Polyprotein strategy for stoichiometric assembly of nitrogen fixation components for synthetic biology

Jianguo Yang\textsuperscript{a,b,1}, Xiaqing Xie\textsuperscript{a,b,1}, Nan Xiang\textsuperscript{a,b}, Zhe-Xian Tian\textsuperscript{a,b}, Ray Dixon\textsuperscript{c,2}, and Yi-Ping Wang\textsuperscript{a,b,2}

\textsuperscript{a}State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, 100871 Beijing, China; \textsuperscript{b}School of Advanced Agriculture Sciences, Peking University, 100871 Beijing, China; and \textsuperscript{c}Department of Molecular Microbiology, John Innes Centre, Norwich NR4 7UH, United Kingdom

Edited by Caroline S. Harwood, University of Washington, Seattle, WA, and approved July 6, 2018 (received for review March 22, 2018)

Re-engineering of complex biological systems (CBS) is an important goal for applications in synthetic biology. Efforts have been made to simplify CBS by refractoring a large number of genes with rearranged polycistrons and synthetic regulatory circuits. Here, a post-translational protein-splicing strategy derived from RNA viruses was exploited to minimize gene numbers of the classic nitrogenase system, where the expression stoichiometry is particularly important. Operon-based \textit{nif} genes from \textit{Klebsiella oxytoca} were regrouped into giant genes either by fusing genes together or by expressing polyproteins that are subsequently cleaved with Tobacco Etch Virus protease. After several rounds of selection based on protein expression levels and tolerance toward a remnant C-terminal ENLYFQ-tail, a system with only five giant genes showed optimal nitrogenase activity and supported diazotrophic growth of \textit{Escherichia coli}. This study provides an approach for efficient translation from an operon-based system into a polyprotein-based assembly that has the potential for portable and stoichiometric expression of the complex nitrogenase system in eukaryotic organisms.

nitrogen fixation | polyprotein | synthetic biology

\textbf{Complex biological systems (CBS)} encode useful functions of interest for biotechnology, including nutrient assimilation, metabolic pathways, and natural product biosynthesis (1). However, genetic engineering of such systems with large numbers of genetic components is difficult, particularly when there is a stoichiometric requirement for balanced expression of the encoded protein components. To date, one approach toward engineering CBS has involved the complete refractoring of each individual gene in which all of the original native regulatory components have been removed (2–4). The disadvantage of this approach is the increased fragility of refracted systems, particularly in cases where there is an essential requirement to maintain stoichiometry of the protein-encoded components. An alternative approach is to reassemble the system as polycistronic modules, which maintain protein complex stoichiometry (5). However, large polycistronic operons cannot easily be utilized to express bacterial CBS in eukaryotic cells. One way to solve this problem is to retain the advantage of bacterial operon coexpression by expressing polyproteins that can be subsequently cleaved by a site-specific protease to release individual enzyme components. A natural example is provided by some RNA viruses that express a single giant gene, which is translated as a polyprotein and is subsequently cleaved by viral protease to form functionally viral components with consistent stoichiometry (6). Here, the well-studied Tobacco Etch Virus protease (TEVp) was used for engineering a CBS using a polyprotein-based strategy. TEVp can be able to cleave its substrate ENLYFQS between the Q and S residues with high specificity and efficiency, leaving an ENLYFQ-tail on the C terminus of the protein encoded upstream and a serine residue on the N terminus of the downstream-encoded protein (Fig. L4) (6). There has been a long-standing interest in reducing dependence on industrial nitrogen fertilizers through engineering nonlegume crops that can “fix” their own nitrogen (7, 8). This could be achieved by transferring nitrogen fixation (\textit{nif}) genes to plant organelles (e.g., chloroplasts, root plastids, or mitochondria) (9–13).

The main challenge to achieving this goal is balanced expression of the many gene products involved in the biological nitrogen fixation (BNF) process (2, 5, 14). Nitrogenase is a complex enzyme consisting of two metalloprotein components: the Fe protein (dinitrogen reductase) and the MoFe protein (dinitrogenase) (15). Although only three genes—\textit{nifH}, \textit{nifD}, and \textit{nifK}—are required to encode the structural subunits of the enzyme, nitrogenase maturation requires the assembly and insertion of three different metal cofactors in a complex multistep process (16). The functionality of the Fe protein is conferred by a [Fe\textsubscript{4}S\textsubscript{4}] cluster (synthesized by NifU and NifS) that bridges the NifH subunits in the Fe protein homodimer and is also dependent on the maturation protein NifM (17, 18). The mature MoFe protein holozyme is a heterotetramer formed from the NifD (\textit{α}) and NifK (\textit{β}) structural subunits that contains an [Fe\textsubscript{4}S\textsubscript{4}] cluster at the \textit{α–β} interfaces, known as the P cluster, and a complex heterometallic cofactor, known as FeMo-co, that has an interstitial carbon atom at its core and also contains an organic moiety, homocitrate [Fe\textsubscript{6}S\textsubscript{4}C-Mo-homocitrate] (19). The assembly pathway for FeMo-co biosynthesis, which contains one of the most complex heterometal clusters in biology, is highly complex, requiring at least nine \textit{nif} genes in vivo (20). The heterotrimer formed by NifE, which is structurally and functionally related to NifD, plays a crucial role as a node in the FeMo-co maturation pathway (21). Maintaining

\textbf{Significance}

The requirement of maintaining balanced expression of a large number of gene products represents a major challenge to the engineering of nitrogen fixation in cereal crops, necessitating reiterative combinatorial assembly cycles to optimize monocistronic gene expression. In this study, we have explored a “fuse-and-cleave” virus-derived polyprotein strategy to reduce gene numbers and achieve balanced expression of protein components required for nitrogenase biosynthesis and activity. After testing and regrouping assemblies on the basis of expression profiles, cleavage patterns, and activity, 14 essential genes were selectively assembled into 5 giant genes that enable growth on dinitrogen. This strategy has potential advantages, not only for transferring nitrogen fixation to plants, but also for the engineering of other complex systems of profound agronomic and ecological importance.

