start codons in the low TE genes *rimM* and *trmD* in *E. coli* and *V. natriegens*. Specifically, for *trmD*, we found a less accessible ribosome binding site ($A_{SD}^{Ecol} = 0.07$, $A_{SD}^{Vnat} = 0.25$ compared to $A_{SD}^{Bsub} = 0.48$) and start codon ($A_{AUG}^{Ecol} = 0.14$, $A_{AUG}^{Vnat} = 0.04$ compared to $A_{AUG}^{Bsub} = 0.80$). For *rimM*, we found a less accessible ribosome binding site ($A_{SD}^{Ecol} = 0.04$, $A_{SD}^{Vnat} = 0.13$ compared to $A_{SD}^{Bsub} = 0.43$), but no large difference in start codon accessibility ($A_{AUG}^{Ecol} = 0.46$, $A_{AUG}^{Vnat} = 0.36$ compared to $A_{AUG}^{Bsub} = 0.35$).

For *secY* in the S10 operon, we also observe additional structure in *B. subtilis, E. coli*, and *V. natriegens* compared to *C. crescentus* (for which the S10 operon is interrupted by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

transcription terminators for *secY*), although to a lesser extent than in the *rpsP* operon. *E. coli* and *V. natriegens* (though not *B. subtilis*) had a less accessible SD sequence than *C. crescentus* ($A_{SD}^{Ecol} = 0.10$, $A_{SD}^{Vnat} = 0.07$ compared to $A_{SD}^{Ccre} = 0.22$; $A_{SD}^{Bsub} = 0.18$). In addition, *B. subtilis* (but not *E. coli* and *V. natriegens*) had a less accessible start codon than *C. crescentus* ($A_{AUG}^{Bsub} = 0.31$ compared to $A_{AUG}^{Ccre} = 0.50$; $A_{AUG}^{Ecol} = 0.51$, $A_{AUG}^{Vnat} = 0.55$).

Our simple computational assessment of SD and start codon accessibility is consistent with previous targeted studies (Wikström et al., 1992) and *in vivo* structural analyses (Burkhardt et al., 2017) of the *rpsP* operon in *E. coli*. The above provides additional evidence, though indirect, supporting the translational versus transcriptional control of differential expression that we observe in different species.

Synthesis rate for the two EF-Tu copies

Ribosome profiling, relying on sequencing of ribosome-protected mRNA fragments, provides information relating to the relative production of closely related paralogs which can be distinguished at the nucleotide level, such as the two copies of EF-Tu in *E. coli* (13 mismatches out of 1185 positions at the nucleotide level) and *V. natriegens* (24 mismatches out of 1185 positions at the nucleotide level). These mismatches can be leveraged for relative quantification of the synthesis rate of the two copies from ribosome profiling data. In particular, note that many more positions than the number of mismatches can be used for the purpose of quantification (e.g., if the mismatches are far from each other, one expects ℓ times the number of mismatches, where ℓ is the minimum read size, to be usable for quantification). We note that the two copies of EF-Tu are 100% at the nucleotide level in *C. crescentus*, which makes it impossible to compare the protein synthesis rates for each respective copy (Rend-seq reads mapping to the untranslated regions of the corresponding transcripts shows suggests that the two copies have near identical mRNA levels, Fig. S4).

For the purpose of quantification, we first generated a tiling of the genome of length 15 nt (i.e., sequence from 1 to 15, sequence 2 to 16, etc.). This tiling was aligned back to the genome with bowtie using options -v 0 -m 1 (retaining sequences mapping without mismatch to at most one position). The 5’ end position of the mapped sequences were stored in the form of a “mask" (1 at position x if the 5’ end of the sequence starting at x was retained under the above procedure, 0 otherwise).

We thus aligned ribosome footprint reads in bowtie using the -v 1 -k 2 -m 2 option. The 5’ ends of reads with a single reported alignment were counted as before. Reads with two reported alignments were treated as follows. If only one of the reported alignment had no mismatch, the perfect alignment was kept. If both reads had a single mismatch but one had a mismatch corresponding to non-template addition at the 5’ end, the alignment with non-template addition mismatch was kept. Otherwise (two perfect alignments, two alignments each with non-template addition mismatches, or two alignments with mismatches not corresponding to non-template addition), reads were discarded from the current analysis.

The relative synthesis rate of the two copies was then taken as the mean reads counts mapping to positions with mask values of 1 (and excluding 15 nt from the start and stop

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

codon). We find 63% and 82% relative synthesis for *tufA* vs. *tufB* in *E. coli* and *V. natriegens* respectively.

We finally note that the three copies of EF-G in *V. natriegens* were sufficiently different (54% amino acid identity between *fusA* (on the first chromosome) and PN96_01780; 30% (renamed *fusB* in the current work) identity between *fusA* (first chromosome) and *fusA* (on the second chromosome, renamed *fusC* in the current work) to be quantified using the usual approach.

Position-dependent gene dosage

The loss of operon linkage for many of the metabolic pathways shown in Fig. 6 and corresponding scattering of the genes across the chromosome highlight a point generally valid for most of our curated pathways: chromosomal positions, and specifically, relative position from the origin of replication, are not well conserved for conserved proteins between *B. subtilis* and *E. coli*, even for proteins of the translation class (Fig. S5A–C).

Due to the phenomenon of multi-fork replication during fast growth, a well-established effective gene dosage gradient from the origin of replication to the terminus occurs, with the mean gene copy number for a gene at position x, denoted by g_x, equal to $g_x = 2^{(Cx+D)/\tau}$ (we use a normalized coordinate where x is equal to 1 when x is at the origin and 0 when at the terminus) (Bremer and Dennis, 1987). C and D here denote the length of the C and D period respectively, and τ is the cell doubling time. The relative (to the terminus) copy number, or relative dosage ($\overline{g}_x$) for the current analysis, is then equal to $\overline{g}_x = 2^{Cx/\tau}$. The differential gene dosage from origin to terminus is thus dictated by the ratio C/τ. For fast growth, experimental values reported in *E. coli* for C/τ vary between 1.65 to 2 (Bremer and Dennis, 1987), corresponding to $\overline{g}_{oric}/\overline{g}_{Ter}$ ranging between 3 to 4. Reported values of $\overline{g}_{oric}/\overline{g}_{Ter}$ are closer to 4 for *B. subtilis* in LB, e.g., (Soufo et al., 2008). As a conservative estimate (that likely underestimates the effect of interest), we pick $\overline{g}_{oric}/\overline{g}_{Ter} = 3$ for both *E. coli* and *B. subtilis* for the current analysis.

This differential gene dosage from origin to terminus is on the same order as the spread around conserved expression stoichiometry for our curated pathways, and substantially larger than the precision of our synthesis rate measurement. We computed the expression conservation score (defined as the fraction of genes expressed with a fold-deviation smaller than 2, e.g., the fractions highlighted in the insets of Fig. 1'A and B) for the various pathways for both (1) synthesis rates ($\alpha^{-1} k_i^{Ecol}/k_i^{Ecol}$) and (2) synthesis rates with reshuffled gene dosages.

The fold-deviation for a given homolog i for reshuffled dosages was calculated by dividing the measured synthesis rate by the relative dose, and then multiplied by a reshuffled dose. Mathematically, this corresponds to $\alpha^{-1}\left(k_i^{Ecol}\overline{g}_{\hat{x}_i}^{Ecol}/\overline{g}_{x_i}^{Ecol}\right)/\left(k_i^{Bsub}\overline{g}_{\hat{x}_i}^{Bsub}/\overline{g}_{x_i}^{Bsub}\right)$ (α is the differential expression factor for the pathway of interest, identical to that in the discussion in "Expression in gene clusters: species pairs"). $\overline{g}_{\hat{x}_i}^{Ecol}$ and $\overline{g}_{\hat{x}_i}^{Bsub}$ denote the reshuffled dosages

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(random permutations of the observed dosages, but keeping dosages for homologs between the two species paired to maintain possible residual correlations in dosages). For endogenous dosages, the expression conservation score was 87%, 82% and 83% for translation, DNA maintenance and the metabolic pathways considered. For reshuffled dosages, the respective median expression conservation score (across reshuffling) were 55%, 61% and 62% (Fig. S5D). Overall, less than one in 10^4, 10^3 and 10^4 reshuffled dosages (translation, DNA maintenance and metabolic pathways respectively) had a higher expression conservation score than the one observed for the actual dosages, even when correcting for co-clustering of genes (i.e., treating the genes co-localized within 0.1 Mb in both *B. subtilis* and *E. coli*, such as the > 20 the genes in the S10 operon, as a single entity for this analysis).

