Supplementary Materials and Methods:

Description of the Genome Calligrapher algorithm. The Genome Calligrapher web tool is freely accessible at https://christenlab.ethz.ch/GenomeCalligrapher and its service is available at no cost for non-commercial users. The Genome Calligrapher algorithm includes the following processing steps for refactoring DNA sequences for *de novo* DNA synthesis. (1) Adjustment of GC- and AT-contents of the target sequence within a moving window of 21 or 99 bases. (2) Removing unfavourable or obstructive DNA sequences (see disallowed sequence patterns listed in Table S1). (3) Replacing hairpins and direct repeats. The algorithm removes user defined disallowed sequence patterns, as well as predefined homopolymeric sequence stretches, dinucleotide and trinucleotide repeats. For hairpins and repeats, minimum length and maximum gap size can be specified as input parameters. In addition to the three main tasks, the Genome Calligrapher allows the user to customize the codon table by specifying immutable codons and forbidden codons (codons to be erased). In the first part of the program, user defined parameters and disallowed sequences are read into the parameter file. While reading the disallowed sequences from the parameter file, a list of all exception patterns is build up including their corresponding reverse complements.

Accepted input sequence files - The Genome Calligrapher web tool accepts GenBank files with a single sequence record not exceeding 5Mb in file size. The Genome Calligrapher is primarily intended for refactoring of prokaryotic sequences, but eukaryotic sequences can also be processed. GenBank file must conform to the GenBank file specification and list at least one CDS feature and a DNA sequence of more than 100 bases in length. Furthermore, GenBank files with ambiguous DNA letters, out-of-phase CDS or discontinuous CDS (i.e. CDS with a feature location field that contains operators such as ‘join’ commonly found in eukaryotic CDS features with intron-exon structure) are not accepted as input.

The Recoding process - In order to streamline a given DNA sequence for *de novo* DNA synthesis, the algorithm replaces sequence patterns within a given CDS that are unfavorablee for chemical DNA synthesis. The recoding process is based on Monte Carlo synonymous codon substitutions to approximate the natural codon distribution as defined by organism specific pre-computed codon usage tables. Recoding is performed using a global recoding probability specified as an input parameter number between zero and one. With this parameter, the user can control the mean percentage of synonymous codon replacement performed using a random number generator. A value of zero means that no codons are switched, a value of one means that all codons of a given CDS are switched into synonymous ones. If the recoding probability is set to zero, recoding is limited to synonymous codon substitutions to eliminate synthesis constraints and erase disallowed sequence patterns, hairpins or GC content violations.

Codon table generation - The program adjusts the codon usage table of a given organism according to user specifications. The codon usage table is pre-sorted for the amino acid column and codon

frequencies. While reading the file, the internal codon table is build up and completed by appending on every line the number of synonymous codons, the list of synonymous codons and their probabilities. Codons from the list of immutable codons are treated as a separate amino acid group. To identify them, the string "_i " (for immutable codons) will be appended to the corresponding amino acid abbreviation. Immutable codons are never recoded and no other codon from the same amino acid group will be substituted into an immutable codon. Similarly, the amino acid abbreviations of forbidden codons are marked with the appendix "_f". Forbidden codons are always erased and replaced by synonymous codons upon recoding.

To complete the internal codon table, two additional copies of the codon probability list are created: One prefers the selection of GC lowering alternative codons (GC-skewed codon usage table) and the other favours AT lowering codons (AT-skewed codon usage table). The preference is achieved by introducing a weight factor (Fweight) for the individual codon probabilities according to the following equations:

$$\mathrm{F}_{\mathrm{weight}_{\mathrm{GC}}} = 2^{\left(\mathrm{F}_{\mathrm{skew}} \times \left(\mathrm{GC}_{\text{alternative codon}} - \mathrm{GC}_{\text{original codon}}\right)\right)}$$

$$\mathrm{F}_{\mathrm{weight}_{\mathrm{AT}}} = 2^{\left(\mathrm{F}_{\mathrm{skew}} \times \left(\mathrm{AT}_{\text{alternative codon}} - \mathrm{AT}_{\text{original codon}}\right)\right)}$$

In this weight factor equation, GC corresponds to the number of G's and C's, and AT corresponds to the number of A's and T's observed in alternative and original codons respectively. F_{skew} corresponds to the user-defined input parameter 'skew factor', which can be specified in the advanced parameter settings section of the Genome Calligrapher web tool (Figure 1). If F_{skew} is set to zero, no change in the codon probabilities will occur. The pre-set parameter value for the skew factor is 2.

