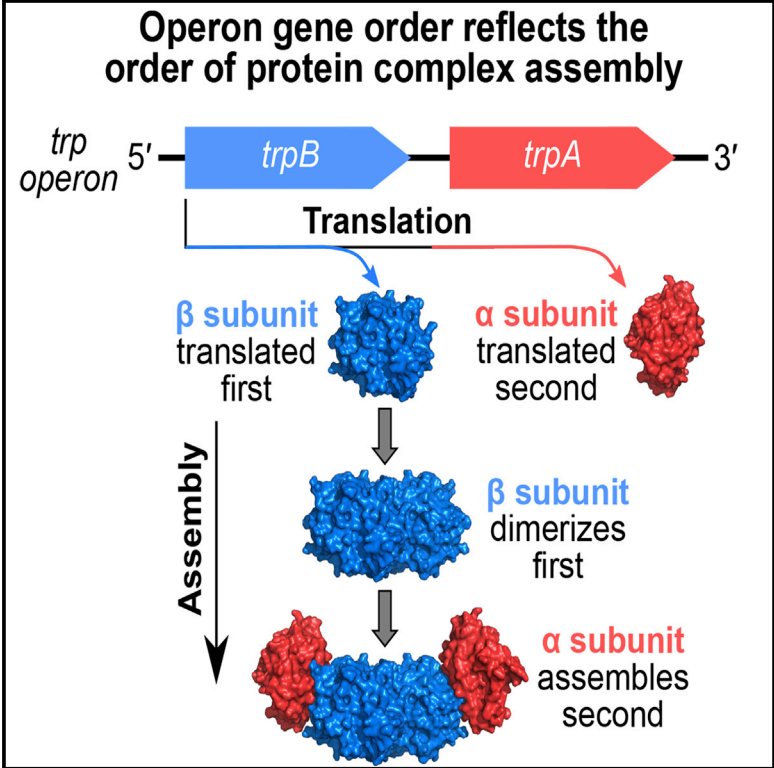


Operon Gene Order Is Optimized for Ordered Protein Complex Assembly

Graphical Abstract



Authors

Jonathan N. Wells,
L. Therese Bergendahl, Joseph A. Marsh

Correspondence

joseph.marsh@igmm.ed.ac.uk

In Brief

Many prokaryotic protein complexes are operon-encoded, so that subunits of the same complex will be translated from the same mRNA. Wells et al. show that the order in which genes are arranged in operons tends to be optimized for the order in which protein complex subunits assemble.

Highlights

- Operon-encoded subunits tend to be encoded by neighboring genes and form large interfaces
- Operon gene order is often optimized for the order of protein complex assembly
- Exceptions are mostly highly expressed proteins for which assembly is less stochastic

Operon Gene Order Is Optimized for Ordered Protein Complex Assembly

Jonathan N. Wells,¹ L. Therese Bergendahl,¹ and Joseph A. Marsh^{1,*}

¹MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom

*Correspondence: joseph.marsh@igmm.ed.ac.uk

<http://dx.doi.org/10.1016/j.celrep.2015.12.085>

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

SUMMARY

The assembly of heteromeric protein complexes is an inherently stochastic process in which multiple genes are expressed separately into proteins, which must then somehow find each other within the cell. Here, we considered one of the ways by which prokaryotic organisms have attempted to maximize the efficiency of protein complex assembly: the organization of subunit-encoding genes into operons. Using structure-based assembly predictions, we show that operon gene order has been optimized to match the order in which protein subunits assemble. Exceptions to this are almost entirely highly expressed proteins for which assembly is less stochastic and for which precisely ordered translation offers less benefit. Overall, these results show that ordered protein complex assembly pathways are of significant biological importance and represent a major evolutionary constraint on operon gene organization.

INTRODUCTION

The assembly of proteins into complexes is integral to a wide range of biological processes. Although we now have extensive knowledge of the diverse quaternary structures formed by protein complexes (Goodsell and Olson, 2000; Janin et al., 2008; Marsh and Teichmann, 2015; Ahnert et al., 2015), much less is known about how they assemble and how assembly is regulated. In recent years, advances in electrospray mass spectrometry techniques have provided major new insights into in vitro assembly, allowing the assembly and disassembly pathways of protein complexes with diverse quaternary structure topologies to be elucidated in detail (Hernández and Robinson, 2007). In homomers, formed from the self-assembly of a single type of polypeptide chain, experimentally identified assembly intermediates often correspond to putative evolutionary precursors, so that the evolutionary history of a complex is reflected in its assembly pathway (Levy et al., 2008). Heteromers, formed from multiple distinct subunits, also tend to assemble and disassemble via ordered pathways that have a strong tendency to be evolutionarily conserved (Marsh et al., 2013). Although these experiments can be time-consuming, ordered assembly path-

ways can usually be predicted with very good accuracy from the known three-dimensional structure of a complex (Levy et al., 2008; Marsh et al., 2013). Given the many thousands of protein complex structures that are now available, this enables the study of assembly on a larger scale using computationally predicted assembly pathways.

Within the cell, assembly is much more complex and stochastic than in vitro, particularly in heteromers where multiple protein-coding genes must first be transcribed to mRNA and translated into protein, and those proteins must then find each other and assemble. Assembly is especially difficult for lowly expressed proteins, for which the stochastic variations in relative subunit concentrations are greater and the probability of interaction is lower (Kovács et al., 2009; Swain et al., 2002). How do cells cope with this? Does assembly within the cell follow similar ordered pathways as those observed in vitro and predicted computationally? Where does assembly occur within the cell? Has the regulation of gene expression been optimized for protein complex assembly order, as appears to be the case for the large multi-subunit bacterial flagella (Kalir et al., 2001)? Here we were able to address all of these questions by considering the relationship between protein complex assembly and gene organization in prokaryotic operons.

RESULTS

Operon-Encoding of Protein Complexes Is Likely to Enhance the Efficiency of Assembly

Many operons contain genes encoding different subunits of the same protein complex (Dandekar et al., 1998; Mushegian and Koonin, 1996) that can then be transcribed onto the same polycistronic mRNA. We first searched for heteromeric protein complexes of known structure from all prokaryotic organisms where at least two of the subunits are encoded by different genes from the same operon. In total, we identified 368 non-redundant pairs of subunits from the same heteromer encoded by different genes from the same operon (Figure 1A, left) from 70 different bacterial and archaeal species. This compares to 711 pairs encoded by different transcriptional units (*i.e.* translated from different mRNAs) from the same species (Figure 1A, right).

It has been suggested previously that a major advantage of operon-encoded complexes is their more efficient assembly because of smaller stochastic fluctuations in relative concentration than would occur if separate transcription steps were required for each subunit (Shieh et al., 2015; Sneppen et al.,