The significance of the tighter distribution of synthesis rates for endogenous compared to reshuffled dosages suggests that compensatory mutations, subsequent to movement along the chromosome (or the presence of feedback mechanisms) were necessary for the tight expression stoichiometry conservation we observe to be possible.

QUANTIFICATION AND STATISTICAL ANALYSIS

In Fig. 4E, *n* refers to the number of tuned terminators identified in *B. subtilis* (see section “Contribution of tuned terminators"). In Fig. 4F and S3N, *n* refers to the number (total, and for each U-tract length) of identified terminators (see section “Quantification of read-through fraction"). In Fig. 7C, *n* refers to the number of groups of paralogs in each of the category (number of paralog copies per species for compared proteins, see section “Comparison between yeast and bacteria"). In Fig. S2M and S2N, *N* represents the total number of genomic positions passing the read depth cuts (see section “Detection sensitivity").

To test statistical significance of read-through fraction differences between terminators with different U-tract lengths (Fig. 4F, S3N), two-sample t-tests comparing the read-through fractions from terminators with consecutive U-tract length were used. To assess lack of significant differences in fold deviations for synthesis rates of groups of paralogs with different gene copies (Fig. 7C), a two-sample t-test and a two-sample Kolmogorov-Smirnov test were used.

See section “Expression stoichiometry and position-dependent gene dosage" for the statistical testing relating to the importance of position-dependent dosage and expression. Other tests for significance are described below.

Significance of protein expression conservation

In order to test for the significance of the observed expression conservation, we compared the expression in our pathways and clusters to samplings of the expression space in the species studied.

Specifically, for each pair of proteins (compared across species) of a given pathway or gene cluster, the ratio of synthesis rate between the two species compared was calculated. For

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

each pathway/cluster, we then computed the range of observed ratios, defined as the maximum over the minimum ratio (in words, the span of the fold-deviations). For perfect stoichiometry, this range of ratio would be 1. For a pathway with N members, we sampled randomly N synthesis rates from all proteins expressed above our threshold in the two species compared. Ratios of synthesis for this random sampling were computed. From the ratios, the range of ratios for the random sampling was stored. The sampling was repeated 10^6 times for pathways and 10^5 times for clusters. The fraction of times the range of ratio was smaller than that observed for the pathway of interest was taken as our p-value, which are listed in Table S2 for pathways and for gene clusters. For large pathways (e.g., translation and the DNA maintenance pathways), we used the more conservative metric (which generates a larger p-value) of determining the fraction of proteins with less than 2× deviations in the resampling. The fraction of times that the sampling had a larger fraction within 2× than the pathway was then taken as our significance. Significance for pathways and clusters are listed in Table S2.

In some of the pathways with few members ($N = 3$), expression was conserved, but we did not have the statistical power to assert that homologs were more conserved than randomly sampled proteins.

Correction for co-transcription

Many pathways contain proteins that remain co-transcribed in gene clusters, which might maintain the same historical expression stoichiometry. To account for such apparent contribution to conservation, we collapsed all co-transcribed genes into a single entity and examined its conservation with the rest of the respective pathway. Operationally, we used a conservative definition of co-transcription, considering two genes to be co-transcribed if they were less than 5 kb apart and with mRNA level (determined from mean winsorized Rend-seq signal across the coding sequence, leaving 45 nt gaps around start and stop codons) differing by less than 30%. If these two criteria were satisfied in the two species compared, the genes were treated as co-transcribed. The co-transcription corrected set for statistical testing were then the connected components of the graph whose nodes were the homologs compared and the connectivity matrix defined by the operational co-transcription criterion above. As an example, this criterion collapses conserved members of the S10 operon in *E. coli* and *B. subtilis* to a single entity. The synthesis rate of the co-transcribed cluster was taken as the median synthesis rate of the members of the cluster.

Expression conservation and synteny

To assess whether the number of gene clusters categorized as having conserved expression stoichiometry (Fig. 2C) was significant, we compared the expression of clusters of homologs sampled randomly. Specifically, the spatial clustering of genes (see section “Expression in gene clusters: species pairs") generated a set of M clusters, with cluster i having N_i members. A set of “random clusters" was generated by sampling the synthesis rate of expressed of one-to-one homologs in the species pair compared (from pairwise best BLASTP hits). Thus, M sets of pairs (one per species) of protein synthesis rates were obtained (with the number of sampling per set determined by the cluster sizes, i.e., N_1, N_2, …, N_M). The same categorization as described in section “Expression in gene clusters:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

species pairs" was then performed for each randomly sampled set (“random clusters"). The fraction of clusters categorized as conserved category. This constituted one sampling. 10^3 samplings were performed. The significance of the categorization was taken as the fraction of samplings with fraction of “random clusters" categorized as conserved larger than the fraction observed (Fig. 2C). In parallel (to avoid possible issues coming from categorization criteria), the distribution of fold-deviations for each homolog pair across clusters was determined for each sampling. The fraction of proteins with fold-deviations smaller than 2× was also stored (expression score) for each sampling. The fraction of samplings with expression score larger than the observed one constituted another p-value ($p < 2\times10^{-3}$) for all pairs).

DATA AND SOFTWARE AVAILABILITY

Ribosome profiling and Rend-seq data are available at the Gene Expression Omnibus with accession number GSE95211.

Core scripts used for Rend-seq data analysis were deposited to Github at https://github.com/jblalanne/Rend_seq_core_scripts

Other custom Python and Matlab scripts for analysis are available upon reasonable request.

Extensive Rend-seq validation data, mRNA abundances from Rend-seq, translation efficiencies (from ribosome profiling and Rend-seq), and raw Northern blot images have been deposited to Mendeley data: doi:10.17632/ncm3s3pk2t.1.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank BM Koo for providing *B. subtilis* single-gene deletion strains; M Laub, J Elf, and members of the GWL and A Grossman labs for discussions; A Amon, D Bartel, and V Siegel for comments on the manuscript; MIT BMC for DNA sequencing. This research is supported by NIH R00GM105913, NIH R35GM124732, Pew Biomedical Scholars Program, Searle Scholars Program, Sloan Research Fellowship, Smith Family Awards, NSERC Fellowship (to JBL), HHMI International Student Research Fellowship (to JBL), NSF Graduate Research Fellowship (to JCT), NIH T32GM007287 (to JCT), Helen Hay Whitney Fellowship (to LH), and a Jane Coffin Childs Memorial Fellowship (to MSG).

References

Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016; 44:D7–D19. [PubMed: 26615191]

Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. 2015; 16:197–212. [PubMed: 25707927]

Alon U. An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman Hall. 2007; 10:301.

Andersen GR, Pedersen L, Valente L, Chatterjee I, Kinzy TG, Kjeldgaard M, Nyborg J. Structural Basis for Nucleotide Exchange and Competition with tRNA in the Yeast Elongation Factor Complex eEF1A:eEF1Ba. Mol. Cell. 2000; 6:1261–1266. [PubMed: 11106763]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Artieri CG, Fraser HB. Evolution at two levels of gene expression in yeast. 2014:411–421.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. Mol. Syst. Biol. 2006; 2

Ban N, Beckmann R, Cate JHD, Dinman JD, Dragon F, Ellis SR, Lafontaine DLJ, Lindahl L, Liljas A, Lipton JM, et al. A new system for naming ribosomal proteins. Curr. Opin. Struct. Biol. 2014; 24:165–169. [PubMed: 24524803]

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. Impact of regulatory variation from RNA to protein. Science. 2014; 347:664–667. [PubMed: 25657249]

Blank LM, Kuepfer L, Sauer U. Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol. 2005; 6:R49. [PubMed: 15960801]

Bremer H, Dennis PP. Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. Escherichia Coli Salmonella Cell. Mol. Biol. 1987; 2:1527–1542.