CDS integrity test and GC content optimization - The GenBank input file is read using the routine SeqIO.read from Biopython [1]. This routine delivers many helpful features for GenBank file parsing and extracting the beginning, end or direction of each CDS (protein coding sequence). Using these parameters, all CDS are tested for frame integrity. Compiling a genome out of genomic sequence parts can give rise to CDS remnants that are only partially represented. Due to the overlapping organization of bacterial genomes, extraction of some DNA parts can cause clipping of adjacent CDS. To detect such CDS remnants, the Genome Calligrapher algorithm compares all CDS coordinates against the coordinates of DNA parts. This analysis requires properly formatted DNA parts; with each DNA part represented by a "source" GenBank feature entry in the GenBank file. If either the start or the end of a CDS shares coordinates with a DNA part, the CDS is deemed truncated and if necessary, the reading frames is corrected. In addition, the Genome Calligrapher algorithm tests whether adjacent CDS share overlapping sequences. If a CDS overlap is detected, the overlapping section is excluded from recoding. Next, the main loop is entered, where all CDS are analysed for synthesis constraints and recoded if necessary. This main loop rebuilds the current CDS by stepping through each codon and performs synonymous codon replacements according to the probability specified by the global recoding parameter.

After each step, the rebuilt partial CDS is checked against the GC-content limits for two sliding windows of 99 and 21 base pair in size. Two different sliding windows for GC content adjustment were chosen to offer more flexibility for users to optimize sequences according to diverse synthesis constraints of commercial *de novo* DNA synthesis providers. While other window sizes similar to the 21 and the 99 base pair are also conceivable, we have chosen a 21 base pair and a 99 base pair sliding windows for the following reason. The 21 base pair sliding window is used to detect short sequence stretches that potentially form secondary structures within the roughly 20 base long overlapping regions of oligonucleotides used for subsequent annealing and assembly into dsDNA. Similarly, the 99 base pair window is used to restrict the annealing temperature of the roughly 100bp long oligonucleotides used for dsDNA assembly. In case of a violation, the program aims to fix the GC-content by moving backwards through the CDS and exchanging codons with synonymous codons that reduce the bias. For GC content optimization the probability distributions from the GC or AT skewed codon usage tables are used to select synonymous codon substitutions.

Test for occurrence of disallowed sequence patterns - After optimizing the GC content of the target sequence, the algorithm screens through every CDS to check for occurrence of disallowed sequences, hairpins or direct repeats. If disallowed sequences and repeats are detected at a particular CDS position, the program tries to eliminate them by replacing the last or the second last codon at this particular CDS position with synonymous codons. Here, all combinations of a codon pair from the two synonymous codon lists are checked in the order of their respective frequency occurrence. The order of the algorithm for removing first GC content violation then hairpins and direct repeats and finally the disallowed sequence patterns follows their typical occurrence frequency within bacterial genomes. The most frequently observed constraints in *de novo* DNA synthesis are GC content violations followed by disallowed sequences such as homopolymeric, di-, and trinucleotide repeats, and longer hairpins and direct repeats. There is no guarantee that a removed sequence repeat will not introduce new GC content violations and that a removed disallowed sequence will not introduce new repeat sequences. Therefore, the algorithm generates a statistic output file showing the detailed reduction of all violations. If certain violations have not been fully eliminated, it is possible to run the program again with the same sequence file. Each run produces different results because of the random selection of synonymous codons in the recoding process. Alternatively, the algorithm can be re-run using the recoded sequence output GenBank file and setting global recoding probability to zero. This will often completely eliminate all sequence violations.

Generation of codon usage tables from all available sequenced bacterial genomes. A total of 2776 bacterial genomes were downloaded from NCBI repository (completed bacterial genomes as by

November 2014). Using a custom Biopython script, codon usage tables were calculated for each of these bacterial species by analysing the codon occurrence across all annotated protein-coding genes located on the chromosome or on plasmids sequenced. For this analysis the standard bacterial genetic code was used (NCBI translation table 11). Separate codon tables were calculated from bacterial species with multiple isolates or serovar types sequenced. In addition to bacterial codon usage tables, the Genome Calligrapher tool also includes the codon usage table from *Saccharomyces cerevisiae* S288c to permit sequence refactoring for synthetic biology applications in yeast. To customize the recoding of synthetic DNA constructs, users can select any of the 2777 pre-computed codon tables by an autocomplete-assisted input field (Figure 1). The pre-computed codon usage table can be downloaded from the Genome Calligrapher result page.

