Supporting Information

PASIV - A pooled approach-based workflow to overcome toxicity-induced Design of Experiments failures and inefficiencies

Authors:

Alexis Casas, Matthieu Bultelle, Charles Motraghi, and Richard Kitney *

Department of Bioengineering, Imperial College London
Exhibition Road, London, SW7 2BX, United Kingdom

Corresponding Author: Richard Kitney

E-mail: r.kitney@imperial.ac.uk
1 - Design of the Lycopene Operon

The construct design used in the PASIV study (see Figure S1) is based on an operon pattern, and uses the following components:

- A constitutive promoter to drive the operon. Five constitutive promoters are chosen from the SynBIS library, to span a range from weak (K137085) to Strong (Kelly's reference promoter J23101) to very strong (BioFab promoter apFab32) - see Table S1.
- An insulating element RiboJ - inserted between the promoter and the first UTR-RBS.
- All three RBS in the operon can be varied, and for each position, three RBS parts from a subset of the Biolegio library (referred to as RBS 1, 2 and 3 in the paper) are used.

### Lycopene Operon

![Lycopene Operon Diagram](image)

**Legend**

- Promoter
- RBS
- Terminator
- RiboJ
- CDS

**Figure S1** - The lycopene operon design.

| Name       | Type      | Strength (in RPU) |
|------------|-----------|-------------------|
| K137085    | Constitutive | 0.08              |
| J23106     | Constitutive | 0.35              |
| J23108     | Constitutive | 0.61              |
| J23101     | Constitutive | 1                 |
| apFab32    | Constitutive | 2.5               |

**Table S1**. The five promoters chosen for the lycopene study - ranging from weak (K137085; TPU = 0.08) to very strong (apFAB32, RPU=2.5)
The design does not fix the gene order and allows for permutation instead. Constructs are labelled according to their gene order (e.g. BEI is labelled 1) as shown in Table S2.

| Pool / Combination Number | Gene Order |
|---------------------------|------------|
| 1                         | BEI        |
| 2                         | BIE        |
| 3                         | EBI        |
| 4                         | EIB        |
| 5                         | IBE        |
| 6                         | IEB        |

Table S2. The 6 possible gene orders in the lycopene operon
2 - Variation of RBS Expression with its Context

Effective translation rates were computed for all contexts with the RBS calculator (https://www.denovodna.com/) - see supplementary files Pool1 - BEI, Pool2 - BIE, Pool3 - EBI, Pool4 - EIB; Pool5 - IBE and Pool6 - IEB. Figures S2 and S3 show the variations of translation rates for the most commonly represented pools (EBI and EIB) for all RBS in the Biolegio library. Translation rates vary by up to orders of magnitude and RBS ranking were not conserved.

**Figure S2** - Effective translation rates of the BioLegio RBS in all three positions (denoted as Position 1, Position 2 and Position 3) for the gene order EIB. All values are in arbitrary units as returned by the RBS calculator.

Another practical effect of RBS context is that having CrtI downstream of the RBS drastically reduces the translation rate. Figure S4 compares translation rate with the CrtI sequence used in this study vs a codon optimised ctrl sequence (https://eu.idtdna.com/pages/tools/codon-optimization-tool) - showing the translation rate is an order of magnitude large with the optimised sequence (although it was not optimised to increase translation).
Figure S3 - Effective translation rates of the BioLegio RBS in all three positions (denoted as Position 1, Position 2 and Position 3) for the gene order EBI.

Figure S4 - Translation rate with CrtI upstream of the RBS - comparison between the original sequence and a codon-optimised sequence
3 - Design Space - Visualization

3.1 - Design Space Visualization - By Gene Order

Influence on gene order on construct distribution can be observed on Figures S5 to S10. The 15-strong Biolegio RBS library (RBS-A01 to RBS-A12) and RBS 1,2 and 3 was used for this purpose. Some orders lead to clustering along some axes, whereas EBI and EIB sample the space more evenly - offering a simple explanation for their over-representation in constructs yielded by the pooled approach. Figure S11 displays the whole design space accessible with the lycopene operon and the whole BioLegiop library; Figure S12 the design space accessible with the RBS 1,2 and 3 subset.