Brinsmade SR, Alexander EL, Livny J, Stettner AI, Segrè D, Rhee KY, Sonenshein AL. Hierarchical expression of genes controlled by the Bacillus subtilis global regulatory protein CodY. Proc. Natl. Acad. Sci. U.S.A. 2014; 111:2–7.

Burkhardt DH, Rouskin S, Zhang Y, Li GW, Weissman JS, Gross CA. Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. Elife. 2017; 6

Cai L, Dalal CK, Elowitz MB. Frequency-modulated nuclear localization bursts coordinate gene regulation. Nature. 2008; 455:485–490. [PubMed: 18818649]

Cambray G, Guimaraes JC, Mutalik VK, Lam C, Mai QA, Thimmaiah T, Carothers JM, Arkin AP, Endy D. Measurement and modeling of intrinsic transcription terminators. Nucleic Acids Res. 2013; 41:5139–5148. [PubMed: 23511967]

Chen Y-J, Liu P, Nielsen AaK, Brophy JaN, Clancy K, Peterson T, Voigt Ca. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat. Methods. 2013; 10:659–664. [PubMed: 23727987]

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. Saccharomyces Genome Database: The genomics resource of budding yeast. Nucleic Acids Res. 2012; 40

Cho B-K, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ. The transcription unit architecture of the Escherichia coli genome. Nat. Biotechnol. 2009; 27:1043–1049. [PubMed: 19881496]

Christiano R, Nagaraj N, Fröhlich F, Walther TC. Global Proteome Turnover Analyses of the Yeasts S.cerevisiae and S.pombe. Cell Rep. 2014; 9:1959–1966. [PubMed: 25466257]

Commichau FM, Rothe FM, Herzberg C, Wagner E, Hellwig D, Lehnik-Habrink M, Hammer E, Volker U, Stulke J. Novel Activities of Glycolytic Enzymes in Bacillus subtilis. Mol. Cell. Proteomics. 2009; 8:1350–1360. [PubMed: 19193632]

Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J. Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. MBio. 2014; 5:1–12.

Daou-Chabo R, Mathy N, Bénard L, Condon C. Ribosomes initiating translation of the hbs mRNA protect it from 5′-to-3′ exoribonucleolytic degradation by RNase J1. Mol. Microbiol. 2009; 71:1538–1550. [PubMed: 19210617]

Dar D, Shamir M, Mellin JR, Koutero M, Stern-Ginossar N, Cossart P, Sorek R. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science. 2016; 352:aad9822–aad9822. [PubMed: 27120414]

Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005; 436:588–592. [PubMed: 16049495]

DeLuna A, Springer M, Kirschner MW, Kishony R. Need-based up-regulation of protein levels in response to deletion of their duplicate genes. PLoS Biol. 2010; 8:e1000347. [PubMed: 20361019]

DiChiara JM, Liu B, Figaro S, Condon C, Bechhofer DH. Mapping of internal monophosphate 5′ ends of Bacillus subtilis messenger RNAs and ribosomal RNAs in wild-type and ribonuclease-mutant strains. Nucleic Acids Res. 2016; 44:3373–3389. [PubMed: 26883633]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Donovan WP, Kushner SR. Polynucleotide phosphorylase and ribonuclease II are required for cell viability and mRNA turnover in Escherichia coli K-12. Proc. Natl. Acad. Sci. U.S.A. 1986; 83:120–124. [PubMed: 2417233]

Eames M, Kortemme T. Cost-benefit tradeoffs in engineered lac operons. Science. 2012; 336:911–915. [PubMed: 22605776]

Emmerling M, Dauner M, Ponti A, Fiaux J, Hochuli M, Szyperski T, Wüthrich K, Bailey JE, Sauer U. Metabolic flux responses to pyruvate kinase knockout in Escherichia coli. J. Bacteriol. 2002; 184:152–164. [PubMed: 11741855]

Evinger M, Agabian N. Envelope associated nucleoid from Caulobacter crescentus stalked and swarmer cells. J. Bacteriol. 1977; 132:294–301. [PubMed: 334726]

Fell D. Understanding the control of metabolism. Front. Metab. 1997; 2:300.

Gilbert, La, Larson, MH., Morsut, L., Liu, Z., Gloria, A., Torres, SE., Stern-ginossar, N., Brandman, O., Whitehead, H., Doudna, Ja, et al. CRISPR-Mediated Modular RNA-Guided Regualtion of Transcription in Eukaryotes. Cell. 2013; 154:442–451. [PubMed: 23849981]

Hackett SR, Zanotelli VRT, Xu W, Goya J, Park JO, Perlman DH, Gibney PA, Botstein D, Storey JD, Rabinowitz JD. Systems-level analysis of mechanisms regulating yeast metabolic flux. Science. 2016; 354:aaf2786. [PubMed: 27789812]

Harper JW, Bennett EJ. Proteome complexity and the forces that drive proteome imbalance. Nature. 2016; 537:328–338. [PubMed: 27629639]

Harwood, CR., Cutting, SM. Molecular Biological methods for Bacillus. John Wiley; 1990.

De Hoon MJL, Makita Y, Nakai K, Msyasio S. Prediction of transcriptional terminators in bacillus subtilis and species. PLoS Comput. Biol. 2005; 1:0212–0221.

Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS. Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol. Syst. Biol. 2007; 3:86. [PubMed: 17389874]

Ingham CJ, Dennis J, Furneaux PA. Autogenous regulation of transcription termination factor Rho and the requirement for Nus factors in Bacillus subtilis. Mol. Microbiol. 1999; 31:651–663. [PubMed: 10027981]

Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. 2012; 7:1534–1550. [PubMed: 22836135]

Irnov I, Sharma CM, Vogel J, Winkler WC. Identification of regulatory RNAs in Bacillus subtilis. Nucleic Acids Res. 2010; 38:6637–6651. [PubMed: 20525796]

Kacser H, Burns JA. The molecular basis of dominance. Genetics. 1981; 97:639–666. [PubMed: 7297851]

Kafri R, Levy M, Pilpel Y. The regulatory utilization of genetic redundancy through responsive backup circuits. Proc. Natl. Acad. Sci. 2006; 103:11653–11658. [PubMed: 16861297]

Karr JR, Sanghvi JC, MacKlin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW. A whole-cell computational model predicts phenotype from genotype. Cell. 2012; 150:389–401. [PubMed: 22817898]

Keren L, Hausser J, Lotan-Pompan M, Vainberg Slutskin I, Alisar H, Kaminski S, Weinberger A, Alon U, Milo R, Segal E. Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. Cell. 2016; 166:1282–1294.e18. [PubMed: 27545349]

Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, et al. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res. 2016; 45:gkw1003.

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science. 2013; 342:1100–1104. [PubMed: 24136357]

Khosla C, Keasling JD. Metabolic Engineering For Drug Discovery and Development. Nat. Drug Discov. 2003; 2:1019–1025.