Occurrence of DNA synthesis constraints across bacterial genome sequences. To assess the frequency and occurrence of *de novo* DNA synthesis constraints across bacterial genome, GenBank files for a total of 4720 microbial chromosomes and plasmids were processed by the Genome Calligrapher algorithm. Bacterial GenBank files were downloaded from NCBI repository [2]. Standard parameter settings were used for the analysis. The CDS offset was set to one amino acid to exclude the start codon of CDS. Global recoding probability was set to zero. The GC-content within the 99 base pair sliding window was set between 0.3 and 0.7 and for the 21 base pair sliding window was set between 0.15-0.85. Direct repeat length was set to 12 base pair with maximum repeat spacers set to 20 base pair. No immutable or forbidden codons were specified and only the standard disallowed sequence patterns that impede *de novo* DNA synthesis were used. A complete overview of the parameters used can be found in Data SI. For each genome, the Genome Calligrapher algorithm was used to analyze the number of GC-content violations, occurrence of direct and indirect repeats and the occurrence of disallowed sequence patterns impeding *de novo* DNA synthesis. To determine the degree of recoding required to overcome synthesis constraints, the Genome Calligrapher algorithm was used to iterate through each CDS and record the number of synonymous codon replacement upon recoding. Complete statistics from the synthesis constraint analysis across bacterial genome are listed in Data SI.

Design of a synthetic essential genome construct. The comprehensive list of DNA parts corresponding to the entire set of essential and high-fitness sequences required for rich-media growth of *Caulobacter crescentus* was generated using a previously identified essential genome data set [3]. The part list includes DNA sequences encoding proteins, RNA and regulatory features as well as small essential inter-genic sequences. Part boundaries of protein coding genes were set to the CDS coordinates according to the *Caulobacter crescentus* NA1000 genome annotation (NCBI Accession: NC_011916.1) plus additional regulatory and terminator regions. The complete part list is listed in supplementary data

(Data SI). Coordinates of regulatory upstream sequences of essential genes were set according to previously identified essential promoter regions [3] and, when necessary, were enlarged to include strong transcriptional start sites as determined by RNAseq [4]. For essential or high-fitness genes, predicted Rho-independent terminator sequences were included as identified by the WebGeSTer DB [5]. The resulting DNA parts were concatenated in order and orientation as found on the wild-type genome and compiled into a 766828 base pair long synthetic genome constructs. This genome construct was then partitioned into thirty-nine 20 kb long segments that were further partitioned into 3 kb DNA building blocks. The GenBank file of the entire designed synthetic essential genome construct is available upon request from the authors.

Optimization and synthesis feasibility analysis. To perform the synthesis feasibility analysis prior and after recoding, the synthetic essential genome sequence was streamlined by the Genome Calligrapher tool using standard recoding parameter settings. Recoding probability was set to 1.0, the codon table specific for *Caulobacter crescentus* NA1000 was used, forced recoding was selected, 5' CDS offset was set to one amino acid, codon skew factor was set to 2.5 and GC content was set between 0.3-0.7 (for 99 base pair window) and 0.15-0.85 (for the 21 base pair window). In addition, the Genome Calligrapher's built-in exception sequences were used. For homo-polymeric sequences no more than six consecutive G's or six C's and no more than nine T's or nine A's are allowed. Di-nucleotides and tri-nucleotide repeats must not exceed more than ten or six repetitions respectively. The synthesis feasibility analysis was requested from a commercial synthesis provider (IDT, Integrated DNA Technologies, Leuven, Belgium) for both the sequence optimized synthetic essential genome as well as the original synthetic essential genome sequence based on wild-type parts.

Sequence optimization and synthesis of eight 20 kb synthetic genome segments. To optimize the sequence of the eight 20 kb synthetic genome segments (segment 7,8,9,10,11 and segment 29,30,31, for an overview of the segments, see Figure 4) similar parameter settings were used except that the recoding probability across different segments was gradually incremented from 0.125 to 1.0 (Table S6). In addition, the 5' CDS offset was set to four amino acids to prevent recoding within the first twelve bases of a CDS. Disallowed sequences included endonuclease sites for BsaI, AarI, BspQI, PacI and PmeI. Furthermore, the AGT, ATA, AGA, GTA and AGG codons, which are rare codons in *Caulobacter crescentus*, were set as immutable codons. The amber stop codon TAG and the two TTA and TTG codons for leucine were set as forbidden codons and erased upon recoding. The eight 20 kb long synthetic essential genome segments were partitioned and ordered as synthetic dsDNA building blocks from a commercial provider of *de novo* DNA synthesis (Gen9, Inc. Cambridge, MA, USA). Sequences of the partitioned DNA building blocks and information on the 4 out of 61 fragments that failed the first round of synthesis are listed in

Table S5. The missing four 3 kb dsDNA building blocks were reordered from a second supplier of synthetic DNA (GeneArt, Life Technologies, USA). The 3 kb synthetic dsDNA building blocks were subsequently assembled into 20 kb segments and cloned into the low copy plasmid pMR10 using yeast recombineering. The assembled 20 kb synthetic segment were sequence verified using PacBio sequencing.