Figure S5 - Design Space for the BEI Gene Order
Figure S6 - Design Space for the BIE Gene Order
Figure S7 - Design Space for the EBI Gene Order
Figure S8 - Design Space for the EiB Gene Order
Figure S9 - Design Space for the IBE Gene Order
Figure S10 - Design Space for the IEB Gene Order
Figure S11 - Complete Design Space with the entire BioLegio Library - Stratified by Pool/Gene Order
Figure S12 - Design Space for the \{RBS1,RBS2,RBS3\} Subset - Stratified by Pool/Gene Order
3.2 - Design Space Visualization - By Promoter

Another way to visualise the design space is to stratify it according to the leading promoter of the operon. Again, the 15-strong Biolegio RBS library (Figure S13) and RBS 1,2 and 3 (Figure S14) are used for this purpose. Results show spatial distribution is clearly associated with promoter strength. Contrary to RBS, this time the results are more predictable since synthesis rates are proportional to promoter strength with the lycopene operon design. Practically the stronger the promoter, the more likely the construct is to be far from the origin.

Figure S13 - Complete Design Space with the Entire BioLegio Library - Stratified by Leading Promoter
Figure S14 - Design Space for the \{RBS1,RBS2,RBS3\} Subset - Stratified by Leading Promoter
4 - Viability Screening Assay - Results

In phase 2 of the PASIV workflow, viable colonies are picked to be sequenced and cultured for a titration assay. Colony picking was conducted irrespective of the colour of the colony. Only a few colonies were present - confirming previous attempts at cultivating the *E. coli* for lycopene production and how difficult it is. Overall, 99 colonies (from Pools 1-6) were isolated (Table S3). Pools 3 (EBI; 37/99) and 4 (EIB; 54/99) were overwhelmingly represented among these isolates as these pools yielded the most colonies for picking. Pools 1,2 and 6 yielded very few viable colonies (all were picked), while Pool 5 yielded none.

| Pool | Gene Order | Isolate Multiplicity |
|------|------------|----------------------|
| Pool 1 | B-E-I | 1 |
| Pool 2 | B-I-E | 1 |
| Pool 3 | E-B-I | 37 |
| Pool 4 | E-I-B | 54 |
| Pool 5 | I-B-E | 0 |
| Pool 6 | I-E-B | 6 |

Table S3. Distribution of the Isolate Colonies (After Colony Picking). The predominant pools, 3 and 4, have been coloured in blue and purple, respectively.

All picked colonies were sent for Sanger sequencing. Returned sequences were then analysed with CMatch for reconstruction. Eighty-seven were successfully sequenced. Seventy-three out of the sequenced 87 isolates could be identified with sufficient reliability - that is they met the minimum quality requirements for construct matching (each component could be matched with a sufficiently high score). Most of the identified isolates were duplicates and only 23 constructs were unique. Table S4 groups the identified isolates by pool of origin. As in Table S3 the predominant pools are 3 and 4.

| Pool | Gene Order | Construct Multiplicity |
|------|------------|------------------------|
| Pool 1 | B-E-I | 1 |
| Pool 2 | B-I-E | 0 |
Table S4. Results of the Sequence Identification Step by Pool. The predominant pools, again 3 and 4, have been coloured in blue and purple, respectively.

### 5 - Titration Assay - Data Analysis

Two types of measurements were acquired after 24h for each sample: Optical Density at 600nM (standard optical density) and at 471nM. Both measurements were respectively converted into dry cell weight. The yield (ratio of concentration over dry cell weight) was finally calculated, as it is the most commonly used metric for metabolic performance.

Biological replicates can be split into two types. For each isolate, 4 biological repeats (originating from the same colony) are run - they are referred to as 'Type I'. After identification of their genetic material via part matching, the isolates are grouped by distinct constructs ('Type 2' repeats) - in practice, this means colonies are grouped according to their genetic material (the construct). Analysis of the measurements was therefore conducted in two steps - first, for the measurements on the isolates and then after aggregation, for the measurements on the constructs.

#### Level 1 - Measurements on the Isolates

Average values for the yield, DCW and lycopene concentration are plotted, with corresponding error bars, in Figures S15 (Yield), S16 (DCW) and S17 (lycopene concentration). In all cases, the isolates were ranked in descending order of average yield to make comparison easier, while the isolates are identified by their internal Ids (at the first level, the genetic content of the isolates does not matter, only the measurements associated with the isolates).
Figure S15 - Average Yield in mg lycopene/g of dry cell weight (with error bars) of the 73 isolates. Isolates were ranked, in descending order, by value of the average yield.