\textbf{Author contributions:} J.Y. and Y.-P.W. designed research; J.Y., X.X., N.X., and Z.-X.T. performed research; J.Y., X.X., R.D., and Y.-P.W. analyzed data; and J.Y., R.D., and Y.-P.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

See Commentary on page 8849.

\textsuperscript{1}J.Y. and X.X. contributed equally to this work.

\textsuperscript{2}To whom correspondence may be addressed. Email: ray.dixon@jic.ac.uk or wangyp@pku.edu.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804992115/-/DCSupplemental.

Published online July 30, 2018.

www.pnas.org/cgi/doi/10.1073/pnas.1804992115
the stoichiometry of the NifEN and NifDK tetrameric complexes and the requirement of balancing expression ratios of all of the \textit{nif} gene products required for nitrogenase synthesis and activity is a vital prerequisite for engineering nitrogenase in nondiazotrophic hosts (2, 3, 5).

In this study, we have chosen a representative BNF system from \textit{Klebsiella oxytoca} to design a polyprotein-based strategy for stoichiometric expression of components required for the biosynthesis and activity of nitrogenase (Fig. 1). Eighteen \textit{nif} genes required for nitrogen fixation in recombinant \textit{Escherichia coli} (5) were grouped according to their expression levels and tolerance of their respective products to the presence of tails remaining after TEVp cleavage. Giant genes were assembled from individual \textit{nif}-coding sequences linked by TEVp recognition sites. After several rounds of testing expression and functionality of cleaved proteins, we regrouped genes as necessary to optimize each polyprotein and established a nitrogen fixation system containing five giant genes. Following TEVp processing of the polyproteins in vivo, their cleavage products enabled synthesis of active nitrogenase, allowing \textit{E. coli} to grow on dinitrogen as the sole nitrogen source. Thus, the polyprotein-based strategy provides an effective solution for stoichiometric expression of nitrogenase components from a single precursor protein, thus reducing the number of genes required to engineer BNF in foreign hosts.

**Results**

**Assessment of Nif Components for Polyprotein Assembly.** First, to utilize the polyprotein-based strategy, it is necessary to assess the expression levels of each component to determine which proteins are suitable for grouping together in terms of expression stoichiometry. Second, since both N-terminal and C-terminal tails can remain after TEVp cleavage, depending on positioning within the precursor polyprotein (Fig. 1A), it is necessary to determine the tailing tolerance of each gene product to design the arrangement of the coding sequences within the giant genes. Expression levels of each \textit{nif} gene in their native operon locations were quantitated.

---

**Fig. 1.** Process for reconstructing biological nitrogen fixation using the posttranslational protein splicing strategy. (A) Digestion of polyproteins by TEVp leads to a C-terminal ENLYFQ-tail for the lead protein A, an N-terminal S-cap for the last protein C, and both an N-terminal S-cap and a C-terminal ENLYFQ-tail for the sandwiched protein B. (B) Arrangement of structural and accessory \textit{nif} genes in the reconstituted operon-based system (pKU7017) from \textit{K. oxytoca} (5). The regulatory \textit{nifLA} operon, which controls transcription from \textit{fus}-dependent \textit{nif} promoters, is not included in this diagram (a detailed description of pKU7017 is provided in Materials and Methods). Essential \textit{nif} genes were regrouped and assembled into giant genes based on their relative expression levels and tailing tolerance. Test-regroup cycles were carried out to select for optimal combinations of the polyprotein-based system that yielded nitrogenase activity as measured by the acetylene reduction assay.
using in-frame translational fusions under steady-state diazotrophic conditions (SI Appendix, Fig. S1 and Table S1). Although this assay does not take into account the stability of the native nif-encoded proteins, we observed that the ratio of nifH to nifDK expression was 2:1 as demonstrated for their respective accumulat- ing protein products in Azotobacter vinelandii (22). These relative expression levels suggest polyprotein designs in which the coding sequences of NifHDKJ, NifENBY, NifUSV, and NifFM can be grouped into giant genes that express cleavable polyproteins. Since the presence of tails after TEVp cleavage may introduce additional constraints on giant gene design, we assessed the tolerance of each Nif protein to the presence of a C-terminal ENLYFQ tail by assaying the nitrogenase activities of constructs in which the coding sequence for this tail was added to individual genes in the 18-gene reconstituted Nif system (Fig. 1 B and SI Appendix, Table S2). This revealed that NifK cannot tolerate the tail and therefore can be located only at the C terminus of a polyprotein. Most of the other components were tailing-tolerant although this reduced the activity of NifB by about 30%.

To minimize the BNF system and simplify the reassignment of genes encoding polyproteins, we omitted nifF, nifX, nifW, and nifZ, which are not essential for BNF in E. coli, as single mutations in these genes do not influence nitrogenase activity (SI Appendix, Fig. S2B). Similarly, nifQ was also omitted because the function of its gene product can be recovered in the presence of a high concentration of molybdenum (SI Appendix, Fig. S2C), as observed previously (23). Guided by the expression groupings defined above and the tailing tolerance of protein components, giant genes were assembled to encode polyproteins with coding sequences flanked by TEVp recognition sites (Fig. 1B). These giant genes are annotated from here on as nifJoH0DoK, nifJoE0NoBoY, nifJoU0SoV, and nifJoF0M, where o indicates the presence of a TEVp-processing site.

