Balancing expression levels between genes encoding pathway enzymes is a prerequisite for achieving optimized performance of the designed metabolic pathway. However, the random mutagenesis approach is impractical since it requires searching an unnecessarily large number of variants and often results in searching a narrow range of expression levels which are out of optimal level. Here, we developed a predictive combinatorial design method, called UTR Library Designer, which systematically searches a large combinatorial space of expression levels. It accomplishes this by designing synthetic translation initiation region of mRNAs in a predictive way based on a thermodynamic model and genetic algorithm. Using this approach, we successfully enhanced lysine and hydrogen production in *Escherichia coli*. Our method significantly reduced the number of variants to be explored for covering large combinatorial space and efficiently enhanced pathway efficiency, thereby facilitating future efforts in metabolic engineering and synthetic biology.
In this study, we developed a method, called UTR Library Designer, for the combinatorial design of TIR based on a thermodynamic model and genetic algorithm to facilitate systematic optimization of gene expression levels. This method generates TIR sequences (one-to-many) that varies gene expression levels between a minimum and maximum level with a selected number of intermediate points. Furthermore, we applied this method to enhance the lysine and hydrogen productivities by controlling expression levels of ppc and gapA, respectively. Our approach enabled us to efficiently enhance the efficiencies of metabolic pathways compared to random mutagenesis.

Results
Model-driven combinatorial design of mRNA TIR. We developed UTR Library Designer to search the combinatorial space of gene expression levels by designing mRNA TIR sequence that covers a desired range of gene expression levels (Fig. 1). The design solution was optimized by applying a thermodynamic energy model to a genetic algorithm. Briefly, our thermodynamic energy model ($\Delta G_{\text{TIR}}$) calculates the difference in Gibbs free energy before and after the 30S complex assembles onto an mRNA transcript by considering ribosome binding affinity and accessibility to mRNA.

Using a linear relationship between ($\Delta G_{\text{TIR}}$) and log expression level, we can design sequences that meet a specific expression level. We used this relationship for the forward engineering mode of UTR Designer with the genetic algorithm. Briefly, an initial sequence population was randomly generated and the fitness of each sequence was evaluated by the difference between desired and predicted expression levels. Only top $n$ closest sequences in expression with desired level can be remained in the population and others are eliminated. Then, each remained sequence is changed with little mutations on its sequence. These selection and mutation steps were conducted until a sequence was found that has desired level of expression.

UTR Library Designer also utilized this energy model to generate 5′-UTR variants to achieve a desired range of gene expression levels. The difference between UTR Designer and UTR Library designer is that UTR Library designer finds minimum- and maximum-expression sequences at the same time and has post-analysis step to find desired number of intermediates. The algorithm has been optimized to find sequences that can change its expression level from desired maximum to minimum expression levels by the limited number of mutations (Supplementary Fig. 1). Specifically, our method takes a 5′-UTR template and coding sequence, desired minimum/maximum expression levels, and constraints (e.g., expression-level intermediates, nucleotide constraints) for design. Our method finds sequences that could generate desired maximum expression level and minimum expression level with genetic algorithm described above with $n = 1$, separately. Then, it mutates nucleotide sequences in the 5′-UTR to create a pool of mRNA sequences to analyze the effect of mutation in a certain position of the sequences. For example, it ranks mutation positions that change expression levels from major to minor amount. If mutated sequences for maximum and minimum expression achieve the desired levels, the two sequences are combined into a sequence capable of covering a diverse range of expression. As such, the algorithm selectively adds mutations to match the desired number of mutations. After several different trials, the algorithm gives the best solution. Since UTR Library Designer employs a genetic algorithm that mimics an evolutionary process to search optimal sequences in nature, it reaches a desired range of gene expression levels much faster than random trials. For example, UTR Library Designer easily achieved 5′-UTR variants yielding 5,000-fold changes in expression level with 16 expression-level intermediates, which is an extremely rare event (4 in 50,000 trials) using random sequences (Supplementary Fig. S2 and Supplementary Tables S1 and S2). Even when applied to comparisons of greater than 2,000-fold changes, UTR Library Designer is at least 200 times faster than random trials (Supplementary Table S3). We tested two different libraries with 16 expression-level intermediates obtained above using green fluorescent protein (sgfp) in vivo. We could observe that the fluorescence level in vivo was matched well with the predicted expression level in silico (Fig. 2).