Figure S16 - Average Dry Cell Weight in g/L (with error bars) of the 73 identified isolates - computed for each isolate from all 4 repeats. Isolates were ranked, in descending order, by value of the average yield to make results easier to relate to Figure S15.
Figure S17 - Average Lycopene Concentration in mg/ml (with error bars) of the 73 identified isolates. Isolates were ranked, in descending order, by value of the average yield to make results easier to relate to Figure S15.

Results show that:

- Average yields range more than an order of magnitude from 70 mg/g of DCW to only a few mg/g.
- A large amount of variation could be observed in the dry cell weight data (from 0.2 to 1.2 mg DCW/ml) - showing that some of the isolates yielded by the pooled approach are very stressed, while some grow seemingly unencumbered. It is notable that the isolates with the best yields also have low DCW.
- The data were noisy - and coefficient of variations (CVs) for all metrics were high:
  - Yield: Mean CV : 0.176 - Std : 0.100
  - DCW: Mean CV : 0.136 - Std : 0.084
  - [Lycopene]: Mean CV : 0.142 - Std : 0.102

Isolate data were further processed to remove outliers on the basis of the yield-CV. 7 isolates were removed due to a value above 50% - they were all very low-expressing isolates. It was expected that some low-expressing isolates would exhibit high yield CVs, because low-yield isolates are also low lycopene-concentration isolates and are therefore more subject to the influence of measurement process (a delicate process due to the DMSO-extraction step).

The remaining isolates (i.e. the non-outliers) were clustered (kmeans clustering, N=2) yielding a low-yield cluster and higher-yield cluster - see Figure S18 (DCW vs Yield) and S19 (DCW vs [Lycopene]). In both figures, linear regression was performed on both clusters. The higher-yield cluster corresponded to lower values of the DCW (as was previously noted by comparing Figures S15 and S16) and exhibited a clear downward trend in DCW-Yield coordinates (Figure S18). Conversely, the low-yield cluster spread over all values of the DCW and exhibited a modest
downward trend (also Figure S18). Both results were expected and are indicative of production-induced stress. However, in terms of lycopene production (as indicated by the measured lycopene concentration), the higher yield cluster does not contain more good performers than the other cluster (Figure S19).

These clustering results are a clear indication that the chosen constitutive design not only places cells under continuous duress and thus restricts the viable region drastically, but also forces the cells to strike a clear trade-off between growth and production, with big implications in terms of lycopene production - lower yield, higher DCW isolates being able to produce as much as high yield, lower DCW isolates.

![Graph showing relationship between Yield and DCW](Figure S18 - Outcome to the Clustering Routine (Yield Vs DCW).)
Figure S19 - Outcome to the Clustering Routine (Yield Vs DCW).

Level 2 - Measurements on the Constructs

In a second step, yield and dry cell weight statistics are calculated for the 23 distinct constructs. Results for the yield are shown in Table S5. Two statistics are used to indicate the best performers in terms of yield (mean of the replicates, and maximum over the replicates).