**Activity-Based Test and Regroup Cycles.** The functionality of our first-generation polyproteins, both before and after TEVp cleavage, was determined by measuring nitrogenase activities obtained from each giant gene when complemented with the remainder of the nif genes, respectively. Cleavage was achieved by introducing a cassette for expressing TEVp under the control of the Psen promoter after induction with isopropyl-β-D-thiogalactopyranoside (IPTG). Induction of TEVp expression did not influence the functionality of native Nif proteins. When assayed for acetylene reduction, the giant nifJoH0DoK gene, expressed from the nifH promoter, resulted in only 5% activity after TEVp induction (SI Appendix, Fig. S3A). The nifJoU0SoV and nifJoE0NoBoY genes also exhibited low activities of 22 and 19%, respectively (Fig. 2B and SI Appendix, Fig. S3C), in contrast to the nifJoF0M gene, which showed almost full activity (99%) after TEVp induction (Fig. 2B).

The low activities of proteins following cleavage from polyproteins prompted us to test multiple gene combinations in a series of regrouping and retesting cycles. We observed that coexpression of NifJ with NifHDK led to decreased protein levels of NifH, NifD, and NifK compared with proteins expressed from the native

### Table A

| Gene Arrangement | Relative nitrogenase activity (Acetylene reduction assay, %) |
|------------------|-------------------------------------------------------------|
| a1 nifHDK        | 0 25 50 75 100 125                                          |
| a2 nifHDKK (+)    | 0 25 50 75 100 125                                          |
| a3 nifHDK (+)     | 0 25 50 75 100 125                                          |
| a4 nifHDKK (+)    | 0 25 50 75 100 125                                          |
| a5 nifH0DoK      | 0 25 50 75 100 125                                          |

### Table B

| Grouping | Gene Arrangement | Relative nitrogenase activity (Acetylene reduction assay, %) |
|----------|------------------|-------------------------------------------------------------|
| a1 nifHDK |                  | b1 nifE0N, B                                                 |
| a2 nifHDKK (+) |            | 25 50 75 100 125                                             |
| a3 nifHDK (+) |                | 25 50 75 100 125                                             |
| a4 nifHDKK (+) |               | 25 50 75 100 125                                             |
| a5 nifH0DoK |                | 25 50 75 100 125                                             |

**Fig. 2.** Assessment of giant genes for complementation of nitrogenase activity and cleavage of their encoded polyproteins. In all cases, giant genes were expressed from native nif promoters specific to the first gene located in their coding sequences (Materials and Methods). Giant genes taken forward for assembly of the complete polyprotein-based system are indicated in red with an additional asterisk. (A) Example of a dataset for the nifHDK group. The gene arrangement for each giant gene is displayed as colored arrows with letters representing the corresponding nif gene product. TEVp sites are shown as di- monds, and linkers within fusion proteins are indicated by wavy ribbons. The acetylene reduction assay was used to measure complementation by each giant gene of the remaining nif genes in the operon-based system either in the absence of the TEVp-coding sequence (blue bars) or when TEVp expression was introduced using in-frame translational fusions under steady-state diazotrophic conditions (Materials and Methods). Giant genes taken forward for assembly of the complete polyprotein-based system are indicated in red with an additional asterisk. (A) Example of a dataset for the nifHDK group. The gene arrangement for each giant gene is displayed as colored arrows with letters representing the corresponding nif gene product. TEVp sites are shown as diamonds, and linkers within fusion proteins are indicated by wavy ribbons. The acetylene reduction assay was used to measure complementation by each giant gene of the remaining nif genes in the operon-based system either in the absence of the TEVp-coding sequence (blue bars) or when TEVp expression was induced with 20 μM of IPTG (green bars). Acetylene reduction activities by the reconstituted operon-based system in E. coli were assigned as 100% (specific activities in the absence and presence of TEV were 30.4 ± 2.6 and 29.2 ± 1.7 nmol C2H4/min/mg total protein, respectively). Error bars indicate the SD observed from at least two biological replicates. Samples were immediately collected after the acetylene reduction assay for Western blotting (antibodies used are listed in Materials and Methods). Complete gels of the Western blots are provided in SI Appendix, Fig. S9. **(B)** Example of a dataset for the nifHDK group. The gene arrangement for each giant gene is displayed as colored arrows with letters representing the corresponding nif gene product. TEVp sites are shown as diamonds, and linkers within fusion proteins are indicated by wavy ribbons. The acetylene reduction assay was used to measure complementation by each giant gene of the remaining nif genes in the operon-based system either in the absence of the TEVp-coding sequence (blue bars) or when TEVp expression was induced with 20 μM of IPTG (green bars). Acetylene reduction activities by the reconstituted operon-based system in E. coli were assigned as 100% (specific activities in the absence and presence of TEV were 30.4 ± 2.6 and 29.2 ± 1.7 nmol C2H4/min/mg total protein, respectively). Error bars indicate the SD observed from at least two biological replicates. Samples were immediately collected after the acetylene reduction assay for Western blotting (antibodies used are listed in Materials and Methods). Complete gels of the Western blots are provided in SI Appendix, Fig. S9. "EV" represents empty vector, used as a negative control. ImageJ software was used for quantification of NifH protein, and relative expression levels are shown in arbitrary integrated optical density units (B) Summary data for the nifE0N, nifUSV, nifUVWZ, nifU0S, and nifFM groups. The gene arrangement is shown in short format with rows b1, c1, d1, e1, and f1 representing the original operon-based system in E. coli. Commas indicate separate operons, "o" indicates a TEVp site, and "~" indicates a fusion protein. "†" indicates SD values lower than 0.5. Complete datasets for these groupings including the associated Western blots are shown in SI Appendix, Fig. S4.
nifHDK operon, although the cleavage products appeared similar (SI Appendix, Fig. S3B). Consequently, when nifW was excluded from the giant gene, the resultant nifHDoK assembly enabled 75% nitrogenase activity after TEVp cleavage of its polypeptide product, with a similar NifD:NifK protein stoichiometry to that expressed from the native genes (Fig. 2A). As NifB showed weak tolerance (71%) to the C-terminal ENLYFQ-tail (SI Appendix, Table 3), we removed nifY from nifB:nifSoY, and the resultant nifEnoB gene restored 72% of nitrogenase activity, following TEVp induction (Fig. 2B). When the NifY-coding sequence was reassigned to nifFoMoY, cleavage of the larger polypeptide encoded by nifFoMoY resulted in 89% of the nitrogenase activity exhibited by the native components, and increased levels of NifV were observed (Fig. 2B and SI Appendix, Fig. S4E). For the nifUS group, posttranslational splicing of the NifUoSv polypeptide led to decreased levels of NifU and nearly undetectable amounts of NifS, but the level of NifV did not apparently change (compare lanes c1 and c2 in SI Appendix, Fig. S4B). In an attempt to improve this, we removed the nifY sequence from the giant gene, but retained NifUs coexpression with nifUoS. Surprisingly, this restored native levels of NifU and NifS after cleavage and recovered nitrogenase activity to 82% (compare lanes c2 and c5 in SI Appendix, Fig. S4B). To assign nifV to another giant gene, we regrouped it with nifS to form nifOv. Interestingly, in this case the polyprotein product was active even before protease cleavage, resulting in 65% nitrogenase activity, which increased to 95% after splicing with TEVp. Although native levels of NifU were released under these conditions, the amount of NifV decreased (compare lanes d1 and d2 in SI Appendix, Fig. S4C).