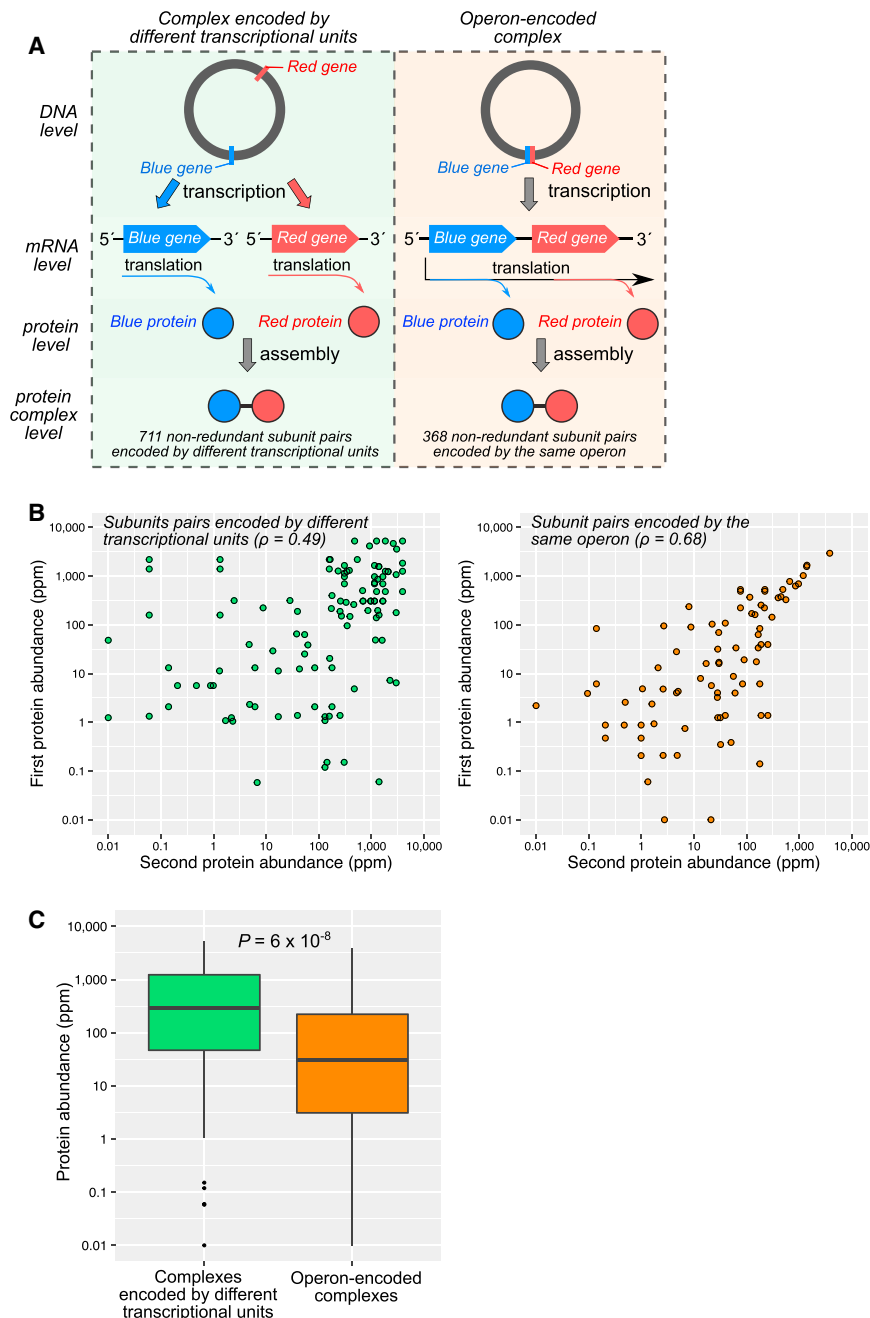


Figure 1. Operon Encoding of Protein Complex Subunits Enhances the Efficiency of Assembly

(A) Comparison of assembly for heterodimers where different subunits are encoded by different transcriptional units and where genes encoding both subunits are present on the same operon. (B) Correlation (Spearman's ρ) between abundance measurements from subunit pairs encoded by different transcriptional units or by the same operon. The correlation for subunit pairs encoded by the same operon is significantly higher than for those encoded by different transcriptional units ($p = 0.002$), as calculated by randomly shuffling the pairs between two groups of the same size 10^5 times. (C) Comparison of protein abundance measurements for subunits from operon-encoded complexes versus other subunits from complexes encoded by different transcriptional units. Boxes represent quartile distributions, and whiskers extend up to $1.5\times$ the interquartile range. The p value was calculated with Wilcoxon rank-sum test.

Figure S1 shows these comparisons using protein abundance measurements combined from multiple organisms and with *E. coli* protein synthesis rates.

et al., 2002). This is supported by a highly significant ($p = 6 \times 10^{-8}$) tendency for operon-encoded subunits to be lower in abundance than subunits from complexes encoded by different transcriptional units (Figure 1C). Although there is an overlap between the groups, this suggests that lowly expressed genes encoding interacting subunits may have experienced stronger evolutionary pressure to be located on the same operon because of their more stochastic assembly. Alternatively, because of the efficiency of their assembly, operon-encoded subunits may only need to be expressed at lower levels.

Operon-Encoded Subunits Tend to Be Encoded by Neighboring Genes and Form Large Interfaces

In addition to simply having genes encoding interacting subunits on the same

operon, another way to enhance the efficiency of assembly would be to position the genes close together. If two genes encoding interacting subunits are close, then the newly translated subunits will also be close and more likely to encounter each other than if the two genes are farther apart (Figure 2A). In fact, the tendency for adjacent genes to code for interacting proteins has long been recognized (Dandekar et al., 1998; Mushegian and Koonin, 1996).

In Figure 2B, we plot the number of subunit pairs from the same complex by the distance between their genes within the operon. Strikingly, we see that 220 of 368 subunit pairs (59.8%) are

operon, another way to enhance the efficiency of assembly would be to position the genes close together. If two genes encoding interacting subunits are close, then the newly translated subunits will also be close and more likely to encounter each other than if the two genes are farther apart (Figure 2A). In fact, the tendency for adjacent genes to code for interacting proteins has long been recognized (Dandekar et al., 1998; Mushegian and Koonin, 1996).

In Figure 2B, we plot the number of subunit pairs from the same complex by the distance between their genes within the operon. Strikingly, we see that 220 of 368 subunit pairs (59.8%) are

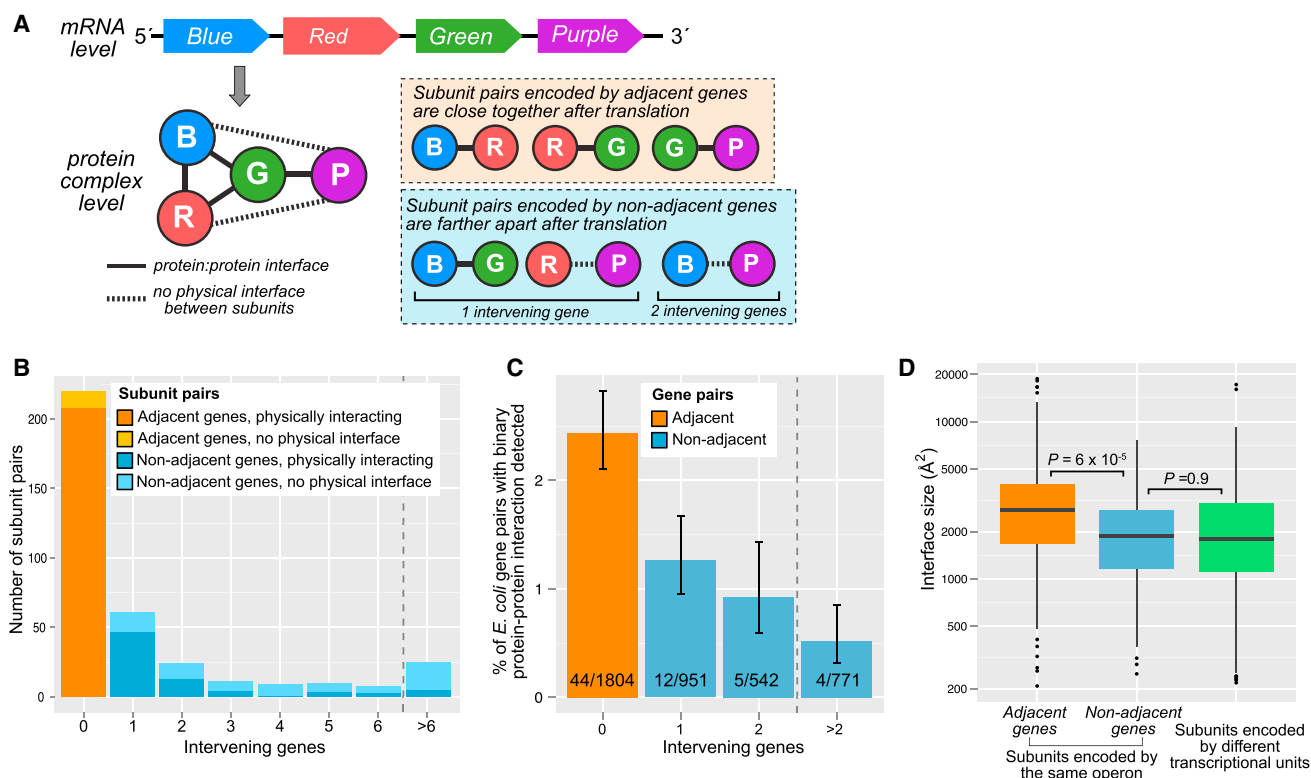


Figure 2. Genes Encoding Interacting Subunits of the Same Complex Tend to Be Close Together on an Operon

(A) Illustration of how operon structure can be related to quaternary structure with a hypothetical four-subunit heteromer. Pairs of subunits from the same complex can be encoded by genes that are adjacent on an operon or farther apart.