Kisselev L. Polypeptide release factors in prokaryotes and eukaryotes: Same function, different structure. Structure. 2002; 10:8–9. [PubMed: 11796105]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Klumpp S, Scott M, Pedersen S, Hwa T. Molecular crowding limits translation and cell growth. Proc. Natl. Acad. Sci. U.S.A. 2013; 110:16754–16759. [PubMed: 24082144]

Kondrashov FA, Koonin EV. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. Trends Genet. 2004; 20:287–291. [PubMed: 15219392]

Koo BM, Kritikos G, Farelli JD, Todor H, Tong K, Kimsey H, Wapinski I, Galardini M, Cabal A, Peters JM, et al. Construction and Analysis of Two Genome-Scale Deletion Libraries for Bacillus subtilis. Cell Syst. 2017; 4:291–305.e7. [PubMed: 28189581]

Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

Larrabee KL, Phillips JO, Williams GJ, Larrabee AR. The relative rates of protein synthesis and degradation in a growing culture of Escherichia coli. J. Biol. Chem. 1980; 255:4125–4130. [PubMed: 6989832]

Li G-W, Oh E, Weissman JS. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature. 2012; 484:538–541. [PubMed: 22456704]

Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014; 157:624–635. [PubMed: 24766808]

Locke JCW, Young JW, Fontes M, Hernández Jiménez MJ, Elowitz MB, Raj A, Oudenaarden A, van, Rosenfeld N, Young JW, Alon U, et al. Stochastic pulse regulation in bacterial stress response. Science. 2011; 334:366–369. [PubMed: 21979936]

Lomakin IB, Shirokikh NE, Yusupov MM, Hellen CUT, Pestova TV. The fidelity of translation initiation: reciprocal activities of eIF1, IF3 and YciH. EMBO. J. 2006; 25:196–210. [PubMed: 16362046]

Lorenz R, Tafer H, Höner zu Siederdissen C, Stadler PF, Bernhart SH, Hofacker IL, Flamm C. ViennaRNA Package 2.0. Algorithms Mol. Biol. 2011; 6:26. [PubMed: 22115189]

McManus CJ, May GE, Spealman P, Shteyman A. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. Genome Res. 2014; 24:422–430. [PubMed: 24318730]

Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juárez K, Contreras-Moreira B, et al. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. PLoS One. 2009; 4

Michna RH, Zhu B, Mäder U, Stülke J. SubtiWiki 2.0--an integrated database for the model organism Bacillus subtilis. Nucleic Acids Res. 2016; 44:D654–62. [PubMed: 26433225]

Moffitt JR, Pandey S, Boettiger AN, Wang S, Zhuang X. Spatial organization shapes the turnover of a bacterial transcriptome. Elife. 2016; 5

Mondal S, Yakhnin AV, Sebastian A, Albert I, Babitzke P. NusA-dependent transcription termination prevents misregulation of global gene expression. Nat. Microbiol. 2016; 1:1–7.

Morohoshi F, Hayashi K, Munakata N. Bacillus subtilis alkA gene encoding inducible 3-methyladenine DNA glycosylase is adjacent to the ada operon. J. Bacteriol. 1993; 175:6010–6017. [PubMed: 8376346]

Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J. Bacteriol. 1974; 119:736–747. [PubMed: 4604283]

Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science. 2012; 335:1103–1106. [PubMed: 22383849]

Ohno S. Evolution by Gene Duplication. 1970 (1970).

Oromendia AB, Dodgson SE, Amon A. Aneuploidy causes proteotoxic stress in yeast. Genes Dev. 2012; 26:2696–2708. [PubMed: 23222101]

Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO. A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. Mol. Syst. Biol. 2011; 7:535–535. [PubMed: 21988831]

Oussenko IA, Abe T, Ujiie H, Muto A, Bechhofer DH. Participation of 3'-to-5' exoribonucleases in the turnover of Bacillus subtilis mRNA. J. Bacteriol. 2005; 187:2758–2767. [PubMed: 15805522]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature. 2003; 424:194–197. [PubMed: 12853957]

Peters JM, Mooney RA, Grass JA, Jessen ED, Tran F, Landick R. Rho and NusG suppress pervasive antisense transcription in Escherichia coli. Genes Dev. 2012; 26:2621–2633. [PubMed: 23207917]

Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, Hawkins JS, Lu CHS, Koo BM, Marta E, et al. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell. 2016; 165:1493–1506. [PubMed: 27238023]

Saini P, Eyler DE, Green R, Dever TE. Hypusine-containing protein eIF5A promotes translation elongation. Nature. 2009; 459:118–121. [PubMed: 19424157]

Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat. Methods. 2012; 9:671–675. [PubMed: 22930834]

Schrader JM, Zhou B, Li GW, Lasker K, Childers WS, Williams B, Long T, Crosson S, McAdams HH, Weissman JS, et al. The Coding and Noncoding Architecture of the Caulobacter crescentus Genome. PLoS Genet. 2014; 10

Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010; 330:1099–1102. [PubMed: 21097934]

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature. 2010; 464:250–255. [PubMed: 20164839]

Shunsuke I, Kuroki K, Imamoto F. tRNAMetf2 gene in the leader region of the nusA operon in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 1983; 81:409–413.

Sierro N, Makita Y, De hoon M, Nakai K. DBTBS: A database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res. 2008; 36

Sohmen D, Chiba S, Shimokawa-Chiba N, Innis CA, Berninghausen O, Beckmann R, Ito K, Wilson DN. Structure of the Bacillus subtilis 70S ribosome reveals the basis for species-specific stalling. Nat. Commun. 2015; 6:6941. [PubMed: 25903689]

Soufo CD, Soufo HJD, Noirot-Gros MF, Steindorf A, Noirot P, Graumann PL. Cell-Cycle-Dependent Spatial Sequestration of the DnaA Replication Initiator Protein in Bacillus subtilis. Dev. Cell. 2008; 15:935–941. [PubMed: 19081080]

Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, Huang S-X, Ma M, Shen B, Qian S-B, Hengel H, et al. Decoding human cytomegalovirus. Science. 2012; 338:1088–1093. [PubMed: 23180859]

Sugimoto N, Nakano S, Katoh M, Matsumura A. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry. 1995; 34:11211–11216. [PubMed: 7545436]

Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Rep. 2016; 14:1787–1799. [PubMed: 26876183]

Wikström PM, Lind LK, Berg DE, Björk GR. Importance of mRNA folding and start codon accessibility in the expression of genes in a ribosomal protein operon of Escherichia coli. J. Mol. Biol. 1992; 224:949–966. [PubMed: 1569581]