To further optimize activity, we carried out additional regrouping of genes encoding polyproteins and also tested the incorporation of fused genes as a means to simplify the ensembles. Since splicing of NifHDoK by TEVp resulted in a decreased amount of NifH (66%), but in wild-type levels of NifD and NifK (compare lanes a1 and a2 in Fig. 2A), we attempted to restore the optimal 2:1 ratio of NifH:NifD (22) through assembly of a giant gene (nifHDoHoDk) expressing two copies of nifH. Although this ensemble slightly increased the level of NifH after cleavage (70%), it did not result in increased nitrogenase activity (compare lanes a2 and a4 in Fig. 2A). To further attempt optimization of NifHDK levels, we also incorporated fusion proteins with different linkers guided by previous studies and natural existing examples (24–27). Fusion NifD–K proteins showed broad tolerance to different lengths of GGGGS linkers, with a maximum activity of 91% when 5x GGGGS linkers were added (SI Appendix, Fig. S5A). We also found that two copies of the NifH protein could be functionally fused with an ArsA linker retaining 89% nitrogenase activity (SI Appendix, Fig. SSB). However, integrating NifD–K or NifH–H fusions into two further ensembles (nifHDoK–K and nifH–HoDk) did not result in higher activities than the original nifHDoK gene (compare lane a2 with lanes a3 and a5 in Fig. 2A).

For the nifENB group, functional fusions of NifE–N and NifN–B were obtained using 5x and 3xGGGGS linkers, which exhibited activities of 91% and 115%, respectively (SI Appendix, Fig. S5 C and D). We replaced the corresponding parts in the nifEnoB gene to generate three more assemblies (nifE–Nob, nifE–noB, and nif–E–N–B). The incorporation of either the NifE–N or NifN–B fusion resulted in higher nitrogenase activities (76 and 89%, respectively) compared with the nifEnoB gene (compare lanes b3 and b4 with b2 in SI Appendix, Fig. S4A). However, when all three genes were fused to express the NifE–N–B protein, only 50% nitrogenase activity was obtained. This decrease may reflect the presence of truncated NifE–N translation products expressed from nifE–N–B (see lane b5 in SI Appendix, Fig. S4A). We also attempted fusion of NifU with NifS and obtained 50% nitrogenase activity with a 5x GGGGS linker when expressed in the native operon arrangement nifU–svwzm (SI Appendix, Fig. S5E). However, this fusion protein had lower activity than that obtained after cleavage of the NifUoS polyprotein (compare lanes c3 and c5 in SI Appendix, Fig. S4B and lanes c2 and c3 in SI Appendix, Fig. S4D).

Assembly and Characterization of Complete Polyprotein-Based Nitrogenase Systems. To combine polyproteins into a functional Nif system, we sequentially replaced the native gene parts with giant gene assemblies (Fig. 3A). Sequential combination of nifHDoK with nifFoMoY and nifEno–B (thus reducing the number of genes from 16 to 9) resulted in relatively small decreases in nitrogenase activity as measured both by acetylene reduction and ^15N assimilation (Fig. 3A, rows I–IV). However, replacement of the native nifUSvWZ genes with nifUosv (thus reducing the number of genes to 5) resulted in a dramatic decrease in activity (10% of the native system) when acetylene reduction was measured (compare rows IV and V in Fig. 3A). The greater influence on acetylene reduction is probably the result of a kinetic effect as the ^15N assay was conducted over a longer time window. Nitrogenase activity improved when nifV was switched from nifUoSv to nifV in the five-gene system (Fig. 3A, compare rows V and VI) as anticipated from the analysis of single polyproteins (Fig. 2B). Nevertheless, the decreased activity observed in the absence of nifW and nifZ prompted us to reconsider their involvement. Although the nifWZ gene products do not apparently have an impact on the activity of our reconstituted system (SI Appendix, Fig. S2D), previous studies suggest that they are required for full activity of the MoFe protein (28, 29). Bearing in mind the native location and expression levels of nifW and nifZ, we assembled additional giant genes designed to express their gene products as polyproteins with NifJ and NifV (nifJov, nifJoVoW, and nifJoVoWoZ). When assayed to complement the native genes, the highest activity was obtained with the polyprotein expressing NifJvW (98%), and no benefit was obtained by incorporating NifZ (SI Appendix, Fig. S4C, compare lanes d3, d4, and d5). These activities were mirrored when these giant genes were complemented with the other four genes encoding polyproteins, with nifJovW again giving the highest level of activity (51% for acetylene reduction; compare arrangements VI, VII, VIII, and IX in Fig. 3A).