Verification of UTR Library Designer for library generation. The ability of UTR Library Designer was also validated by generating large number of variants using fluorescent proteins. Using non-optimized coding sequences of two different fluorescent reporters (green and red fluorescent proteins encoded by gfp and rfp), we designed 5′-UTR libraries containing 128 sequences for each reporter gene to have broad range of expression levels (designed library) and compared them with random library which randomize consecutive five nucleotides around the Shine-Dalgarno (SD) sequence (Supplementary Fig. S3). The maximum expression level covered by these libraries (designed library) was predicted to be 3- to 10-fold higher than that of libraries generated by consecutive random mutations (random library) (Supplementary Fig. S3b and c). When we randomly sampled 1,000 variants among all possible combinations of gfp and rfp libraries in silico, the range of expression levels covered by designed library was predicted to be much larger than that of random library (Fig. 3a and b). Furthermore, combinatorial design after additional codon optimization of the N-terminal region of each reporter gene further increased the predicted maximum expression level of the libraries (reoptimized codon-based designed library) by 3- to 5-fold compared to that of designed library (Supplementary Fig. S3a). Consequently, the difference was 10- to 50-fold higher than that of random library. Thus, the predicted expression-level space covered by reoptimized codon-based designed library was much larger even than that of designed library (Fig. 3c).

To verify the utility of UTR Library Designer in vivo, we analyzed three combinations of gfp and rfp libraries (random, designed, and reoptimized codon-based designed libraries) by co-transformation of two plasmids simultaneously. As shown in Materials and Methods, we first conducted low-throughput analysis by manually isolating 50 clones of each case that have expression levels as broad as possible based on the fluorescence colors. They showed that the patterns of expression-level coverage were similar to those from in silico analysis (Fig. 3d–f). Each fluorescent protein variant showed a linear relationship between the $\Delta G_{\text{TIR}}$ predicted by UTR Designer and log fluorescence intensities, and the average error was 1.98 kcal/mol with standard deviation of 1.60 kcal/mol (Fig. 4). To observe the entire population, we analyzed non-isolated
clones by using two-color FACS. The difference in expression levels between libraries was similar to that expected from in silico analysis as well as low-throughput analysis. The range of expression levels of reoptimized codon-based designed library (Supplementary Fig. S4c) was larger than that of random library (Supplementary Fig. S4a) and even than that of designed library (Supplementary Fig. S4b). Although we also found non-uniform density of sampling points between libraries from FACS analysis (Supplementary Fig. S4), we...

Figure 2 | In silico and in vivo expression levels of libraries with 16 expression-level intermediates using sfp fluorescent reporter. Two different 16-bp library sequences were used for validation assay. The y-axis indicates log fluorescence measured and predicted expression level. Variants of 5’-UTR sequences are as follows: (a) CCTRTTGTCTAAAGKAGSATCGCCM and (b) GCTGMCAGAAAGSAGCRTCMTTG. Experiments were performed in triplicate and error bars indicate standard deviation.

Figure 3 | Verification of UTR Library Designer for library generation. The range of expression levels of 5’-UTR libraries containing 128 sequences for each reporter gene generated by different design methods ((a) and (d): random library; (b) and (e): designed library; (c) and (f): reoptimized codon-based designed library) was examined both in silico (a–c) and in vivo (d–f). For the in silico analysis, 1,000 events were randomly sampled out of a possible 16,384 libraries for each case. For the in vivo analysis, 50 clones randomly selected from agar plates were grown, and fluorescence for each case was measured. Experiments were performed in triplicate and error bars indicate standard deviation.
believe that it is originated from experimental artifact during library construction because the higher density region is concentrated at the lower expression level. It can be due to either transformation of plasmid library with non-functional fluorescent protein (error caused by library preparation during PCR) or non-uniform synthesis of degenerate oligonucleotides. Nevertheless, these results indicate that our method facilitates the investigation of specific expression-level combinatorial-space in a predictive manner.