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

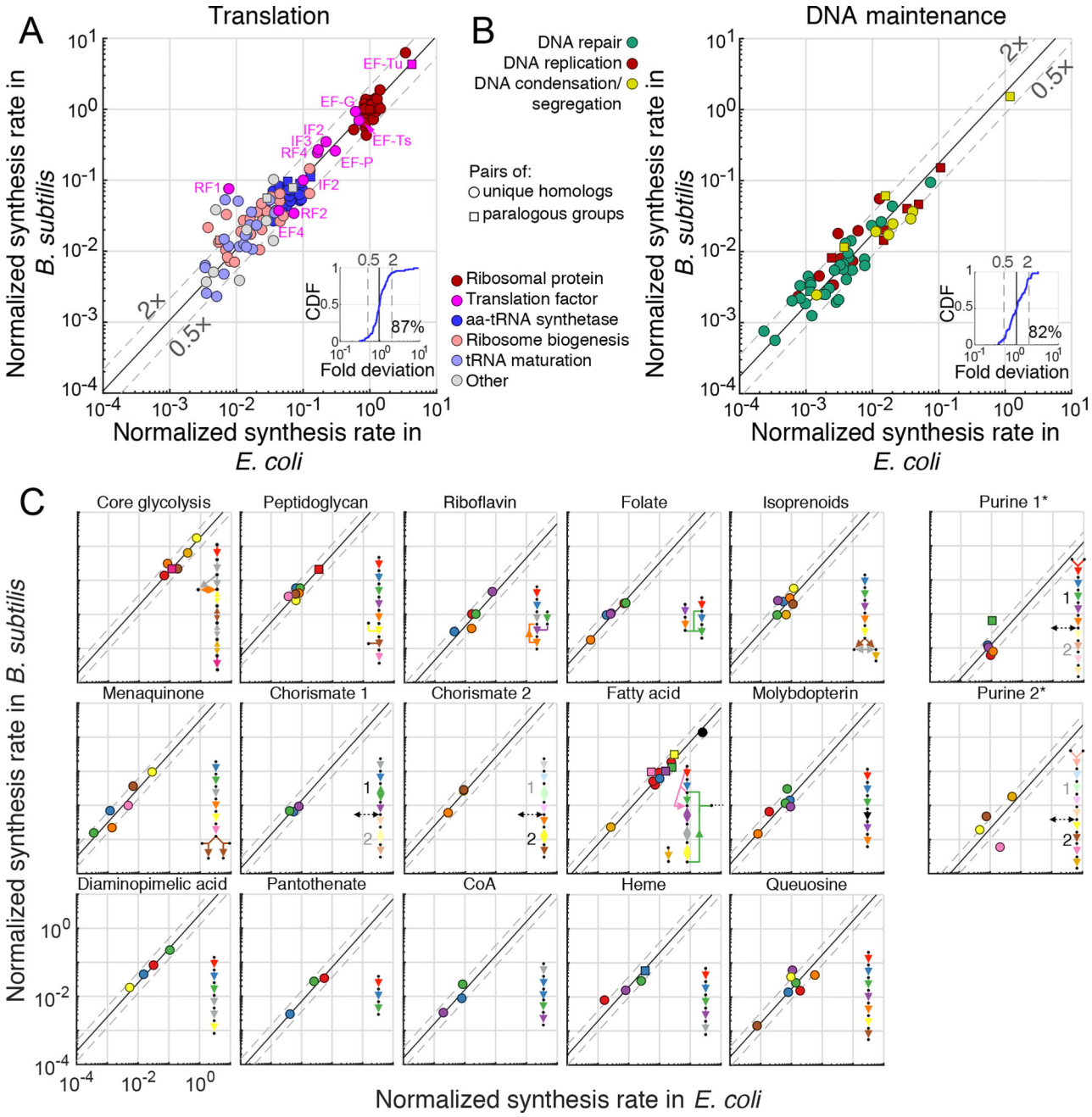


Fig. 1. Conservation of protein expression stoichiometry for ancient pathways in bacteria

(**A, B**) Synthesis rates for proteins involved in translation (A) and DNA maintenance (B) plotted for *B. subtilis* and *E. coli*. Each dot represents a pair of either homologous proteins (circle) or paralogous groups (square, aggregated synthesis rates). Synthesis rates are normalized to the median of ribosomal proteins. Translation factors are highlighted in (A). The black line is a linear fit with a slope of 1 through logarithmically transformed synthesis rates. Dashed lines indicate twofold deviation from the fit. Insets in (A) and (B) show cumulative distribution functions (CDF) of fold-deviation from the regression line. (**C**) Comparison of synthesis rates for curated metabolic pathways. Each dot is an enzyme,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

color-coded by the pathway schemes shown in insets. Reactions performed by non-homologous enzymes are in grey. See Table S2 for the list of enzymes and statistical testing for significance. See Fig. S1 for detailed pathway schemes. *acpP* (acyl carrier protein, black circle) is included in the fatty acid category though not formally an enzyme. Chorismate and purine biosynthesis pathways are each split in two due to intervening fluxes (dashed arrows). Asterisk indicates pathways with non-conserved stoichiometry. The intercept of linear fit indicates differential expression of pathways relative to ribosomal proteins. See also Figure S1 and Table S2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

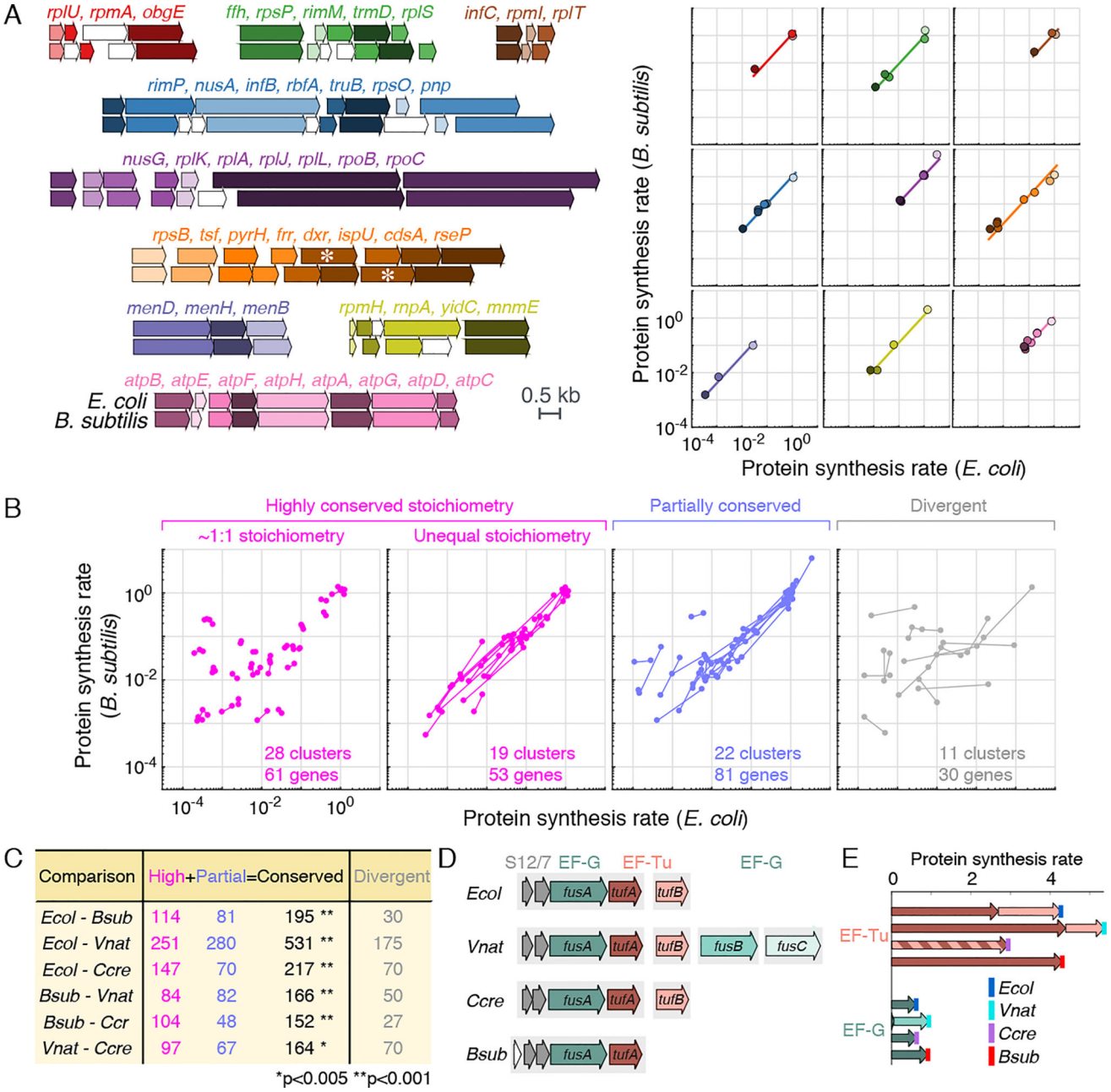


Fig. 2. Differential protein expression within ancient gene clusters is quantitatively conserved

(**A**) Examples of conserved gene clusters between *E. coli* (top) and *B. subtilis* (bottom). Homologous pairs (gene names from *E. coli*) are color-coded for plots on the right. Intervening non-conserved genes are not colored. White asterisk (*) highlights *dxr* whose order within the cluster is shifted. Panels on the right show synthesis rates for proteins in the conserved clusters, plotted for *B. subtilis* and *E. coli*. (**B**) Global analysis of expression stoichiometry for conserved gene clusters between *B. subtilis* and *E. coli*. Each cluster is classified by highly conserved (magenta), partially conserved (blue), and divergent (grey) protein expression stoichiometry. Within the group of highly conserved stoichiometry, clusters are further divided by having equal or unequal synthesis rates (STAR Methods).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Genes (dots) that belong to the same cluster are connected by lines. (**C**) Pair-wise comparison of in-cluster stoichiometry across different bacterial species. The number of genes in each category is listed. Statistical significance for the fraction of clusters with conserved stoichiometry is listed (STAR Methods) *Ecol: E. coli; Bsub: B. subtilis; Vnat: V. natriegens; Ccre: C. crescentus*. (**D**) Gene copy number variation for EF-Tu and EF-G. Paralogous copies outside the conserved S12 gene cluster are labeled as *fusB*, *fusC* (for EF-G) and *tufB* (for EF-Tu). (**E**) Total protein synthesis rates for EF-Tu and EF-G in each species. The contribution of each gene locus is indicated by arrows colored according to (D). The nucleotide sequences for *tufA* and *tufB* in *C. crescentus* are 100% identical, and the respective synthesis rates cannot be distinguished. For (A), (B), and (E), protein synthesis rates are normalized as in Fig. 1. See also Figure S4 and Table S2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