(B) Number of subunit pairs encoded by the same operon, grouped by the distance between their encoding genes. Subunit pairs are also divided into those that interact physically, which we define as forming an interface of $>200 \text{ \AA}$, and those that do not interact physically.

(C) Percentage of pairs of *E. coli* genes from the same operon for which a binary yeast two-hybrid interaction could be detected. Error bars represent 68% Wilson binomial confidence intervals.

(D) Distribution of interface sizes formed between physically interacting subunit pairs encoded by adjacent or non-adjacent genes on the same operon or between subunits encoded by different transcriptional units. Boxes represent quartile distributions, and whiskers extend up to $1.5 \times$ the interquartile range. The p values were calculated with Wilcoxon rank-sum test.

encoded by adjacent genes. Furthermore, because not all subunit pairs from the same complex physically interact with each other (e.g., blue-purple and red-purple in Figure 2A), we note that the tendency to form a physical intersubunit interface within the complex is much higher between the adjacent (208 of 220) compared with non-adjacent (77 of 148) pairs ($p = 5 \times 10^{-22}$, Fisher's exact test). Finally, this is supported further through analysis of a large set of *E. coli* binary protein-protein interactions (Rajagopala et al., 2014) where we confirmed that proteins encoded by adjacent genes are much more likely to interact (Figure 2C). Importantly, we show in Figure S2 that the tendency for interacting proteins to be close within an operon is highly significant compared with a null model in which gene order is randomized.

Figure 2D compares the sizes of interfaces formed between subunits encoded by adjacent genes, subunits encoded by non-adjacent genes from the same operon, and subunits encoded by different transcriptional units. We observe a highly significant tendency for adjacent subunits to be larger, although the interface size distribution is very broad and there is considerable overlap between the groups. This is especially interesting when

considering that larger interfaces within a complex will usually assemble earlier than smaller interfaces (Levy et al., 2008; Marsh et al., 2013). This provides further evidence that operon structure appears to have been evolutionarily optimized for protein complex formation. Even when we consider only physically interacting proteins, those that form larger interfaces and are, therefore, likely to assemble earlier are much more likely to be encoded by adjacent genes.

The above observation could potentially have implications for our previous finding that evolutionary gene fusion events tend to conserve existing assembly pathways (Marsh et al., 2013) because fusion often occurs between adjacent genes. However, we show in Figure S3 that, even if only subunit pairs encoded by adjacent genes are considered, there still appears to be evolutionary selection for assembly-conserving fusions.

Operon Gene Order Is Optimized for the Order of Protein Complex Assembly

The above results suggest that operon-encoded subunits will often be synthesized very close to each other within the cell.

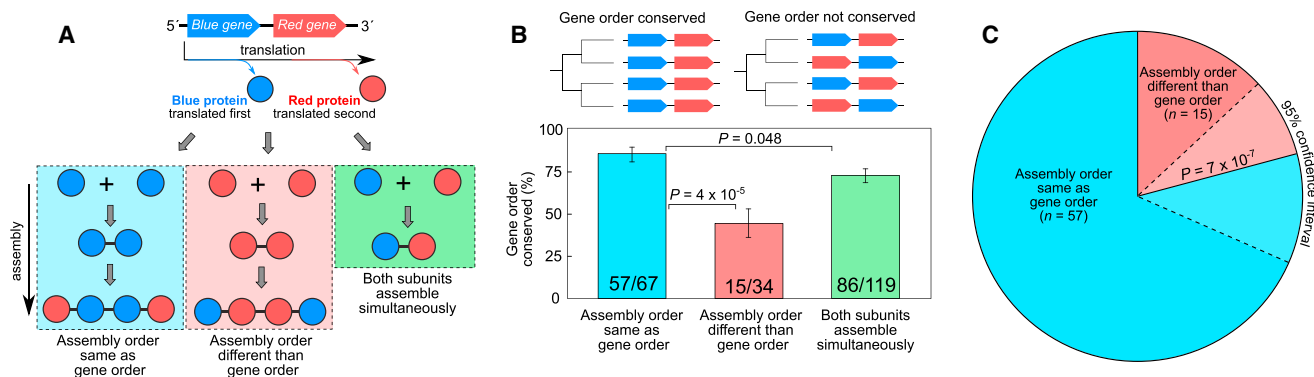


Figure 3. Operon Gene Order Reflects the Order of Protein Complex Assembly

(A) Illustration of the three possible relationships between gene pair order and subunit assembly order.

(B) Evolutionary conservation in pairs of adjacent genes encoding subunits of the same complex. The p values were calculated with Fisher's exact test. Error bars represent 68% Wilson binomial confidence intervals.

(C) When considering adjacent gene pairs with evolutionarily conserved gene order that encode different subunits of the same protein complex, the predicted assembly order is the same as the gene order in 57 of 72 cases. The p value was calculated with a binomial test.

However, there is also a temporal component to this in that upstream genes will tend to be translated before downstream genes. This is first due to coupled transcription and translation, where the upstream gene that is transcribed first will also be translated first (Gowrishankar and Harinarayanan, 2004), and second to translational coupling, in which translating ribosomes can continue on to downstream genes (Oppenheim and Yanofsky, 1980). Therefore, if genes are arranged so that the gene order matches the order of subunit assembly, then the newly translated subunits will be more likely to interact quickly.

We illustrate this in Figure 3A with the example of a hypothetical operon containing two adjacent genes, *blue* and *red*. If these genes encode different subunits of the same complex, then there are three possible relationships between gene order and assembly order. First, the assembly order could be the same as the gene order if the blue subunit that is translated first also assembles first. Second, the assembly order could be different than the gene order if the blue subunit assembles last. Finally, both subunits could assemble simultaneously, as would be the case for a simple heterodimer where the first step of assembly is the heteromeric interaction between different subunits.

Using our previous observation that assembly pathways can be predicted using interface sizes from three-dimensional structures of protein complexes (Marsh et al., 2013), here we predicted the assembly pathways for all operon-encoded heteromers in our dataset and classified each of the 220 adjacent gene pairs into one of these three groups. We then considered the tendency for gene order to be evolutionarily conserved in each group (Figure 3B). Interestingly, the evolutionary conservation of gene order is significantly higher in cases where it is the same as the predicted assembly order. This suggests that the evolutionary constraint on gene order is much stronger when it is optimized for assembly.

Next, we consider 72 gene pairs where gene order is evolutionarily conserved and where one subunit is predicted to assemble before the other. Figure 3C illustrates the striking correspondence between gene order and assembly order, with 57

pairs (79.2%) having the same assembly order as gene order ($p = 7 \times 10^{-7}$, binomial test). In contrast, when the gene order is not evolutionarily conserved, only 10 of 29 gene pairs show correspondence between gene order and assembly order. Therefore, selection for ordered protein complex assembly appears to be a major evolutionary determinant of operon gene order.

We can also consider the relationship between gene order and assembly order for non-adjacent genes. Although the dataset is smaller, the relationship between gene order and assembly order appears to get weaker between genes that are more distant (Figure S4). This is likely due to weaker spatial and temporal coupling between non-adjacent genes that are translated farther apart from each other, as evidenced by the fact that subunits encoded by non-adjacent genes are much less likely to physically interact with each other (Figure 2B). Interestingly, the relationship between gene order and assembly order is stronger for proteins that interact physically, particularly those that form large interfaces. Similarly, subunit pairs encoded by adjacent genes where gene order and assembly order are the same tend to have significantly larger interfaces (Figure S4).