| Id   | Promoter | RBS E | RBS B | RBS I | Gene Order | Mean Yield | Max Yield | Multiplicity |
|------|----------|-------|-------|-------|------------|------------|-----------|--------------|
| C_01 | J23101   | RBS1  | RBS1  | RBS2  | I-E-B      | 26.10      | 31.85     | 1            |
| C_02 | J23101   | RBS1  | RBS1  | RBS3  | I-E-B      | 58.03      | 80.69     | 3            |
| C_03 | J23101   | RBS1  | RBS2  | RBS2  | E-I-B      | 1.29       | 2.42      | 1            |
| C_04 | J23106   | RBS1  | RBS1  | RBS2  | E-B-I      | 4.83       | 9.66      | 3            |
| C_05 | J23106   | RBS1  | RBS2  | RBS1  | E-B-I      | 0.94       | 1.31      | 1            |
| C_06 | J23106   | RBS1  | RBS2  | RBS2  | E-B-I      | 9.86       | 20.08     | 2            |
| C_07 | J23106   | RBS1  | RBS2  | RBS3  | E-B-I      | 7.63       | 8.26      | 1            |
| C_08 | J23106   | RBS2  | RBS3  | RBS1  | E-I-B      | 7.19       | 34.78     | 2            |
| Construct | Replicates | Yield (mg/g of dry cell weight) | Multiplicity |
|-----------|-----------|-------------------------------|--------------|
| C_09      | J23108    | RBS1 RBS1 RBS2 E-B-I          | 2.84 4.49    | 1 |
| C_10      | J23108    | RBS1 RBS2 RBS2 E-I-B          | 8.25 11.29   | 1 |
| C_11      | J23108    | RBS1 RBS3 RBS2 E-B-I          | 9.17 25.31   | 2 |
| C_12      | K137085   | RBS1 RBS1 RBS2 E-I-B          | 22.73 86.07  | 9 |
| C_13      | K137085   | RBS1 RBS1 RBS2 I-E-B          | 3.44 4.39    | 1 |
| C_14      | K137085   | RBS1 RBS1 RBS3 E-I-B          | 17.54 40.57  | 6 |
| C_15      | K137085   | RBS1 RBS2 RBS2 B-E-I          | 20.85 22.82  | 1 |
| C_16      | K137085   | RBS1 RBS2 RBS2 E-B-I          | 25.05 65.71  | 4 |
| C_17      | K137085   | RBS1 RBS2 RBS2 E-I-B          | 16.54 67.63  | 9 |
| C_18      | K137085   | RBS1 RBS2 RBS3 E-B-I          | 9.49 11.52   | 2 |
| C_19      | K137085   | RBS1 RBS2 RBS3 E-I-B          | 24.48 73.78  | 12 |
| C_20      | K137085   | RBS1 RBS3 RBS1 E-I-B          | 5.80 6.58    | 1 |
| C_21      | K137085   | RBS1 RBS3 RBS2 E-I-B          | 23.23 55.33  | 5 |
| C_22      | K137085   | RBS1 RBS3 RBS3 E-B-I          | 63.18 108.62 | 2 |
| C_23      | K137085   | RBS1 RBS3 RBS3 E-I-B          | 15.38 38.28  | 3 |

**Table S5:** Yield results (in mg/g of dry cell weight) for all distinct constructs drawn with the pooled approach. Constructs of higher multiplicity (4 and above) are coloured in red. Best performers in terms of yield are coloured in purple (max yield) and blue (mean yield of the replicates).

Both statistics, mean or maximum, identify a similar list of strong performers - the discrepancies between both lists being related to the multiplicity of the construct. Average yields for the constructs with highest multiplicity (hence with the highest number of replicates) were often significantly lower than their max value (by as much as 80%) - indicating that genotypically identical constructs in the pool were liable to feature in disparate metabolic states. Conversely, construct C_01 with a multiplicity of only one returned similar values of the mean and max.

In most cases these best performers are in all but two cases among the constructs with higher multiplicity (located in the cluster closest to the origin). The best performer, C_22 (K137085,RBS1,RBS,RBS3) is also located in the cluster. This was unexpected, as constructs in the cluster (with some fitness advantage) were not expected to be strong producers (an activity imposing duress). The best yields corresponded to the strongest RBS (RBS2 and RBS3) placed in front of the final enzyme of the pathway (crtI) - hinting that better yields could be found with stronger RBS in front of crtI and that our original choice of RBS was too restrictive (stronger RBS placed in front of crtI should be tolerated and lead to better yields).
Although, as previously discussed, inter-isolate variation (i.e. between the four replicates of the isolate) is substantial for most isolates, Figures S20 and S21 show that inter-construct variation (between isolates corresponding to the same construct) can be much larger, depending on construct multiplicity. In Figure S20, constructs are ranked according to their mean yield. The best three performers C_02, C_22 and C_01 have low multiplicity (3, 2 and 1 respectively), and are followed by a set of constructs with extremely high CVs due to their higher multiplicity. Comparison with Figure S21, showing the maximum yield, shows that these constructs are capable of high yields themselves, unlike C_01 and C_15 (multiplicity 1).