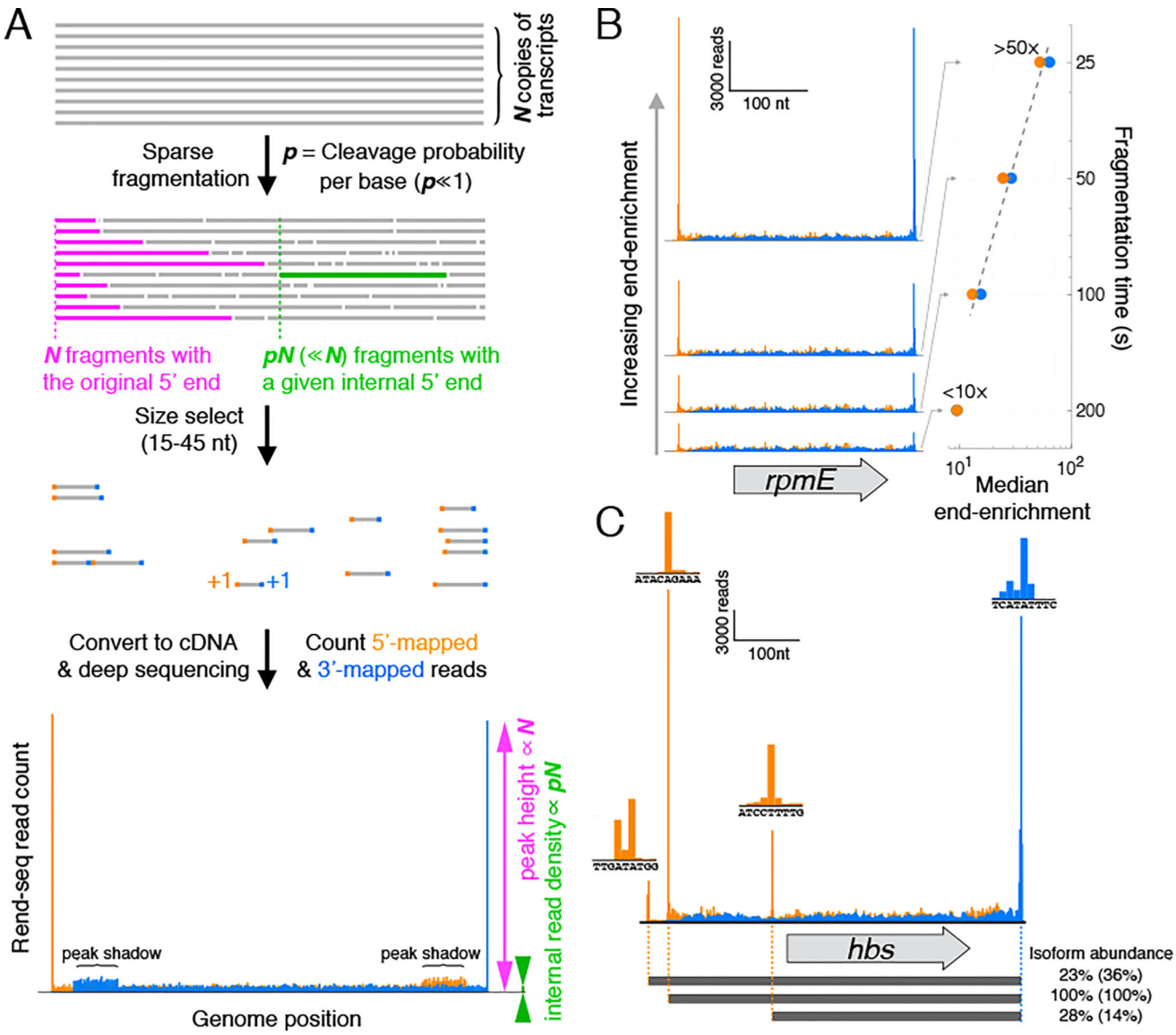


Fig. 3. Rend-seq defines and quantifies mRNA isoforms with single-nucleotide resolution

(**A**) Schematic of end-enrichment strategy. *N* molecules of mRNA are randomly cleaved with a small probability per base ($p \ll 1$). Fragmented RNA is selected for short sizes, converted to a cDNA library, and deep-sequenced. The 5'-mapped (orange) and 3'-mapped (blue) read counts are then plotted separately, revealing peaks at the ends of transcripts and a largely constant read density across the transcript body (simulated data). Peak shadows are shown here but computationally removed for data visualization (see STAR Methods). (**B**) Example of Rend-seq data showing increased end-enrichment with decreased fragmentation. The *rpmE* gene in *B. subtilis* is shown for Rend-seq libraries generated with different amount of fragmentation time *t*. Quantitation in the right panel shows that the median end-enrichment across the transcriptome scales as *1/t* (dashed line). (**C**) Example of Rend-seq data showing multiple mRNA isoforms with alternative 5' ends (*B. subtilis*). Relative isoform abundance can be estimated both by read density between peaks and by peak height (parenthesis). Zoomed-in views illustrate peak width. See also Figure S2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