Enhanced pathway efficiency driven by UTR Library Designer. Next, we applied UTR Library Designer to control the expression level of a gene encoding one of the pathway enzymes that plays a key role in determining pathway efficiency. When the high-throughput colorimetric, fluorescent, and growth-coupled screening methods are available, the number of expression-level intermediates of library can be large enough depending on the number of genes along the pathway to cover a larger feasible space. However, when the throughput of screening is limited, it should be low enough to verify the performance of the constructed variants.

First, we chose the lysine metabolic pathway, seeking to balance the flux distributions of anaplerotic and glycolytic pathways around the phosphoenolpyruvate (PEP) node using a high-throughput screening method based on the Lysine Riboselector comprised of a lysine-responsive riboswitch and a selection marker gene to enable cells showing high lysine production to survive under selection pressure. To achieve this, we attempted to determine the expression level of ppc for increased lysine production by combinatorially designing 5'-UTR libraries (256 variants) of ppc encoding PEP carboxylase, a key anaplerotic enzyme, yielding more than a 10-fold range of expression levels (Fig. 5a; Supplementary Fig. S5). As noted, the library was large enough to cover a broader, but feasible, space since a high-throughput screening method was available. After three rounds of screening process as in our previous study, twenty colonies were randomly selected, and the plasmids were extracted and sequenced. All isolated clones had the same 5'-UTR sequence at the promoter region, and the expression level of the particular sequence was predicted to be in approximately the middle of the range of the entire library (Supplementary Fig. S5). As expected, this enriched strain (WLREU) showed a dramatic increase in lysine production compared to the parental strain (WL3), which showed very little lysine formation was not linearly correlated with the activity of GAPDH in this particularly designed system and rather had local maxima in H2 and H7 strains and the highest maximum in H5 strain, indicating that hydrogen production shows dramatic non-linear behavior with changes in gapA expression levels as other systems did (Fig. 5d). The one variant (H5) showed a 2.5-fold increase in the yield of hydrogen production. The total amount of hydrogen evolved in the H5 strain was also about 4-fold higher than that of the control strain (H0) (Supplementary Fig. S6c). These results indicate that the increase in yield was not due to a retarded growth rate or
reduced glucose consumption rate of the H5 strain. Interestingly, although the difference of the predicted expression level between H5 and H6 strains was around 30% and that of the measured activity was around 20%, they showed completely different capacity of hydrogen production meaning that this change was substantial for controlling the metabolic flux in the cell.

**Discussion**

Identifying biologically relevant maximum and minimum levels of enzyme expression and exploring the optimal level between them is a key to successful optimization of enzyme expression for performance of designed tasks1–5. However, applying a random approach to solve this long-standing issue is often impractical because of two reasons. First, the designable number of variants for optimizing expression level is limited when relying on a random approach because it is virtually impossible to cover such a large solution space. Second, the size of the potential solution space dramatically exceeds the physiologically obtainable search space once the number of mutations are increased6. However, in this study, we showed that our method, UTR Library Designer, could vary expression levels of a target gene across a broad range, while minimizing the number of mutations, through generation of 5’-UTR variants and further optimization of 5’-proximal coding sequences (TIR).

Because of its model-driven library design, our method could be used to examine a broad range of expression levels of target genes (ppc and gapA) to enhance pathway efficiencies. In case of *in silico* analysis, UTR Library Designer generated sequence libraries that achieved 10-fold and 100-fold expression changes for ppc and gapA, respectively. In contrast, with the same number of expression-level intermediates as used for UTR Library Designer (256 for ppc and 8 for gapA), a random approach that generated 10,000 different sets of library pools was largely unable to achieve such expression changes from *in silico* analysis (P-value < 10−4, Supplementary Fig. S7 and S8). Moreover, the probability of the random library including a variant with an expression level similar (95%–105%) to that of the optimal value obtained was approximately 10% (11.31% for ppc and 5.25% for gapA out of 10,000 library pools), indicating that our method efficiently facilitated achieving the specific value.