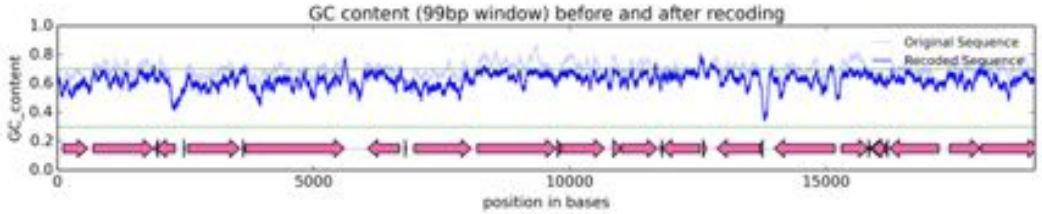
(1) Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics 25*, 1422–1423.
(2) Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004) GenBank: update. *Nucleic Acids Res. 32*, D23–6.
(3) Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Passarelli, B., Coller, J. A., Fero, M. J., McAdams, H. H., and Shapiro, L. (2011) The essential genome of a bacterium. *Mol. Syst. Biol. 7*, 528–528.
(4) Zhou, B., Schrader, J. M., Kalogeraki, V. S., Abeliuk, E., Dinh, C. B., Pham, J. Q., Cui, Z. Z., Dill, D. L., McAdams, H. H., and Shapiro, L. (2015) The global regulatory architecture of transcription during the Caulobacter cell cycle. *PLoS Genet.* (Casadesús, J., Ed.) *11*, e1004831.
(5) Mitra, A., Kesarwani, A. K., Pal, D., and Nagaraja, V. (2011) WebGeSTer DB--a transcription terminator database. *Nucleic Acids Res. 39*, D129–35.

Figure_S1:

GenomeCalligrapher result page

A Download output files: Recoded genbank file, Recoding statistics, Recoding log file, Recoding parameters

B



C

Codon frequencies table after recoding:

TTC	Phe	84	(161)	TCC	Ser	54	(34)	TAC	Tyr	45	(47)	TGC	Cys	19	(40)
TTT	Phe	96	(19)	TCT	Ser	8	(2)	TAT	Tyr	46	(44)	TGT	Cys	26	(5)
TTA	Leu_f	0	(0)	TCA	Ser	3	(3)	TAA	stop	6	(5)	TGA	stop	12	(8)
TTG	Leu_f	0	(16)	TCG	Ser	68	(82)	TAG	stop_f	0	(5)	TGG	Trp	63	(63)
CTC	Leu	208	(96)	CCC	Pro	112	(118)	CAC	His	43	(84)	CGC	Arg	140	(272)
CTT	Leu	79	(16)	CCT	Pro	35	(11)	CAT	His	60	(19)	CGT	Arg	91	(21)
CTA	Leu	22	(5)	CCA	Pro	20	(5)	CAA	Gln	94	(21)	CGA	Arg	42	(12)
CTG	Leu	236	(412)	CCG	Pro	125	(158)	CAG	Gln	78	(151)	CGG	Arg	116	(84)
ATC	Ile	113	(247)	ACC	Thr	121	(177)	AAC	Asn	35	(70)	AGC	Ser	88	(100)
ATT	Ile	143	(9)	ACT	Thr	11	(1)	AAT	Asn	48	(13)	AGT	Ser_i	0	(0)
ATA	Ile_i	1	(1)	ACA	Thr	7	(1)	AAA	Lys	104	(8)	AGA	Arg_i	2	(0)
ATG	Met	95	(95)	ACG	Thr	102	(62)	AAG	Lys	92	(188)	AGG	Arg_i	1	(3)
GTC	Val	209	(238)	GCC	Ala	310	(548)	GAC	Asp	141	(274)	GGC	Gly	197	(364)
GTT	Val	55	(22)	GCT	Ala	107	(30)	GAT	Asp	180	(47)	GGT	Gly	101	(29)
GTA	Val_i	2	(2)	GCA	Ala	41	(12)	GAA	Glu	203	(93)	GGA	Gly	54	(11)
GTG	Val	177	(181)	GCG	Ala	353	(221)	GAG	Glu	126	(236)	GGG	Gly	104	(52)

Note 1: Original codon frequencies are listed in brackets.
Note 2: Amino acid appendix _i indicates immutable codons, amino acid appendix _f indicates forbidden codons.
Note 3: Codons at skipped begin of ORF or within overlapping ORF regions are excluded.