Quantitative analysis of protein levels by Selected Reaction Monitoring (SRM) mass spectrometry revealed that, overall, the stoichiometry of most components from the polyprotein-based system matched remarkably well with the respective levels from the reconstituted operon-based system in E. coli and the native system in the original K. oxytoca host (Fig. 3B; compare green, blue, and yellow bars, respectively). This was particularly true for the NifHDK and NifENB proteins, where stoichiometry is important for nitrogenase biosynthesis and activity [the level of NifK was determined by quantification of Western blots (SI Appendix, Fig. S6)]. Unexpectedly, NifU and NifS levels in the polyprotein system were more similar to those in the native K. oxytoca host, compared with the original recombinant system in E. coli (Fig. 3B and SI Appendix, Fig. S6). Although expression of NifU and NifS as a single polyprotein limits complementation of a nifUS deletion (Fig. 2B), only a small benefit was achieved when nifU and nifS were included as separate genes, with the other four polyproteins (compare rows VIII and X in Fig. 3A). Although NifJ and NifV were located on the same polypeptide, the level of NifV was fivefold higher than that of NifJ, which is similar to the ratio obtained from the native system. This suggests that NifJ may have a relatively short half-life, which is not reflected when expression is measured at the transcriptional level (SI Appendix, Table S1). However, the rearrangement of nifF, nifM, and nifY into a giant gene (nifFoMoY) resulted in increased levels of all three protein products, as anticipated from their inclusion in the same polyprotein (Fig. 3B). This may have consequences for the level of NifH, which, unexpectedly, did not decrease in the polyprotein-based system in contrast to the data obtained in Fig. 2A. The increased level of NifM resulting from posttranslational splicing of the NifMFY polyprotein may be responsible for the increased
NiF and NiM accumulation since NiF and NiM are required for the maturation of NiF and NiM (17).

Since the combined five-gene (Fig. 3A, row VIII) and six-gene (Fig. 3A, row X) polyprotein systems exhibited 72 and 75% 15N assimilation activity, respectively, we anticipated that these combinations of giant genes could support diazotrophic growth by E. coli. As in the case of the initial single-gene system (Fig. 3A, row I), the five-gene and six-gene arrangements (assemblies VIII and X, respectively) enabled E. coli to grow on solid media with dinitrogen as the sole nitrogen source, when the expression of TEVp was induced by IPTG. In contrast, assemblies IX and XI, which exhibited lower nitrogenase activities, grew less well under...
these conditions (Fig. 3C). Control experiments in the absence of IPTG, which results in basal levels of TEVp, resulted in poor growth of assemblies VIII and X as anticipated (SI Appendix, Fig. S7).

Discussion

The use of synthetic biology to re-engineer complex systems has transformed our understanding of fundamental mechanisms and has driven a variety of applications, e.g., materials construction, chemical processing, and energy production (30–33). However, the ability to reliably engineer complex biological systems that behave as expected remains limited due to their intrinsic complexity. Thankfully, biology itself provides natural resources that can be exploited to help satisfy the requirements for re-engineering. In the current study, by employing a posttranslational splicing strategy prevalent in RNA viruses, we have established a method for re-engineering CBS using a polyprotein-based strategy. Although others have used TEVp-based posttranslational cleavage to code deliver a small number of proteins into eukaryotic cells (e.g., refs. 34 and 35), to our knowledge, this strategy has not been used to date to engineer complex biological systems such as BNF. This strategy provides an important approach for efficient adaptation from an operon-based system into a polyprotein-based assembly. Balanced expression of CBS components can be achieved by organizing genes with similar expression levels into giant genes expressed as polyproteins, followed by subsequent cleavage to release single components. The advantages of this strategy are particularly useful for transforming complex systems from prokaryotes into eukaryotes. Eukaryotic promoters and terminators often span hundreds of base pairs, nearly 10 times the amount of their bacterial counterparts (36). Monocistronic expression of large numbers of heterogeneous genes in eukaryotes requires complex combinatorial design and the screening of many gene expression parts to surmount difficulties in coordinating the expression level of each component (14). Taken together, translation of genes into polyproteins reduces the number of translational units and maintains stoichiometric expression, thus simplifying the approach for adapting prokaryotic CBS for expression in eukaryotic cells. Although other approaches are available for coordinating expression of multiple genes in eukaryotes in an operon-like format—for example, using internal ribosome entry sites to direct translational initiation within the mRNA—this may result in substantially lower expression levels compared to those by another gene or NifW protein in the polyprotein-based system may reflect a slower rate of P-cluster formation or maturation. Thus, although our polyprotein combination appears to faithfully reproduce the protein stoichiometry observed with native operons, further iterations of this approach may be necessary to fully optimize nitrogenase biosynthesis and activity.