A possible alternative explanation for the correspondence between gene order and assembly order could be if earlier-assembling subunits need to be expressed at higher levels. Specifically, there is evidence of a linear relationship between expression levels and the proximity of genes to the start of operons (Lim et al., 2011; Nishizaki et al., 2007). This is weakly supported in the dataset used here, with proteins encoded by upstream genes showing a slight but not significant tendency to be more abundant (Table S1). Importantly, we find that protein expression levels show essentially no relationship with assembly order and that gene order is a significantly better predictor of assembly order (Figure S5).

Operon Gene Order Is Most Important for the Assembly of Lowly Expressed Proteins

Despite the strong correspondence between protein complex assembly and operon organization, there is still discordance

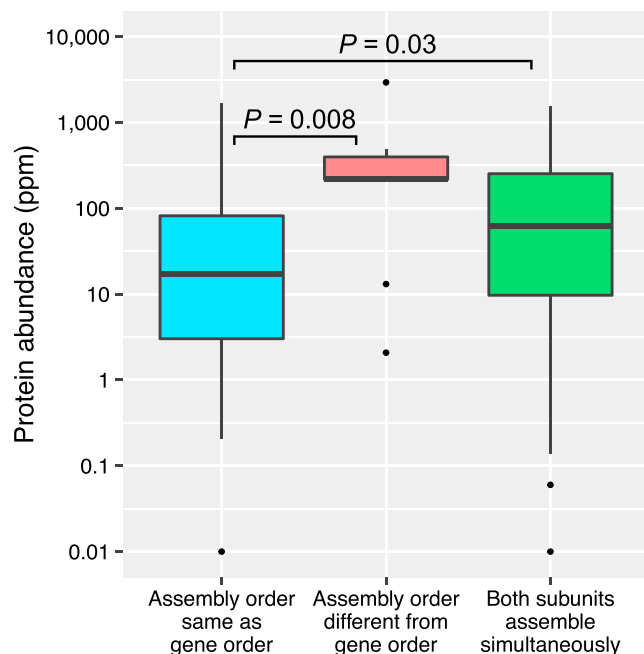


Figure 4. Cases Where Evolutionarily Conserved Gene Order Does Not Follow Assembly Order Tend to be Highly Expressed

Boxes represent quartile distributions of protein abundance measurements, and whiskers extend up to 1.5× the interquartile range. The p values were calculated with a Wilcoxon rank-sum test.

Figure S7 shows these comparisons using protein abundance measurements combined from multiple organisms and with *E. coli* protein synthesis rates.

between gene order and assembly order in >20% of cases where gene order is evolutionarily conserved. This suggests that there must be other factors besides assembly order that influence gene order conservation. For example, the operon order of enzyme genes is known to correlate with metabolic pathway order (Kovács et al., 2009; Zaslaver et al., 2006), although this seems unlikely to explain gene order in operon-encoded complexes. A search for gene ontology terms (Huntley et al., 2015) enriched in subunit pairs where gene order is either the same or different than assembly order revealed little that could account for the results observed here (Figure S6). Furthermore, if gene position can affect expression levels, as mentioned above, then there may be some evolutionary pressure to conserve gene order; for example, to not disrupt the relative subunit stoichiometry (Marsh et al., 2015).

The fact that operon gene order closely follows assembly order suggests that assembly must occur very shortly after protein synthesis because the more time newly synthesized subunits have to diffuse before assembly the less the order of gene expression should matter. Building on this, we hypothesize that the relationship between operon order and assembly order should be stronger for lowly expressed proteins. If they do not assemble quickly, diffusion will reduce the probability of two low-concentration subunits encountering each other. In contrast, the chance of interaction between highly expressed, abundant proteins will be greater, and so there is less need for assembly to occur close to the site of protein synthesis.

In Figure 4, we plot the distributions of intracellular protein abundance measurements for subunits where conserved gene order follows assembly and for those where it does not. Those proteins where assembly order is the same as gene order tend to be much lower in abundance ($p = 0.008$, Wilcoxon test). Therefore, it appears that the correspondence between gene order and assembly order can mostly be attributed to lowly expressed proteins for which assembly is more stochastic. Interestingly, subunits where both assemble simultaneously are intermediate in abundance, consistent with the fact that gene order should show no correspondence with assembly in these cases.

DISCUSSION

Overall, a number of important conclusions can be drawn from these results. First, protein complex assembly within the cell appears to often follow the same ordered pathways that can be characterized experimentally and predicted computationally, at least in the case of operon-encoded complexes. Although there will certainly be some exceptions, particularly in cases where assembly chaperones are involved or subunits are translated in different parts of the cell, these results strongly support the physiological relevance of using *in vitro* or computational methods to study assembly.

Second, the remarkable correspondence between predicted assembly order and gene order further validates the utility of structure-based assembly predictions. Given the huge number of protein complex structures now known, this opens the door to future large-scale analyses of protein assembly pathways and their regulation, evolution, and role in biological function and disease.

This work also tells us something about where assembly occurs within the cell. For the low-abundance, operon-encoded complexes studied here, assembly must occur very close to the site of translation for gene order to have such a significant effect. In some cases, assembly may even occur co-translationally, involving at least one nascent chain still in the process of being translated (Duncan and Mata, 2011; Wells et al., 2015), as has been demonstrated recently for the operon-encoded bacterial luciferase complex (Shieh et al., 2015).

Finally, these results strongly support the biological importance of assembly pathways and suggest that co-ordinating both the timing and location of translation is important for maximizing the efficiency of stochastic protein complex assembly. The fact that operon gene order has been optimized for assembly order in many protein complexes suggests that assembly order is often very important and that there is significant benefit from tightly co-ordinating gene expression and protein assembly. Given that eukaryotes do not have operons that allow multiple protein subunits to be translated from the same polycistronic mRNA, it will be interesting to systematically investigate which other mechanisms might be employed to enhance the efficiency of assembly.

EXPERIMENTAL PROCEDURES

Protein Structural Datasets

We started with the full set of prokaryotic X-ray and electron microscopy structures in the PDB on June 12, 2014. We considered all heteromeric pairs of

subunits from the same complex, defined as having at least two different protein chains of ≥ 30 residues each and mapping to different UniProt sequences from a single species. Complexes with known quaternary structure assignment errors (Levy, 2007) were excluded. Very large complexes with >24 subunits were excluded, because we have not shown that the assembly of these can be predicted accurately from their structures. Heteromeric subunit pairs were filtered for redundancy at the level of 50% sequence identity.

Mapping Subunit Pairs to Operons

Operon datasets were downloaded from the DOOR² database (Mao et al., 2014). Relevant datasets were identified based on the species and strain of each gene pair. After converting GI numbers to UniProt accession identifiers in each dataset, the set of gene pairs was mapped to the operon data. Operons encoding both members of a pair were added to a reference dictionary, with the locus and directionality of each gene being used to arrange constituent genes in order of expression. In rare cases where the copy number of a gene within an operon was found to be greater than one, the position of the gene in the operon was taken to be that of the first copy to be encountered, reading in the 5' to 3' direction. The set was then filtered to remove redundant operons (i.e., identical operons from similar strains or species). In total, 368 gene pairs (220 adjacent) were mapped to 192 unique operons, with the remaining 711 pairs being expressed in different transcriptional units. These are provided in Dataset S1. Similarly, we also mapped a set of 2,562 binary protein-protein interactions (IM-22059) (Rajagopala et al., 2014) to the *E. coli* K-12 W3110 operons to calculate the result in Figure 2C (provided in Dataset S2).

To assess whether the gene order of a pair was evolutionary conserved, we used the STRING v9.1 database (Franceschini et al., 2013). For each pair, we manually assessed, using the STRING online interface, whether all occurrences of a given gene pair shared the same gene order within their local evolutionary group as defined in STRING. This is at the level of phylum (e.g. Firmicutes or Euryarchaeota) or class for proteobacteria, with all groups provided in Dataset S1. Gene pairs present across only a very limited evolutionary range (less than three genera) were not considered to be evolutionarily conserved. Gene pairs associated with evolutionary gene fusion events were identified as those sharing $>40\%$ sequence identity with a gene pair with evidence for fusion in STRING, similar to what has been done previously (Marsh et al., 2013).