Figures S22 and S23 display the mean of the dry cell weight and lycopene concentration (the constructs being ranked in the same order as previously). In both figures, the coefficients of variation for the constructs with higher multiplicity are extremely high - again highlighting high intra-construct variation. It is worth noting that in figure S22, the dry cell weight for the best two performers by yield is low - showing the toll lycopene production is taking on the cells.

To decouple inter-isolate and inter-construct variation, the following statistic was added. The statistic is computed in two stages. First, for all isolates, the average is computed for all features (DCW, lycopene concentration and yield). These average statistics are considered reliable indicators of the behaviour of the isolate. Then for constructs of multiplicity larger than one, the isolate the most representative of the potential of a construct is identified. The averaged features of that isolate are then assigned to the construct. In the present case, the isolate with the largest mean yield is selected as we are interested in lycopene production (mean lycopene concentration could also be used). Table S6 shows the value of this new yield statistic (alongside the mean and max) for all constructs. The new statistic does not penalise constructs with higher multiplicity as much as the mean of all repeats. Conversely, because it is based on some averaging, it does not reward outliers as much as the maximum.
**Figure S20** - Average Yield (with error bars) of the 23 distinct constructs. Constructs were ranked, in descending order, by value of the average yield.

**Figure S21** - Maximum Yield of the 23 distinct constructs. Constructs were ranked, in descending order, by value of the average yield as in Figure 17.
**Figure S22** - Average Dry Cell Weight (with error bars) of the 23 distinct constructs. Constructs were also ranked, in descending order, by value of the average Yield.

**Figure S23** - Average Lycopene Concentration (with error bars) of the 23 distinct constructs. Constructs were also ranked, in descending order, by value of the average Yield.
| Id   | Mean Yield | Max Yield | New Yield | Multiplicity |
|------|------------|-----------|-----------|--------------|
| C_01 | 26.10      | 31.85     | 26.10     | 1            |
| C_02 | 58.03      | 80.69     | 68.92     | 3            |
| C_03 | 1.29       | 2.42      | 1.29      | 1            |
| C_04 | 4.83       | 9.66      | 9.00      | 3            |
| C_05 | 0.94       | 1.31      | 0.94      | 1            |
| C_06 | 9.86       | 20.08     | 17.80     | 2            |
| C_07 | 7.63       | 8.26      | 7.63      | 1            |
| C_08 | 7.19       | 34.78     | 11.52     | 2            |
| C_09 | 2.84       | 4.49      | 2.84      | 1            |
| C_10 | 8.25       | 11.29     | 8.25      | 1            |
| C_11 | 9.17       | 25.31     | 17.62     | 2            |
| C_12 | 22.73      | 86.07     | 57.09     | 9            |
| C_13 | 3.44       | 4.39      | 3.44      | 1            |
| C_14 | 17.54      | 40.57     | 30.64     | 6            |
| C_15 | 20.85      | 22.82     | 20.85     | 1            |
| C_16 | 25.05      | 65.71     | 38.78     | 4            |
| C_17 | 16.54      | 67.63     | 58.40     | 9            |
| C_18 | 9.49       | 11.52     | 10.54     | 2            |
| C_19 | 24.48      | 73.78     | 65.71     | 12           |
| C_20 | 5.80       | 6.58      | 5.80      | 1            |
| C_21 | 23.23      | 55.33     | 44.69     | 5            |
| C_22 | 63.18      | 108.62    | 80.86     | 2            |
| C_23 | 15.38      | 38.28     | 30.97     | 3            |

**Table S6**: Yield results (in mg/g of dry cell weight) for all distinct constructs drawn with the pooled approach.
This new statistic was used to separate the constructs into high and low producers. Practically, K-means clustering was used with all 23 constructs and two target clusters and using all three dimensions (yield, concentration and OD/DCW). Figure S24 displays the result of the clustering in the OD vs Yield plane. Linear regression was performed on both clusters. In the high performers (red) a clear downward trend can be observed (consistent with a classic growth-production trade off), while the trend is almost flat for lower producers. Figure S25 displays the result of the clustering in the OD vs Concentration plane (again, linear regression was performed on both clusters). In the high performers (red) no clear downward trend can be observed, while a clear upward trend can be seen for lower producers.

Figure S24 - Outcome to the Clustering Routine (OD Vs Yield).
Figure S25 - Outcome to the Clustering Routine (OD Vs Lycopene Concentration).