In recent studies, we and others have demonstrated that homologs of the [Fe–S] cluster biosynthesis machinery and electron transport modules present in eukaryotes are compatible with the nitrogenase system (10–12). If the Nif-specific components encoding iron-sulfur cluster biosynthesis (nifU, nifS) and electron transfer (nifH, nifF) can be functionally replaced by their eukaryotic counterparts, the number of components required for, assembly of P clusters (28, 29, 46). However, in our hands, in the heterologous E. coli background, single or double mutants of nifW and nifZ had negligible effects on the activity of the reconstituted operon-based nitrogenase system. In contrast, the polyprotein-based nitrogenase system was more fragile in this respect, with decreased activity being observed in the absence of NifW. It is possible that this represents a kinetic effect since nifW and nifZ mutants delay the induction of nitrogenase activity in K. oxytoca (29). Recently, it has been demonstrated that NifW and NifZ bind to immature forms of apo-NifDK and are likely to function as assembly factors required for P-cluster maturation (47). Since diazotrophic growth is slower in nifW mutants, it has been suggested that either the function of this gene can be partially complemented by another gene or nifW plays a kinetic role in MoFe protein maturation. These observations imply that the operon-based system in E. coli enables optimal maturation of the MoFe protein in the absence of NifW, whereas the requirement for this protein in the polyprotein-based system may reflect a slower rate of P-cluster formation or maturation. Thus, although our polyprotein combination appears to faithfully reproduce the protein stoichiometry observed with native operons, further iterations of this approach may be necessary to fully optimize nitrogenase biosynthesis and activity.

In recent studies, we and others have demonstrated that homologs of the [Fe–S] cluster biosynthesis machinery and electron transport modules present in eukaryotes are compatible with the nitrogenase system (10–12). If the Nif-specific components encoding iron-sulfur cluster biosynthesis (nifU, nifS) and electron transfer (nifH, nifF) can be functionally replaced by their eukaryotic counterparts, the number of components required for, assembly of P clusters (28, 29, 46). However, in our hands, in the heterologous E. coli background, single or double mutants of nifW and nifZ had negligible effects on the activity of the reconstituted operon-based nitrogenase system. In contrast, the polyprotein-based nitrogenase system was more fragile in this respect, with decreased activity being observed in the absence of NifW. It is possible that this represents a kinetic effect since nifW and nifZ mutants delay the induction of nitrogenase activity in K. oxytoca (29). Recently, it has been demonstrated that NifW and NifZ bind to immature forms of apo-NifDK and are likely to function as assembly factors required for P-cluster maturation (47). Since diazotrophic growth is slower in nifW mutants, it has been suggested that either the function of this gene can be partially complemented by another gene or nifW plays a kinetic role in MoFe protein maturation. These observations imply that the operon-based system in E. coli enables optimal maturation of the MoFe protein in the absence of NifW, whereas the requirement for this protein in the polyprotein-based system may reflect a slower rate of P-cluster formation or maturation. Thus, although our polyprotein combination appears to faithfully reproduce the protein stoichiometry observed with native operons, further iterations of this approach may be necessary to fully optimize nitrogenase biosynthesis and activity.

In recent studies, we and others have demonstrated that homologs of the [Fe–S] cluster biosynthesis machinery and electron transport modules present in eukaryotes are compatible with the nitrogenase system (10–12). If the Nif-specific components encoding iron-sulfur cluster biosynthesis (nifU, nifS) and electron transfer (nifH, nifF) can be functionally replaced by their eukaryotic counterparts, the number of components required for, assembly of P clusters (28, 29, 46). However, in our hands, in the heterologous E. coli background, single or double mutants of nifW and nifZ had negligible effects on the activity of the reconstituted operon-based nitrogenase system. In contrast, the polyprotein-based nitrogenase system was more fragile in this respect, with decreased activity being observed in the absence of NifW. It is possible that this represents a kinetic effect since nifW and nifZ mutants delay the induction of nitrogenase activity in K. oxytoca (29). Recently, it has been demonstrated that NifW and NifZ bind to immature forms of apo-NifDK and are likely to function as assembly factors required for P-cluster maturation (47). Since diazotrophic growth is slower in nifW mutants, it has been suggested that either the function of this gene can be partially complemented by another gene or nifW plays a kinetic role in MoFe protein maturation. These observations imply that the operon-based system in E. coli enables optimal maturation of the MoFe protein in the absence of NifW, whereas the requirement for this protein in the polyprotein-based system may reflect a slower rate of P-cluster formation or maturation. Thus, although our polyprotein combination appears to faithfully reproduce the protein stoichiometry observed with native operons, further iterations of this approach may be necessary to fully optimize nitrogenase biosynthesis and activity.

In recent studies, we and others have demonstrated that homologs of the [Fe–S] cluster biosynthesis machinery and electron transport modules present in eukaryotes are compatible with the nitrogenase system (10–12). If the Nif-specific components encoding iron-sulfur cluster biosynthesis (nifU, nifS) and electron transfer (nifH, nifF) can be functionally replaced by their eukaryotic counterparts, the number of components required for, assembly of P clusters (28, 29, 46). However, in our hands, in the heterologous E. coli background, single or double mutants of nifW and nifZ had negligible effects on the activity of the reconstituted operon-based nitrogenase system. In contrast, the polyprotein-based nitrogenase system was more fragile in this respect, with decreased activity being observed in the absence of NifW. It is possible that this represents a kinetic effect since nifW and nifZ mutants delay the induction of nitrogenase activity in K. oxytoca (29). Recently, it has been demonstrated that NifW and NifZ bind to immature forms of apo-NifDK and are likely to function as assembly factors required for P-cluster maturation (47). Since diazotrophic growth is slower in nifW mutants, it has been suggested that either the function of this gene can be partially complemented by another gene or nifW plays a kinetic role in MoFe protein maturation. These observations imply that the operon-based system in E. coli enables optimal maturation of the MoFe protein in the absence of NifW, whereas the requirement for this protein in the polyprotein-based system may reflect a slower rate of P-cluster formation or maturation. Thus, although our polyprotein combination appears to faithfully reproduce the protein stoichiometry observed with native operons, further iterations of this approach may be necessary to fully optimize nitrogenase biosynthesis and activity.
In addition to complex biological systems such as nitrogen fixation, there is considerable interest in engineering bacterial metabolic pathways in plants, for example, to produce antifungal and antibacterial secondary metabolites that provide resistance against pathogens and for effective degradation of xenobiotics. The polypeptide strategy thus may provide a useful approach to engineering complex bacterial pathways for agronomic, ecological, and economic purposes.