Figure S1. Results page of the Genome Calligrapher web tool. (A) The upper panel lists data output files to be downloaded by the user, including recoding statistics, recoding parameters and a detailed log file in the text file format (.txt). The recoded sequence file can be downloaded in the GenBank (.gb) file format. (B) Graphical plot of the GC content from the 20 kb test segment_30 before and after recoding. CDS location and orientation are shown with red arrows. (C) Codon frequencies table of segment_30 before and after recoding.

The supplementary Data SI provided as an excel file (Data_SI_Tables_S1,2,3,5.xlsx) contains the following supplementary tables.

Table S1 (Data SI): List of the parameter settings used by the Genome Calligrapher to assess synthesis feasibility across all sequenced bacterial genomes.

Table S2 (Data SI): Synthesis feasibility analysis across 4720 sequenced bacterial genomes.

Table S3 (Data SI): List of all essential and high-fitness DNA parts used to compile the essential synthetic genome of *Caulobacter crescentus*.

Table S5 (Data SI): List of 3kb building blocks for assembly of synthetic genome segments.

Table S4: Cumulative codon frequencies across eight 20kb synthetic genome segments before and after recoding by the Genome Calligrapher algorithm.

1st base	2nd base: U			2nd base: C			2nd base: A			2nd base: G			3rd base
U	F	862	(1289)	S	372	(331)	Y	419	(478)	C	158	(257)	C
		565	(138)		93	(55)		389	(330)		124	(25)	U
	L	0 [a]	(10)		40	(38)	*	58	(42)	*	82	(56)	A
		0 [a]	(197)		834	(927)	*	0 [a]	(42)	W	515	(515)	G
C	L	1225	(643)	P	809	(810)	H	375	(567)	R	1339	(2153)	C
		560	(189)		332	(117)		326	(134)		704	(365)	U
		151	(35)		215	(55)	Q	628	(194)		308	(75)	A
		2337	(3217)		1043	(1417)		793	(1227)		752	(510)	G
A	I	1305	(2017)	T	1081	(1505)	N	523	(785)	S	660	(648)	C
		799	(87)		110	(39)		405	(143)		24 [b]	(24)	U
		19 [b]	(19)		98	(36)	K	700	(95)	R	25 [b]	(25)	A
	M	987 [c]	(970)		902	(611)		1152	(1757)		36 [b]	(36)	G
G	V	1591	(1789)	A	2615	(3685)	D	1427	(2178)	G	1779	(2885)	C
		381	(210)		867	(275)		1189	(438)		791	(354)	U
		32 [b]	(32)		246	(84)	E	1410	(953)		447	(106)	A
		1395	(1367)		2016	(1700)		1310	(1767)		643	(315)	G

Abbreviations: Numbers in brackets indicate the codon occurrence prior to recoding.

[a] Forbidden codons that were erased from the genetic code are shown in red.

[b] Rare codons set as immutable codons are highlighted in blue.

[c] Due to overlapping CDS a few instances of forbidden codons (TTA,TTG and TAG) could not be erased by the Genome Calligrapher algorithm and were manually removed, resulting in non-synonymous changes in the reading frame of the overlapping CDS.

Table S6: Overview of recoding applied to the eight 20kb synthetic essential genome segments

Synth. Genome Segment	Size [a] [bp]	Recoding Probability	Total Codons	Recoded [b] Codons	GC-content %
Segment_7	20593	0.125	5683	985	63.1
Segment_8	19317	0.15	5184	1075	63.3
Segment_9	21367	0.1875	5965	1554	64.2
Segment_10	20370	0.23	5264	1594	64.3
Segment_11	20828	0.3	6349	2320	64.0
Segment_29	19025	0.4	4924	2278	63.4
Segment_30	19403	0.567	5505	3288	62.1
Segment_31	20247	0.9	5750	4979	55.6

[a] Cumulatively, all eight synthetic essential genome segments cover 161150 bp in size. Total synthesis effort was 168103 bases due to overlapping assembly linker.

[b] Recoded codons include the number of synonymous codon substitution due to removal of synthesis constraints plus additional random synonymous codon replacement as specified by the recoding probability parameter.