Since pathway optimization requires fine-tuning of expression levels over a subtle range7, our method could be used for a grid search with a broad range of expression levels to find a sub-optimal expression level in the first round of the screening process, while providing an opportunity for additional search by further narrowing the range of expression levels to be explored in subsequent screening rounds. When considering a complicated pathway composed of several different enzymes, it is necessary to optimize multiple target genes through simultaneous searches of each gene’s expression level. In

---

**Figure 5 | Pathway optimization driven by UTR Library Designer.** (a) Simplified metabolic pathway leading to lysine production in *E. coli*. The expression level of ppc was controlled by using 5’-UTR variants. Glc, glucose; PEP, phosphoenolpyruvate; AcCoA, acetyl-CoA; OAA, oxaloacetate. (b) Physiological comparison between the parental strain (WL3, left panel) and the enriched strains (WLREU, right panel). The left y-axis represents optical density (black closed rectangles with solid line) and lysine (red open triangles with solid line) concentration (g/L), respectively. The left y-axis represents optical density (black closed rectangles with solid line) at 600 nm in log scale. The x-axis represents the culture time (h). Experiments were performed in triplicate and error bars indicate standard deviation. (c) Simplified metabolic pathway leading hydrogen production in *E. coli*. The expression level of gapA was controlled by using 5’-UTR variants. Glc, glucose; G6P, glucose-6-phosphate; G3P, glyceraldehyde-3-phosphate; Hyd, NADPH-dependent-[FeFe]-hydrogenase; Fd, ferredoxin; NFOR, NAD(P)H:ferredoxin oxidoreductase. (d) Hydrogen production yield of the designed variants. The y-axis represents H2 yield per mole of glucose and the x-axis represents the variants. Experiments were performed in triplicate and error bars indicate standard deviation.
contrast to random approaches, our method would be surprisingly efficient to apply in such cases because it does not rely on searching unpredictable random sequences, which could result in a combinatorial explosion. In addition, designing 25-bp 5′-UTR variants without additional genetic components can reduce failures that arise from using repetitive sequences during the generation of libraries13,14. Since robust methods for assembling DNA constructs and editing genomic DNA with high efficiency are readily available15,16, our method could plausibly be used to optimize the expression of multiple enzymes simultaneously for various purposes by covering a practical range of library space.

**Methods**

**Software implementation.** A software implementation of the combinatorial design method described in this study (UTR Library Designer) is available on our web server (http://ubi.postech.ac.kr/utr_library). Users can apply this software to generate 5′-UTR variants and optimized coding sequences (optimal within the same codon type) depending on the sequences and the number of expression-level intermediates that users have input. We recommend to reoptimize coding content with same codon preference on our web server when the 5′-UTR variants fail to satisfy the desired range of expression levels.

**Reagents, bacterial strains, plasmids, and primers.** Phusion polymerase and restriction endonucleases were purchased from New England Biolabs. pACYC-Duet, pETDuet, and PETDuet vectors were purchased from Novagen. E. coli bacterial strains and plasmids used in this study are listed in Supplementary Table S4. The oligonucleotides used for the construction of plasmids and libraries were synthesized by Bioneer (Daejeon, Korea) and are listed in Supplementary Table S5. All other reagents were obtained from Sigma unless otherwise indicated.