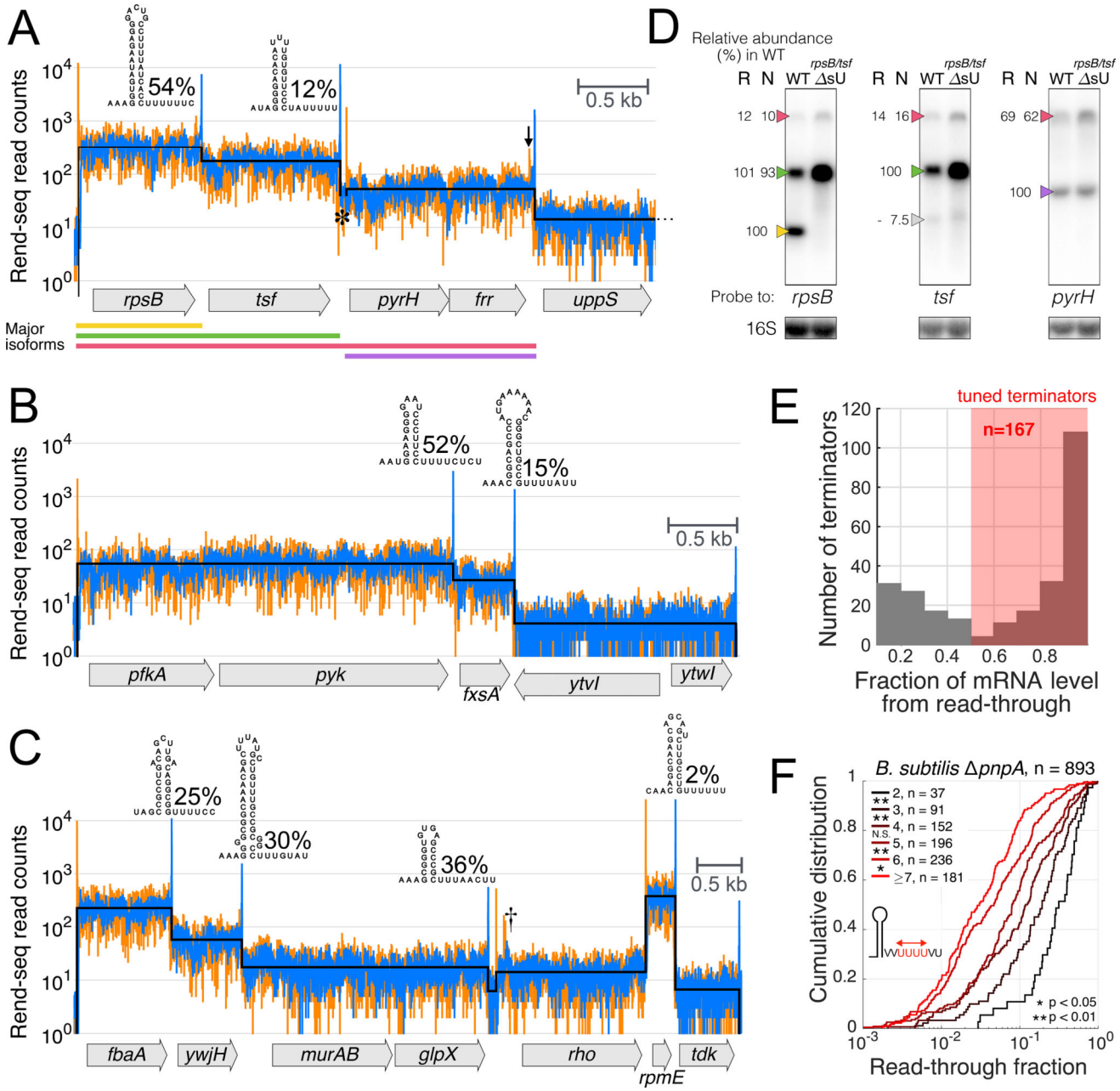


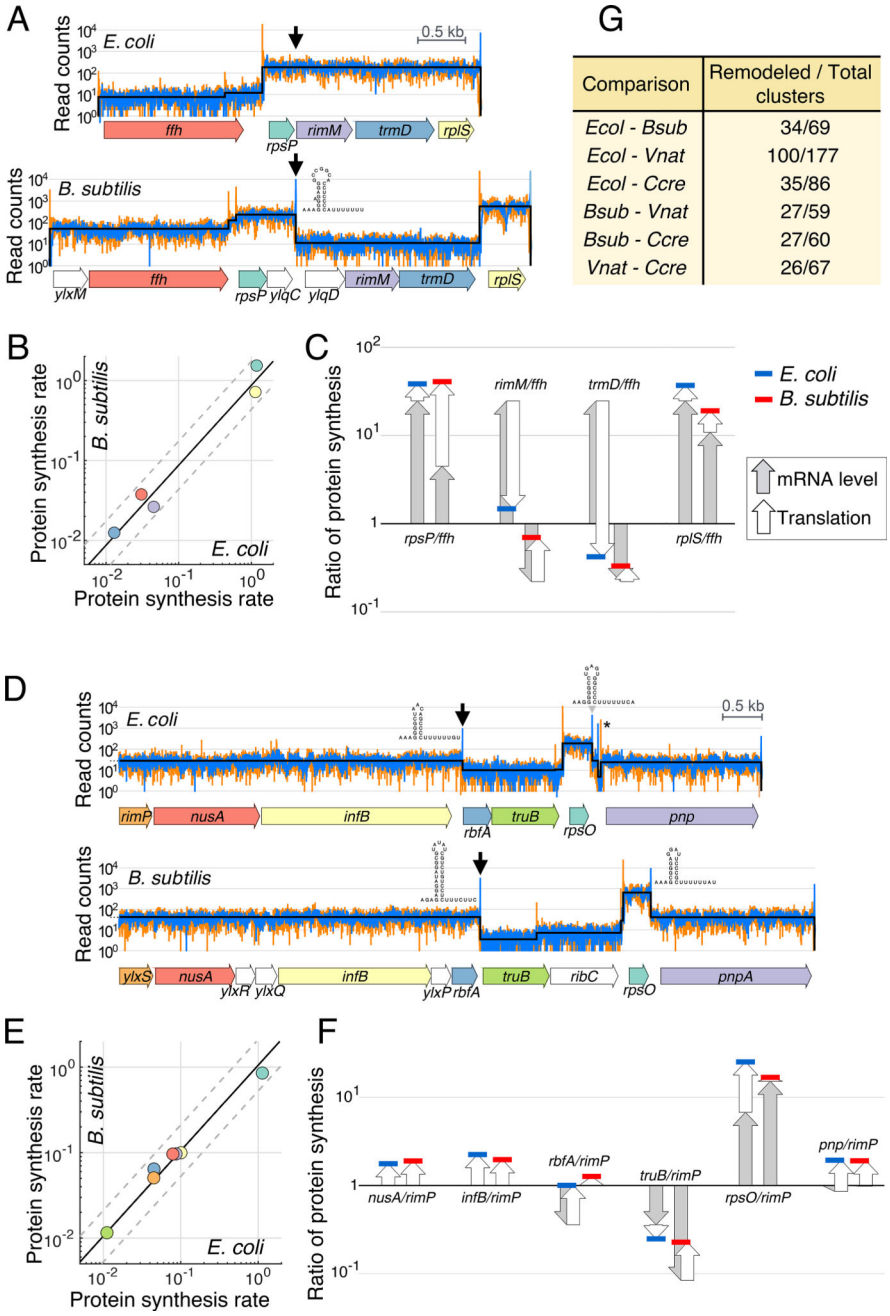
Fig. 4. Rend-seq reveals widespread usage of tuned transcription terminators setting differential expression

(**A-C**) Examples of gene clusters with intervening partial terminators in *B. subtilis*. 5'- and 3'-mapped read counts, shown in logarithmic scale, are plotted in orange and blue, respectively. Black lines indicate average read counts between peaks. Rend-seq data have peak shadows removed for clarity, see Methods. See Table S3 for a comprehensive list of intergenic tuned terminators. Terminator sequences and the corresponding leakiness (fraction of read density remaining past terminator) are shown above each internal 3' peak. Asterisk points to a short intergenic region between a promoter and an upstream tuned terminator, whose leakiness is estimated based on the peak height (Fig. S3, STAR Methods).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Arrow points to a nested promoter immediately upstream of the *frr* terminator. Dagger points to the regulatory region upstream of *rho* (Ingham et al., 1999). (**D**) Northern blotting against different regions of the *rpsB* gene cluster in the wild type (WT) or a strain with perturbed *rpsB/tsf* terminator (ΔsU). Arrows point to different isoforms as indicated in (A). Grey arrow points to an unknown isoform. Relative abundance predicted by Rend-seq is shown under 'R,' and by Northern blotting under 'N.' 16S rRNA is used as a loading control. See Fig. S3 for Northern blots for other gene clusters. **(E)** Distribution of contribution of downstream mRNA level by terminator read-through. All identified terminators contributing to more than 10% of the transcription of downstream genes in *B. subtilis* are included. 167 tuned terminators contribute to more than 50% (red shading) of downstream gene expression for growth in LB. (**F**) Cumulative distribution of read-through fraction for terminators with different U-tract lengths, defined as the number of consecutive U's within 8 nt upstream of the 3' end. Data shown for *B. subtilis* deleted with *pnpA*, which encodes the major 3'-to-5' exoribonuclease (Oussenko et al., 2005). See Fig. S3 for data for wildtype *B. subtilis* and other species. Significance (two sample t-test) was computed between the log-transformed read-through fractions of consecutive U-tract lengths. See also Figure S2, S3 and Table S3.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Comparison	Remodeled / Total clusters
Ecol - Bsub	34/69
Ecol - Vnat	100/177
Ecol - Ccre	35/86
Bsub - Vnat	27/59
Bsub - Ccre	27/60
Vnat - Ccre	26/67

Fig. 5. Bacterial gene clusters have divergent mRNA architecture but conserved protein stoichiometry