### Materials and Methods

**Strains and Media.** *E. coli* Top10 was used for routine cloning and plasmid propagation. *E. coli* JM109 was the strain background used to measure nitrogenase activity either by acetylene reduction or ^15^N assimilation. *E. coli* NCM3722 was used as host for the diazotrophic growth experiments. Luana-Bertiara broth for *E. coli* growth contained 10 g/L of tryptone, 5 g/L of yeast extract, and 10 g/L of NaCl. All nitrogen fixation assays were performed in KPM minimal medium (10.4 g/L of Na_2HPO_4, 3.4 g/L of KH_2PO_4, 26 mg/L of CaCl_2·2H_2O, 30 mg/L of MgSO_4, 0.3 mg/L of MnSO_4, 36 mg/L of ferric citrate, 10 mg/L of para-aminobenzoic acid, 5 mg/L of biotin, 1 mg/L vitamin B1, 0.05% Casamino acids, and 0.8% (v/v) glycerol), and 30 mM of Na_2MoO_4 (according to results from [SI Appendix, Fig. S2C](#)), supplied with 10 mM ammonium sulfate (KPM-HN) for pregrowth or 10 mM glutamate (KPM-LN) for nitrogenase activity assays. Diazotrophic growth experiments were carried out with solid KPM-HN medium or solid KPM medium without 0.05% Casamino acids. Antibiotics were used at the following concentration: 25 µg/mL for kanamycin, 25 µg/mL for chloramphenicol, and 10 µg/mL for tetracycline.

**Plasmids.** Plasmids used in this study are listed in [Dataset S1](#). All plasmids were verified by sequencing before use for further experiments. The pkU7017 plasmid is assembled with seven operon-based biobbies (nifH, nifDKHTY, nifENX, nifUSVZW, nifJ, nifQ, and nifLA) from K. oxytoca, in which regulatory promoters and ribosome-binding sites or additional specific linkers are regulated by the nifL and nifA genes ([5](#)). In this study, each of the operon-based Biobbies, except for the nifLA operon, were removed from the pkU7017 plasmid, respectively, resulting in six different nif-deficient plasmids: pkU7780 (*nifi*), pkU7871 (*nifHDKTY*), pkU7872 (*nifENX*), pkU7873 (*nifUSVZW*), pkU7874 (*nifJQB*) and pkU7875 (*nifJ*). The removed operon-based Biobbies were cloned into pBS402K (BsaI) or pBS102AK (BpiI) plasmid and used as Gateway® (Invitrogen) Gate assembly (48) to generate six plasmids (pkU7878–pkU7881) that complement their pkU7017-deficient derivatives (pkU7870–pkU7875), respectively. The pBS402K and pBS102AK plasmids were assembled with a kanamycin resistance gene, pBR322 origin, Rop sequences, and lacZ coding sequences, flanked with either BsaI or BpiI for TEV protease expression. Twenty microliters of each sample were loaded on 15% SDS-polyacrylamide gels and run with the acetylene reduction assay and suspended in 240 µL PBS buffer supplied with 50 mM Na_2HPO_4·12H_2O, 36 mM KHPO_4·3H_2O, 250 mM NaCl, 10 mM MgSO_4·7H_2O, 10 mM KCl, and 0.05% Tween 20. The gels were stained with Coomassie Blue R250 and imaged with a Transilluminator (Molecular Dynamics). Band intensities were quantified with the Quantity One® software (Bio-Rad Laboratories). The resultant new operons encoding proteins with the ENLYFQ-tail were used for tetracycline-Tailing-Tolerance Assay. To determine the tolerance of each gene product to the C-terminal tail, each nif gene, carrying an extended ENLYFQ-coding sequence, was used to replace the original gene located in each operon. The resultant new operons encoding proteins with the ENLYFQ-tail were verified by sequencing and cotransformed with their corresponding deficient pkU7017 derivatives into the JM109 strain. Transformants were assayed for acetylene reduction activity. Activities obtained from operons without tagged genes were considered as 100% in each case ([SI Appendix, Table S2](#)). For genes that overlap with upstream genes (nifX, nifW, nifZ, nifM, and nifO), the region of overlap together with their respective ribosome-binding sites was amplified and added just downstream of the stop codon of the upstream genes to avoid frameshifts resulting from the addition of the ENLYFQ-coding sequences to the upstream genes.

**Acetylene Reduction and ^15^N Assimilation Assay.** The CH_4 reduction assay and ^15^N assimilation assay were carried out as described previously ([53](#)). To measure the activity of *E. coli* JM109 derivatives expressing reassembled nitrogenase systems, cells were initially grown overnight in KPM-LN medium. The cells were then diluted into 2 mL KPM-LN medium in 20-ML sealed tubes to a final OD_600 of ~0.3. For the acetylene reduction assay, the tubes were repeatedly evacuated and replaced with argon. After incubation at 30 °C for 6–8 h, 2 mL of CH_4 was injected and the gas phase was analyzed ~16 h later with a Shimadzu GC-2014 gas chromatograph. For the ^15^N assimilation assay, air in the tubes was repeatedly evacuated and replaced with high-purity nitrogen gas, and then 3 mL gas was removed and 2 mL of ^15^N (99%, Shanghai Engineering Research Center for Stable Isotopes) was injected. After 48 h of incubation at 30 °C, the cultures were collected and then freeze-dried. Isotope ratios are expressed as δ^15^N where values are a linear transform of the ^15^N/^14^N isotope ratios, representing the per thousand difference between the isotope ratios in a sample and in atmospheric N (N). Data presented are mean values based on at least two biological replicates for both the CH_4 reduction assays and the ^15^N assimilation assays.