Abundance Measurements

We mapped all protein complex subunits in our dataset against the sequences of prokaryotic proteins from PaxDB v4.0 (Wang et al., 2015), selecting abundance measurements with $>90\%$ sequence identity to a subunit. The results in Figures 1 and 4 only use abundance measurements from *E. coli*, but the analyses in the Figures S1, S5, and S7 and Table S1 are repeated using combined measurements from all available prokaryotes and also using protein synthesis rates derived from ribosomal profiling (Li et al., 2014).

Prediction of Assembly Pathways

Ordered protein complex assembly pathways were predicted in a manner very similar to what has been done previously (Marsh et al., 2013). First, the complex is considered in terms of its constituent subunits and the sizes of the interfaces that can be formed between any pair of subunits are calculated with AREAIMOL (Winn et al., 2011). Our model assumes that assembly will proceed via formation of the largest possible interface. The process is then repeated by calculating all possible interfaces that could form between subunits and subcomplexes until the full complex is assembled. To define which of a pair of subunits assembles first and which assembles later, we consider the first step of assembly that brings the two subunits together within the same (sub)complex. Whichever subunit was part of a larger subcomplex prior to this step is defined as assembling first. For example, in the blue pathway in Figure 3A, the blue subunit homodimerizes first and then interacts sequentially with the free red subunits, so the blue subunit is defined as assembling first. If, alternatively, the first step of assembly had been a heterodimerization between the blue and red subunits, then both subunits would be classified as assembling simultaneously. The relative order of assembly for each subunit pair is included in Dataset S1, and all predicted assembly pathways are provided in

Dataset S3. The source code for predicting assembly pathways from protein complex structures is available at <http://github.com/marshlab/assembly-prediction>.

The full set of gene ontology associations for complexes where assembly order and gene order are the same or different is provided in Dataset S4.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, one table, and four datasets and can be found with this article online at <http://dx.doi.org/10.1016/j.celrep.2015.12.085>.

AUTHOR CONTRIBUTIONS

J.M. conceived and designed the research. J.W., T.B., and J.M. performed the computational analyses. J.M. wrote the manuscript with contributions from all authors.

ACKNOWLEDGMENTS

We thank Sarah Teichmann for helpful discussions and comments on the manuscript. This work was supported by a University of Edinburgh Chancellor's Fellowship and Medical Research Council Career Development Award MR/M02122X/1 (to J.A.M.).

Received: August 25, 2015

Revised: November 7, 2015

Accepted: December 17, 2015

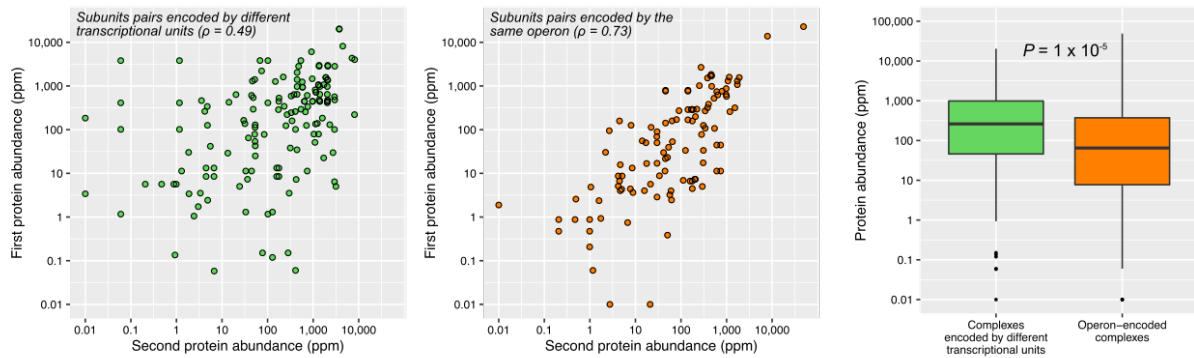
Published: January 21, 2016

REFERENCES

- Ahnert, S.E., Marsh, J.A., Hernández, H., Robinson, C.V., and Teichmann, S.A. (2015). Principles of assembly reveal a periodic table of protein complexes. *Science* 350, aaa2245.
- Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328.
- Duncan, C.D.S., and Mata, J. (2011). Widespread cotranslational formation of protein complexes. *PLoS Genet.* 7, e1002398.
- Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L.J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815.
- Goodsell, D.S., and Olson, A.J. (2000). Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* 29, 105–153.
- Gowrishankar, J., and Harinarayanan, R. (2004). Why is transcription coupled to translation in bacteria? *Mol. Microbiol.* 54, 598–603.
- Hernández, H., and Robinson, C.V. (2007). Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat. Protoc.* 2, 715–726.
- Huntley, R.P., Sawford, T., Mutowo-Muullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., and O'Donovan, C. (2015). The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res.* 43, D1057–D1063.
- Janin, J., Bahadur, R.P., and Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Q. Rev. Biophys.* 41, 133–180.
- Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M.G., and Alon, U. (2001). Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. *Science* 292, 2080–2083.
- Kovács, K., Hurst, L.D., and Papp, B. (2009). Stochasticity in protein levels drives colinearity of gene order in metabolic operons of *Escherichia coli*. *PLoS Biol.* 7, e1000115.
- Levy, E.D. (2007). PiQSi: protein quaternary structure investigation. *Structure* 15, 1364–1367.

- Levy, E.D., Boeri Erba, E., Robinson, C.V., and Teichmann, S.A. (2008). Assembly reflects evolution of protein complexes. *Nature* 453, 1262–1265.
- Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J.S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* 157, 624–635.
- Lim, H.N., Lee, Y., and Hussein, R. (2011). Fundamental relationship between operon organization and gene expression. *Proc. Natl. Acad. Sci. USA* 108, 10626–10631.
- Mao, X., Ma, Q., Zhou, C., Chen, X., Zhang, H., Yang, J., Mao, F., Lai, W., and Xu, Y. (2014). DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res.* 42, D654–D659.
- Marsh, J.A., and Teichmann, S.A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* 84, 551–575.
- Marsh, J.A., Hernández, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., and Teichmann, S.A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* 153, 461–470.
- Marsh, J.A., Rees, H.A., Ahnert, S.E., and Teichmann, S.A. (2015). Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nat. Commun.* 6, 6394.
- Mushegian, A.R., and Koonin, E.V. (1996). Gene order is not conserved in bacterial evolution. *Trends Genet.* 12, 289–290.
- Nishizaki, T., Tsuge, K., Itaya, M., Doi, N., and Yanagawa, H. (2007). Metabolic engineering of carotenoid biosynthesis in *Escherichia coli* by ordered gene assembly in *Bacillus subtilis*. *Appl. Environ. Microbiol.* 73, 1355–1361.
- Oppenheim, D.S., and Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of *Escherichia coli*. *Genetics* 95, 785–795.
- Rajagopala, S.V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S.B., Phanse, S., Ceol, A., et al. (2014). The binary protein-protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.* 32, 285–290.
- Shieh, Y.-W., Minguez, P., Bork, P., Auburger, J.J., Guilbride, D.L., Kramer, G., and Bukau, B. (2015). Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* 350, 678–680.
- Sneppen, K., Pedersen, S., Krishna, S., Dodd, I., and Semsey, S. (2010). Economy of operon formation: cotranscription minimizes shortfall in protein complexes. *MBio* 1, e00177–e10.
- Swain, P.S. (2004). Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J. Mol. Biol.* 344, 965–976.
- Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA* 99, 12795–12800.
- Wang, R., Prince, J.T., and Marcotte, E.M. (2005). Mass spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res.* 15, 1118–1126.
- Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15, 3163–3168.
- Wells, J.N., Bergendahl, L.T., and Marsh, J.A. (2015). Co-translational assembly of protein complexes. *Biochem. Soc. Trans.* 43, 1221–1226.
- Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* 67, 235–242.
- Zaslaver, A., Mayo, A., Ronen, M., and Alon, U. (2006). Optimal gene partition into operons correlates with gene functional order. *Phys. Biol.* 3, 183–189.