**Construction of the 5′-UTR library and strains.** pCDF-mCherry and pCDF-mCherryOpt plasmids were generated by amplifying pCDF-Duet using the 5′-phosphorylated pCDF-M-F-P and pCDF-PET-M-R-P primers followed by blunt-end ligation. The internal XbaI site was removed by site-directed mutagenesis using the pCDF-del-XbaF and pCDF-del-XbaR primers, as described in a previous study17. The mCherry and mCherryOpt genes were amplified using XbaI-mCherry/FspI-mCherry-R and XbaI-mCherryOpt/FspI-mCherry-R primer pairs, respectively, and were inserted into the XbaI and SpI sites of the modified pCDF vector. To test algorithm’s ability, we generated 5′-UTR libraries with 16 expression-level intermediates by using pACYC-ssgfpOpt as a template in a temperature gradient chain reaction (TGR) employing 5′-phosphorylated primers. Each TGR reaction was performed using 50 ng of template, 10 pmol of each primer, 0.5 U Phusion DNA polymerase, 250 mM each dNTP, 10 μl of 5% BSA buffer provided by the manufacturer, and H2O to a final volume of 50 μl. Reactions were carried out on an Applied Biosystems Thermal Block (Applied Biosystems, Foster City, CA, USA) using the following conditions: 30 s at 98 °C followed by 20 cycles of 10 s at 98 °C, 15 s at an annealing temperature determined based on the Tm of the primers and 3 min at 72 °C, followed by a final extension at 72 °C for 10 min. The resulting PCR products were purified using a QiAquick PCR Purification Kit (Qiagen GmbH, Germany), and the template DNA was eliminated by treating with DpnI at 37 °C for 1 h. The PCR products were blunt-end ligated using T4 DNA ligase (TaKaRa, Kyoto, Japan) at 16 °C overnight, and then used to transform the E. coli ElectroMAX DH5α-E strain (Invitrogen, Carlsbad, CA, USA). Purified plasmids were sequenced by Solgent (Daejeon, Korea) using an ABI 3730XL capillary DNA sequencer. In case of 5′-UTR evolved after culturing at 25 °C for 24 h, the culture broth was analyzed by the internal phosphorylated pCDF-M-F-P and pCDF-pET-M-R-P primers followed by blunt-end ligation. The internal phosphorylated primers (gapA-UTR-lib-F-P and gapA-UTR-lib-R-P) were used to amplify each clone as a template with 5′-phosphorylated gapA-UTR library for E. coli HCT101 strain (Invitrogen, Carlsbad, CA, USA). The purified plasmids were sequenced by Invitrogen (Carlsbad, CA, USA) as described in a previous study11 with a modification of supplemented antibiotics, and hydrogen gas evolving transglutaminase (E. coli) was expressed in E. coli BL21 containing the pCDF-Duet vector after culturing at 37 °C for 20 h with 10% of the cells harvested and immobilized. The culture broth was analyzed by the internal phosphorylated primers (gapA-UTR-lib-F-P and gapA-UTR-lib-R-P). The remainders were performed using Phusion DNA polymerase and biomasses were performed using Phusion DNA polymerase and significantly higher specific activity (Units/mg total protein) was observed.

**Statistics and error analysis.** Statistical tests (P-value) were conducted by Python (SciPy stats). The squared correlation coefficient R2 for linear regression is calculated as

\[ R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \]

where y = log (expression level) and GAP activity x = ΔGAPU or relative predicted expression level and N = number of data points.

1. Ajikumar, P. K. et al. Isoprenoid pathway optimization for Taxol precursor overproduction in Erwinia chrysanthemi. Cell 330, 70–74 (2010).
2. Yadav, V. G. et al. The future of metabolic engineering and synthetic biology: towards a systematic practice. Metab. Eng. 14, 233–241 (2012).
3. Pfleger, B. F., Pitera, D. J., Smolke, C. D. & Keasling, J. D. Combinatorial engineering of intragenic regions in operons tunes expression of multiple genes. Nature Biotechnol. 24, 1027–1032 (2006).
4. Blazek, J. & Alper, H. S. Promoter engineering: recent advances in controlling transcription at the most fundamental level. Biotechnol. J. 8, 46–58 (2013).
5. Dietrich, J. A. et al. Automated design of synthetic ribosome binding sites to control protein expression. Nat. Biotechnol. 27, 946–950 (2009).
6. Ajikumar, P. K. et al. Isoprenoid pathway optimization for Taxol precursor overproduction in Erwinia chrysanthemi. Cell 330, 70–74 (2010).
Acknowledgments
This work was supported by grants from the Advanced Biomass R&D Center (ABC-2010-0029800), the Korean National Research Foundation (2013018606, 2012054744, and R31201100101 of the World Class University program) and the Marine Biomaterials Research Center of the Marine Biotechnology Program and the Manpower Development Program funded by the Ministry of Land, Transport and Maritime Affairs, Korea.

Author contributions
S.W.S. and J.-S.Y. designed and performed the experiments, analyzed the data and wrote the manuscript. H.-S.C., J.Y. and S.C.K. performed the experiments. J.P. analyzed the data and edited the manuscript. S.K. and G.Y.J. designed experiments, analyzed the data and wrote the manuscript.

Additional information
Supplementary information accompanies this paper at http://www.nature.com/scientificreports
Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Seo, S.W. et al. Predictive combinatorial design of mRNA translation initiation regions for systematic optimization of gene expression levels. Sci. Rep. 4, 4515; DOI:10.1038/srep04515 (2014).

This work is licensed under a Creative Commons Attribution 3.0 Unported License. The images in this article are included in the article's Creative Commons license, unless indicated otherwise in the image credit; if the image is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the image. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/