(**A**) Rend-seq data for the conserved gene cluster *ffh-rpsP-rimM-trmD-rplS* showing divergent transcript architecture between *E. coli* (top) and *B. subtilis* (bottom). Data are displayed as in Fig. 4(A-C). Black arrows point to the tuned terminator in *B. subtilis*, which is absent in *E. coli*. (**B**) Conservation of synthesis rates for the corresponding proteins, with coloring based on (A). Black and dashed lines are as described in Fig. 1. (**C**) Contribution of mRNA level (from Rend-seq) and translation efficiency (from ribosome profiling and Rend-seq) to conserved expression stoichiometry. Rates of protein synthesis relative to *ffh* is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

plotted for *E. coli* (blue) and *B. subtilis* (red). The contributions of differential mRNA levels and translation efficiency are shown by grey and white arrows, respectively. (**D** to **F**) Same as (A to C), but for the gene cluster containing *rbfA*. Black arrows indicate the positions of tuned terminator either upstream (*E. coli*) or downstream (*B. subtilis*) of *rbfA*. The asterisk (*) between *E. coli*'s *rpsO* and *pnp* in (D) highlights a known processing site by RNase III (see Mendeley Data). **(G)** Genome-wide comparison of transcript architecture for the gene clusters with conserved protein expression stoichiometry. Species names abbreviated as in Fig. 2C. See STAR Methods for definition of remodeled clusters. See also Figure S4 and Data S1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

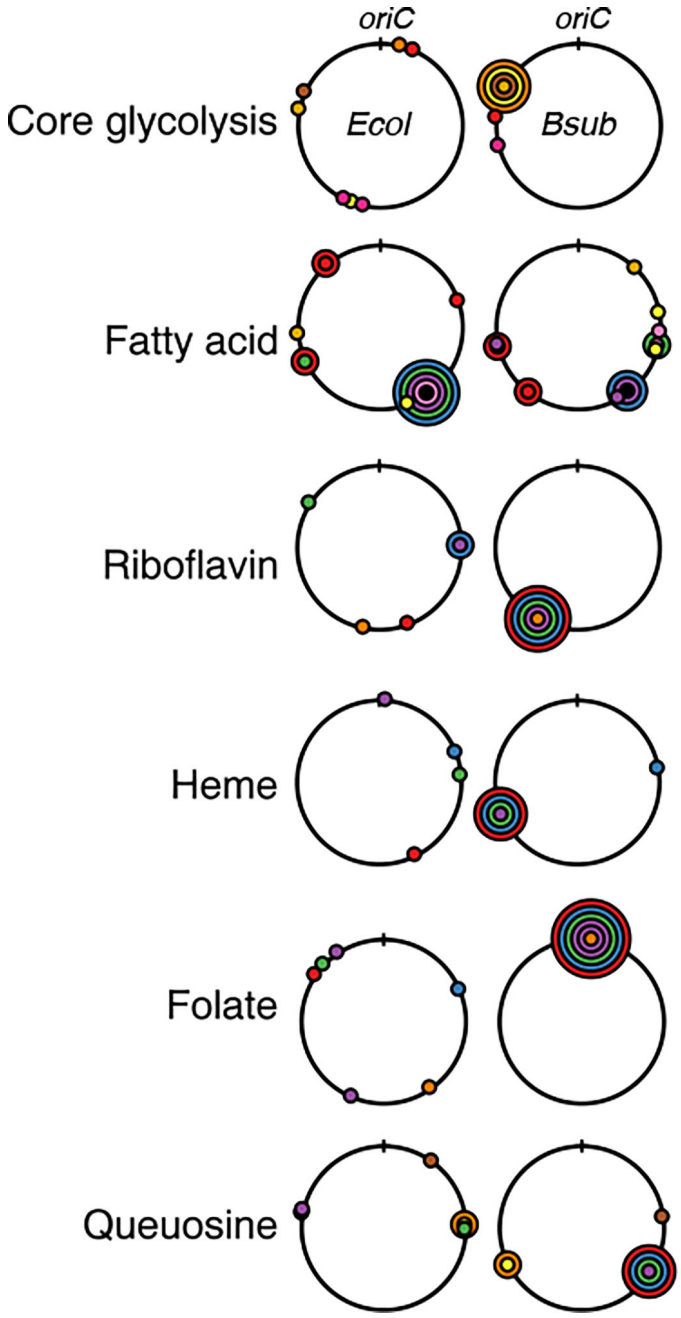


Fig. 6. Dispersion of gene clusters is compensated to maintain conserved protein stoichiometry

Divergent operon organization for a subset of pathways shown in Fig. 1. For each pathway, gene positions are highlighted by colored circles on the circular chromosome diagram (*oriC*: origin of replication). Color coding is the same as Fig. 1. Genes in the same operon are represented as concentric circles. For example, folate biosynthetic genes are all clustered in *B. subtilis*, but are scattered around the chromosome in *E. coli*. *Ecol: E. coli; Bsub: B. subtilis.* See also Figure S5.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

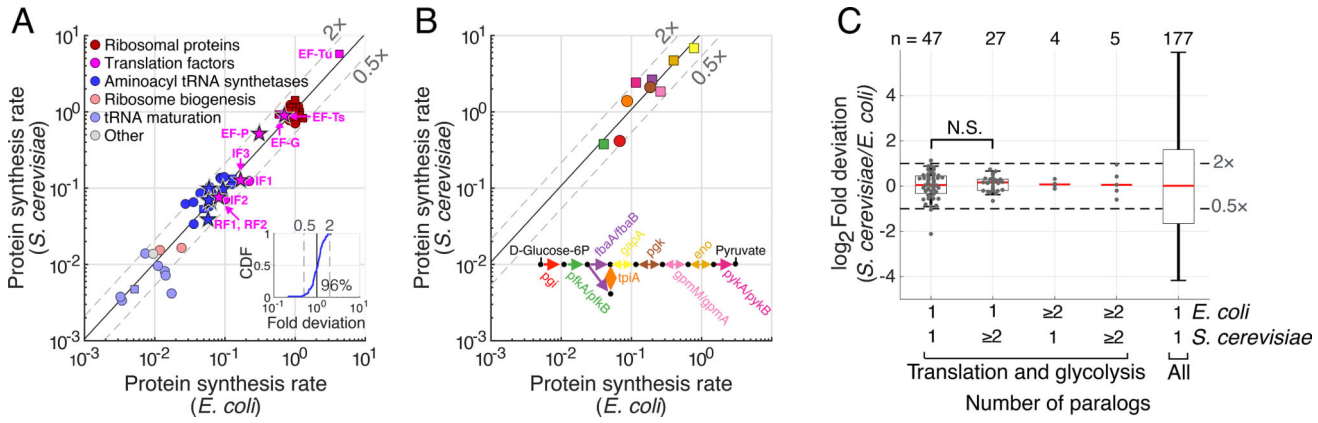


Fig. 7. Conservation of pathway-specific protein stoichiometry across the prokaryote/eukaryote divide

(**A, B**) Synthesis rates (normalized as in Fig. 1) for proteins involved in cytosolic translation (A) and glycolysis (B) are plotted for *S. cerevisiae* and *E. coli*. Synthesis rates in *S. cerevisiae* were estimated based on ribosome profiling data reported by (Weinberg et al., 2016) (STAR Methods). Functional analogs (proteins with similar function but with pairwise BLASTP score<45) are shown as stars. Plotting convention as in Fig. 1. Inset in (A) shows cumulative distribution function (CDF) of fold-deviation from the regression line. Inset in (B) shows the pathway diagram for core glycolysis. (**C**) Distribution of fold-deviation from pathway-specific regression lines for proteins with different genes dosage. Proteins involved in translation and glycolysis are grouped by the numbers of paralogous genes in *E. coli* and *S. cerevisiae*. Medians are indicated by red lines, and whiskers correspond to the 5th and 95th percentiles. The fold-deviation for singleton and duplicated genes in *S. cerevisiae* is tightly distributed and not statistically different (p = 0.53 two-sample t-test, p = 0.35 two-sample Kolmogorov-Smirnov test). As a comparison, the ratio of synthesis rates for all one-to-one homologs across the two genomes (excluding mitochondrial proteins) is shown with a much wider distribution. See also Figure S6.