A) Abundance data from all organisms



B) Protein synthesis rates from *E. coli*

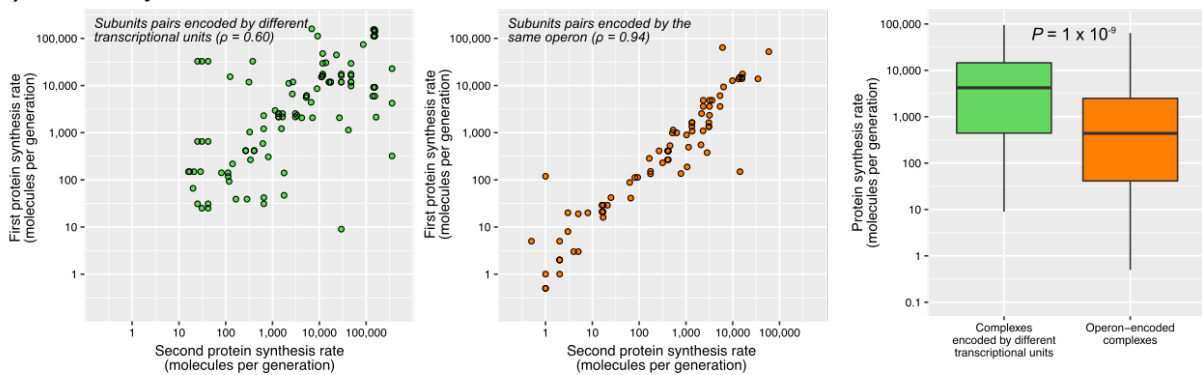


Figure S1: Comparison of subunit pairs encoded by different transcriptional units vs. those encoded by the same operon, using abundance measurements combined from multiple organisms, or absolute protein synthesis rates from *E. coli*, related to Figure 1. A) Same as Fig. 1B-C, except using PaxDB abundance measurements from all prokaryotes instead of just *E. coli*. B) Same as Fig. 1B-C, except using *E. coli* absolute protein synthesis rates derived from ribosomal profiling experiments (Li et al., 2014). The correlations for subunit pairs encoded by the same operon are significantly higher than for those encoded by different transcriptional units in both datasets ($P = 0.004$ for A and $< 10^{-5}$ for B), calculated by randomly shuffling the pairs between two groups of the same size 10^5 times.

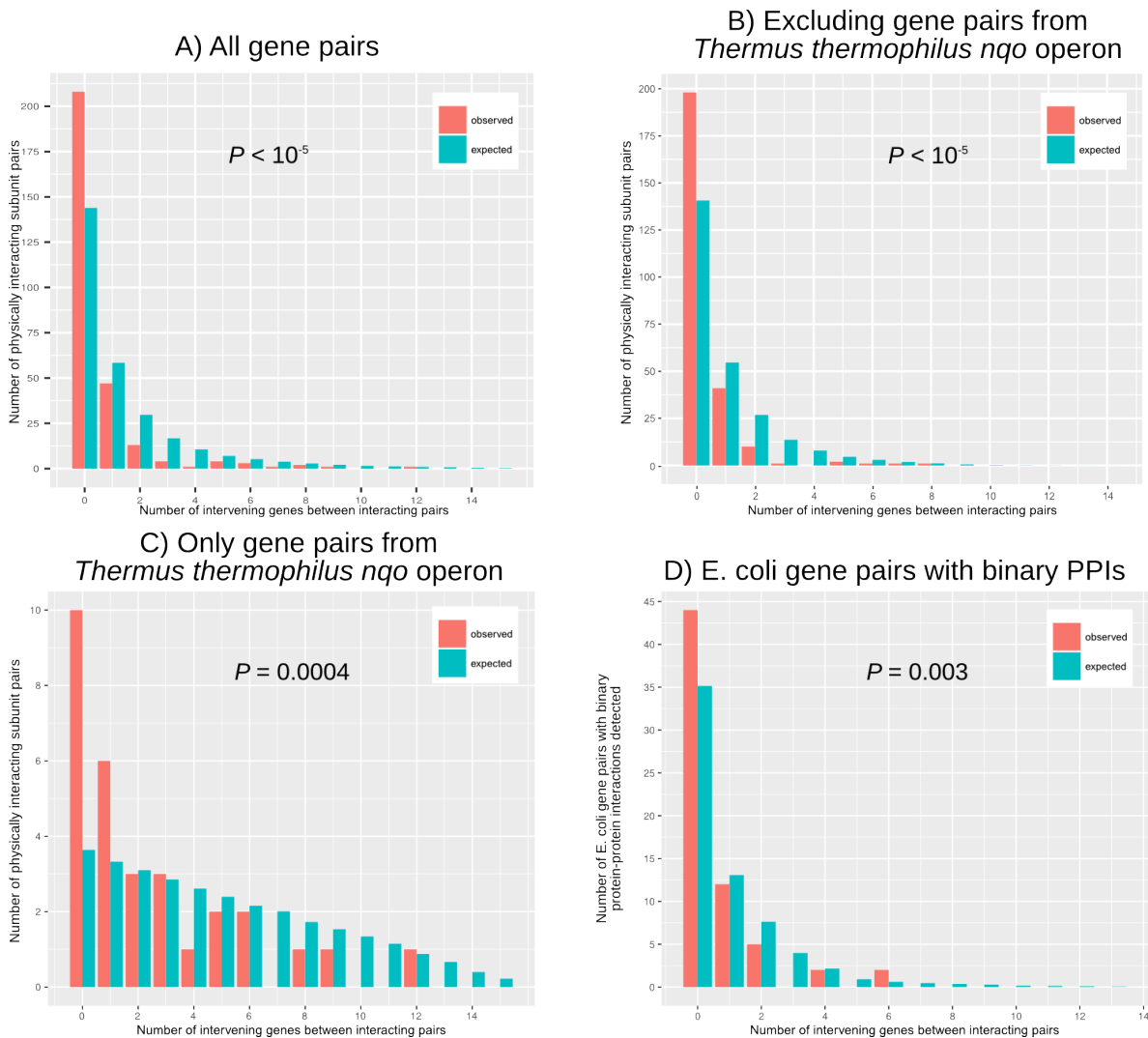


Figure S2: Relationship between operon gene proximity and tendency to physically interact, related to Figure 2. These plots show the number of physically interacting protein pairs (intersubunit interface >200 Å) observed (red), compared to the number expected if the gene order of the operon is randomly shuffled (turquoise). Panel **A**) shows the comparison with all operon gene pairs from our main dataset (as in Fig. 2B), and **D**) shows the *E. coli* pairs with binary protein-protein interactions (as in Fig. 2C). More than half (78/148) of the non-adjacent gene pairs come from a single operon: the *nqo* operon from *Thermus thermophilus* encoding complex I (operon #116617 in Dataset S1). No other operon contributes more than 6 gene pairs. Thus we also present these analyses excluding gene pairs from the *nqo* operon in **B**) and only including *nqo* in **C**). For each plot, the operons were shuffled 10^5 times, and the P -values represent the probability of seeing \leq the total number of intervening genes between interacting proteins.