**Diazotrophic Growth.** *E. coli* JM109 was the strain background used to measure nitrogenase activity either by acetylene reduction or ^15^N assimilation. However, as this strain carries many genetic deficiencies and requires Casamino acids and vitamin B1 for normal growth, it is not suitable for studying diazotrophic growth. In this case, the prototrophic K. oxytoca K-12 strain NCM3722 was used for diazotrophic growth, which is more robust than strains of the popular MG1655/WS3110 lineage (55). Reconstructed plasmids were transformed into the *E. coli* NCM3722 strain, and the transformants were spread on LB plates with appropriate concentration of antibiotics. After incubation at 37 °C for 16 h, single colonies were picked and streaked onto KPM-HN and KPM-LN plates. For aerobic growth, the plates were directly incubated at 30 °C. For anaerobic growth, the plates were moved to a 2.5-L anaerobic jar (Oxoid AG0017; Thermo Scientific), and subsequently anaerobic gas-generating sachets (Oxoid AN0023; Thermo) and an oxygen indicator (Oxoid BR00028; Thermo) were added. Anaerobic jars were immediately locked and incubated at 30 °C for 3–4 d.

**Western Blot Assay.** In each case, 2-ML cultures combined from four parallel samples (8 mL total volume, OD_600 ~ 0.6) were collected immediately after the acetylene reduction assay and suspended in 240 µL PBS buffer supplied with 60 µL 5X SDS buffer (#C508320; Sangon). After boiling for 10 min, samples were cooled to room temperature and centrifuged at 12,000 × g for 2 min. Twenty microliters of each sample were loaded on 15% SDS-polyacrylamide gels with 5 µL of PageRuler Prestained Protein Ladder (#26616; Thermo) as a marker. Proteins on the gels were subsequently transferred to PVDF membranes (#IPVH00010; Merck). The membranes were blocked with 5% skimmed milk (#223100; BD Difco) in PBS. The antibodies for each Nif protein were used at a dilution of 1:3,000. The secondary antibody goat anti-rabbit IgG-HRP (#ZB-2301; ZSGB-BIO) or goat anti-mouse IgG-HRP (#ZB-2305; ZSGB-BIO) was also used at 1:3,000. Development was accomplished by using two different methods: (i) an enhanced chemiluminescent substrate for HRP (#34077; Thermo) captured by Automatic Chemiluminescence Imaging Analysis System (Tanon-4200) or (ii) directly developed on the PVDF membranes by using the DAB Chromogen/HRP Substrate kit (#ZLI-9018; ZSGB-BIO). Antibodies used for blotting and quantification in this study were as follows: (i) antibodies detecting NifH, NifD, NifN, NifU, NifS, NifV, NifM, and NifN were raised against purified preparations of the corresponding K. oxytoca proteins in rabbits; (ii) antibodies detecting NifK and NifE were raised in rabbits against synthetic peptides corresponding to residues 501–514 (KLDSTDTSGLKTDY) of K. oxytoca NifK protein and residues 1–14 (MKGKNEALLDEPA) of K. oxytoca NifE protein; (iii) antibodies raised in rabbits against synthetic peptides corresponding to residues 179–192 (DGERRSGLRGEAL) of K. oxytoca NifB and residues 31–44
is fully characterized. The selected peptides were chemically synthesized (BANKPEPTIDE LTD) in both stable isotope-labeled (C6-N2 lysine and C6-N4 arginine) and unlabeled forms. The synthesized peptides were subjected to LC-SRM, and an optimal method based on each reference peptide was developed. The absolute quantity of targeted proteins in samples was determined by liquid chromatography–mass spectrometry (LC-MS/MS).

**SRM-Based Quantification of Targeted Proteins by Mass Spectrometry.** Cells were immediately collected after the acetylene reduction assay. Samples were subsequently prepared according to Wscientific. This research was supported by the National Science Foundation of China (51303008), the 973 National Key Basic Research Program in China (2015CB755700 and 2010CB126503), the China Postdoctoral Science Foundation (2016T100013), and the SLS-Qidong Innovation Fund; and the State Key Laboratory of Protein and Plant Gene Research, no. B02. Y.-P. W. is a recipient of a grant from the National Science Fund for Distinguished Young Scholars (NSFC, 39925017). R.D. was funded by the UK Biotechnology and Biological Sciences Research Council (BBS/E/000PR9797).

**ACKNOWLEDGMENTS.** We thank Professor Jilun Li (China Agriculture University, Beijing) for the antisera against nitrogenase proteins. Research was supported by the National Science Foundation of China (NSFC; 31530081) and 973 National Key Basic Research Program China (2015CB755700 and 2010CB126503). The SLS-Qidong Innovation Fund; and The State Key Laboratory of Protein and Plant Gene Research, no. B02. Y.-P.W. is a recipient of a grant from the National Science Fund for Distinguished Young Scholars (NSFC, 39925017). R.D. was funded by the UK Biotechnology and Biological Sciences Research Council (BBS/E/000PR9797).

**REFERENCES.**
53. Cannon FC, Dixon RA, Postgate JR (1976) Derivation and properties of F-prime factors in Escherichia coli carrying nitrogen fixation genes from Klebsiella pneumoniae. J Gen Microbiol 93:111-125.

54. Montoya JP, Voss M, Kahler P, Capone DG (1996) A simple, high-precision, high-sensitivity tracer assay for N\(_2\) fixation. Appl Environ Microbiol 62:986-993.

55. Brown SD, Jun S (2015) Complete genome sequence of Escherichia coli NCM3722. Genome Announc 3:e00879-15.

56. Wiśniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6:359–362.

57. Kirkpatrick DS, Gerber SA, Gygi SP (2005) The absolute quantification strategy: A general procedure for the quantification of proteins and post-translational modifications. Methods 35:265–273.

58. MacLean B, et al. (2010) Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26:966–968.