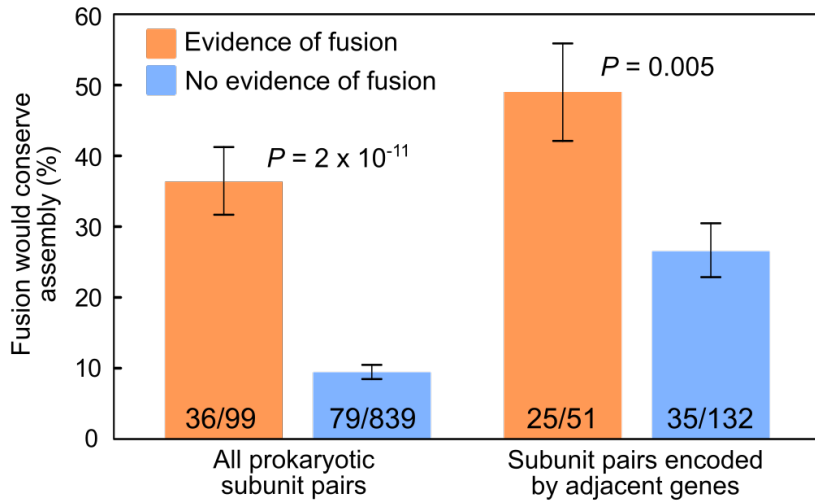


Figure S3: Evolutionary conservation of protein complex assembly pathways by gene fusion events, related to Figure 2. Percentage of cases in which a covalent fusion between protein complex subunits would conserve existing assembly pathways for subunit pairs from heteromers with >2 total subunits, where there is evidence of an evolutionary fusion in some other species (orange), and those where there is no evidence of fusion (blue). On the left, all non-redundant subunit pairs from our dataset are considered, whereas on the right, only subunit pairs encoded by adjacent genes from the same operon are included. *P*-values are calculated using Fisher's exact test. Error bars represent 68% Wilson binomial confidence intervals.

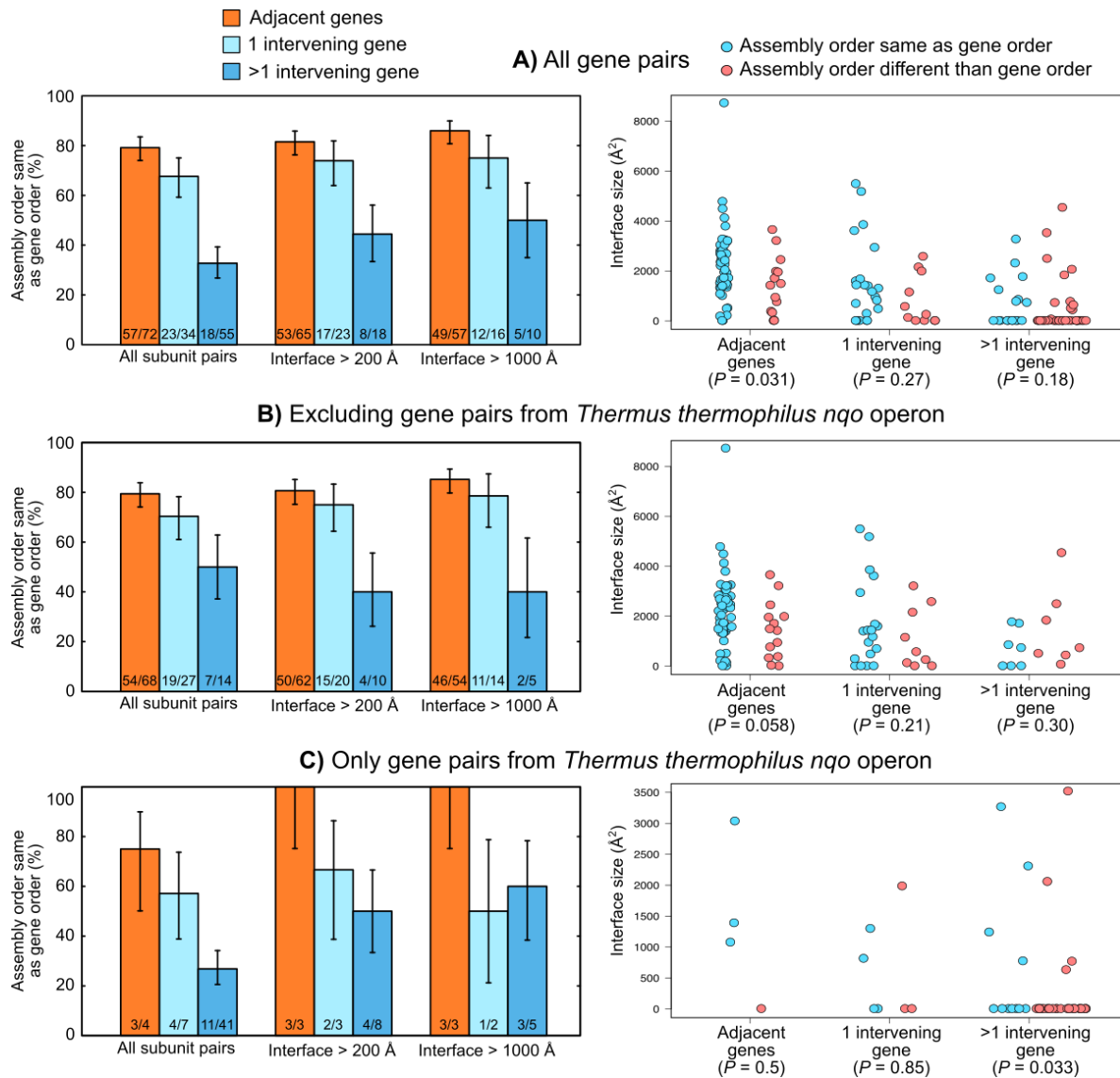


Figure S4: Comparison of the relationship between gene order, assembly order and interface size for adjacent and non-adjacent gene pairs, related to Figure 3. Plots on the left side show the % of gene pairs where evolutionarily conserved gene order is the same as assembly order. They are split into adjacent gene pairs, gene pairs with a single intervening gene, and gene pairs with multiple intervening genes. Error bars represent 68% Wilson binomial confidence intervals. Plots on the right side show a comparison between interface sizes for subunit pairs where assembly order is either the same as gene order or different. P -values are calculated with the Wilcoxon rank-sum test. Since nearly half of the non-adjacent gene pairs come from the *T. thermophilus nqo* operon, as mentioned in Fig. S2, we also show these comparisons excluding pairs from this operon, or only including these genes.

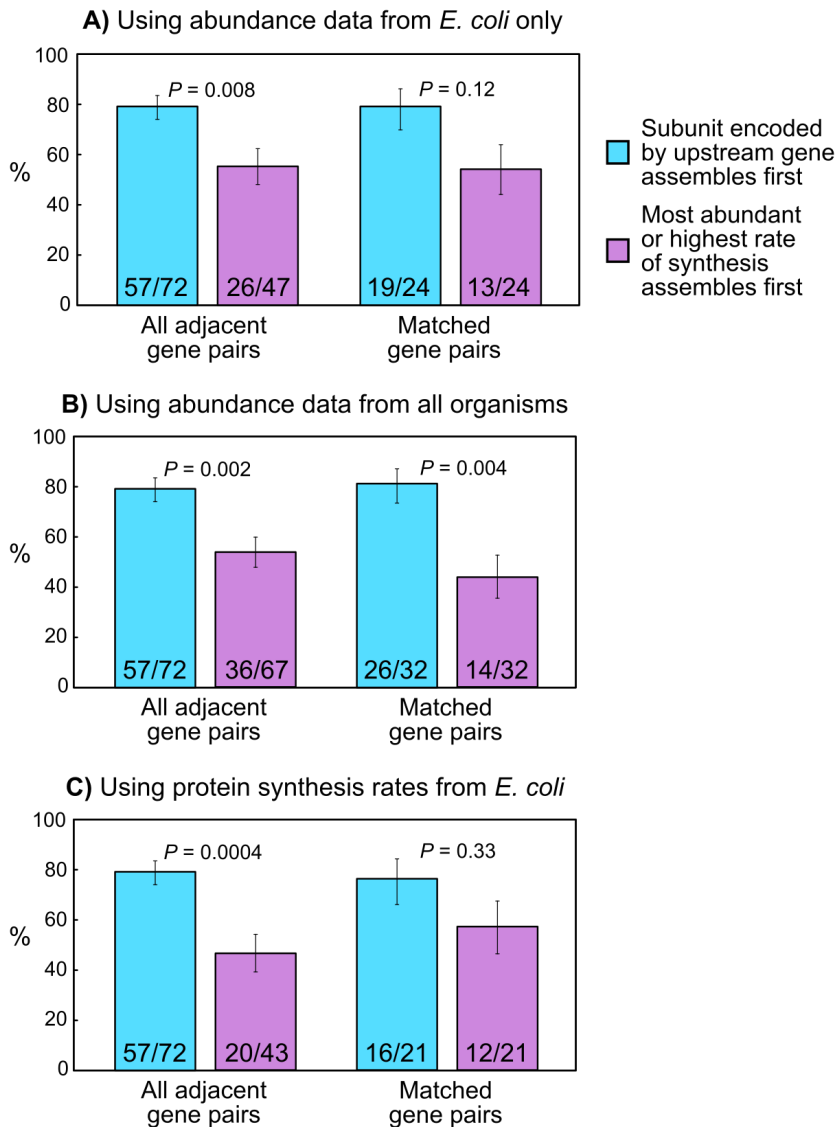


Figure S5: Gene order is a stronger predictor of assembly order than protein expression levels, related to Figure 3. Due to the previous observations that upstream genes in operons tend to be more highly expressed, we tested whether this could possibly affect our conclusion that gene order is optimized for assembly order. We tested whether upstream subunits are more likely to assemble first compared to subunits that are more abundant or have a higher rate of synthesis. “*All adjacent gene pairs*” includes all evolutionarily conserved gene pairs with abundance or synthesis measurements available for both proteins, whereas “*Matched gene pairs*” only includes the much smaller subsets of pairs where abundance measurements are available for both and one subunit is predicted to assemble before the other, so that the same pairs are used for each comparison. Comparisons were performed using (A) abundance data from *E. coli* only, (B) abundance data from all organisms, and (C) absolute protein synthesis rates from *E. coli*. Error bars represent 68% Wilson binomial confidence intervals. *P*-values are calculated with Fisher’s exact test.

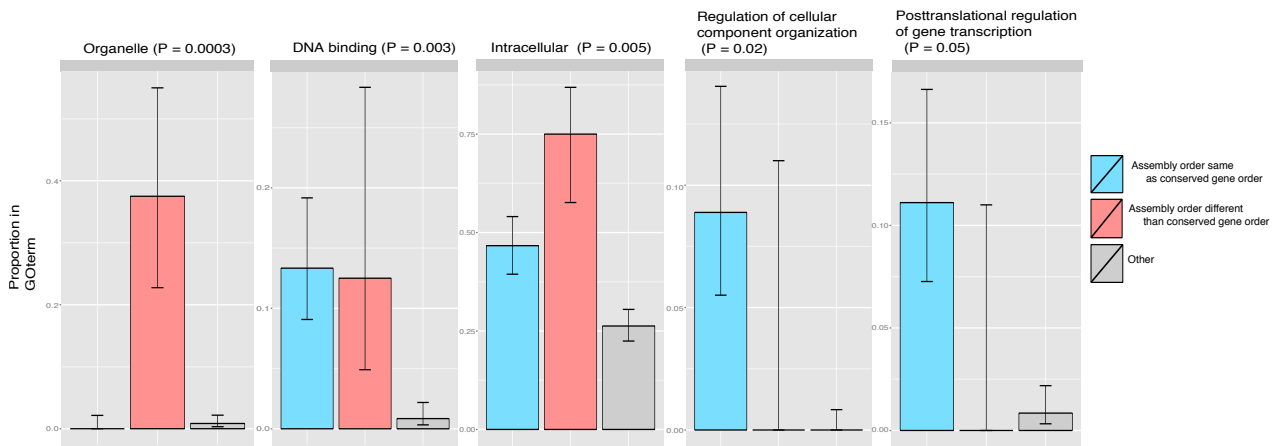
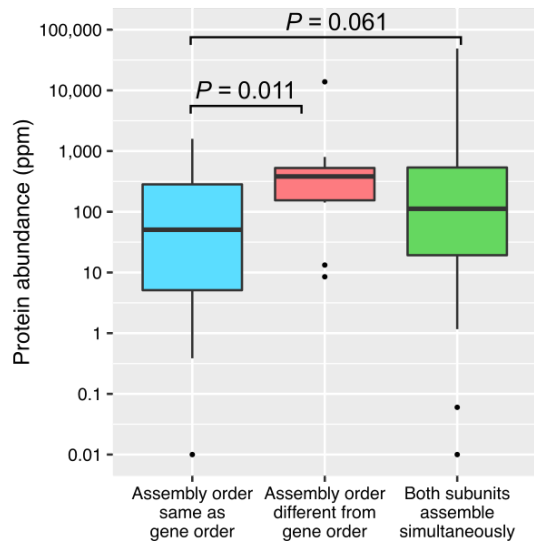


Figure S6: Enrichment of Gene Ontology terms, related to Figure 3. Three groups of complexes are considered in terms of their functional annotations: assembly order same as conserved gene order (blue), assembly order different than conserved gene order (red) and other (grey). The other group includes the complexes where the gene order is not conserved and/or where there is no ordered assembly. The significance of the distribution of each GO term with respect to the three groups was calculated with Fisher’s exact test, using P -values simulated with 2×10^6 Monte Carlo iterations. The GO terms were then filtered for redundancy: if two terms co-occur in $>50\%$ of all proteins in the GOA database where either of the two terms are observed, then only the term with the lower P -value was considered in the non-redundant set. Error bars represent 68% Wilson binomial confidence intervals. The top five non-redundant GO terms with the lowest P -values are shown here, with results for all GO terms given in Dataset S4. Although the enrichment of “organelle” for complexes where assembly order is different than gene order appears striking, it is based upon only three complexes covering quite disparate activities (ribosomal L7/12 stalk, topoisomerase VI and CheY:CheZ), so it is difficult to ascribe much biological meaning to this.

A) Abundance data from all organisms



B) Protein synthesis rates from *E. coli*

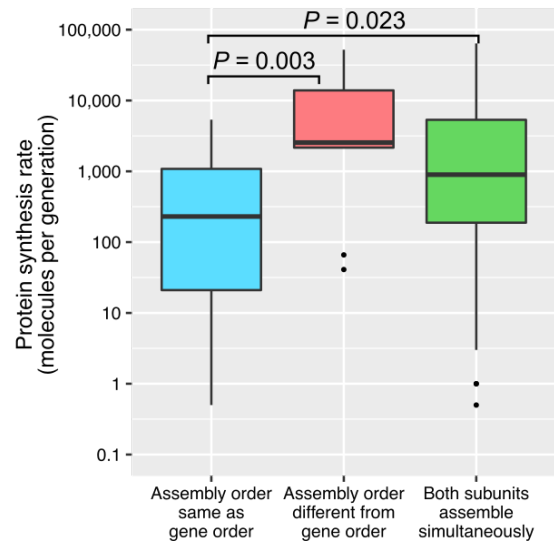


Figure S7: Same comparison as in Figure 4 using abundance measurements from all organisms or absolute protein synthesis rates from *E. coli*, related to Figure 4.

Table S1: Relationship between protein abundance and gene order for adjacent genes encoding different subunits of the same heteromeric complex, related to Figure 3.

Overall, there may be a slight tendency for proteins encoded by upstream genes to be more abundant, although none of these results are close to statistically significant with the binomial test. In addition, when considering rate of protein synthesis derived from ribosomal profiling (Li et al, 2014), the upstream genes actually show lower rates of synthesis, although this is also not significant.

	Subunit encoded by upstream gene is more abundant (<i>E. coli</i> only)	Subunit encoded by upstream gene is more abundant (All species)	Subunit encoded by upstream gene has a higher rate of protein synthesis
Gene order is evolutionarily conserved	26/47 (55.3%)	36/67 (53.7%)	20/43 (46.5%)
Gene order is not evolutionarily conserved	5/10 (50.0%)	10/16 (62.5%)	3/9 (33.3%)
Total	31/57 (54.5%)	46/83 (55.4%)	23/42 (44.2%)

Dataset S1: List of all gene/subunit pairs used in this study and their relevant properties, related to Figure 1.

Dataset S2: Pairs of *E. coli* genes from the same operon with or without evidence of binary protein-protein interaction, related to Figure 2.

Dataset S3: Predicted assembly pathways for heteromeric protein complexes, related to Figure 3.

Dataset S4: Enrichment of Gene Ontology terms associated with protein complex subunits where assembly is either the same or different than operon gene order, related to Figure 3.