Specific Aims

RNA viruses that have been attenuated make excellent live vaccines, and several such vaccines have had major public health impact. However, major stumbling blocks need to be overcome if live vaccines are to be used more broadly. In particular, we need to know how to predict levels of attenuation and predict reversion: attenuated viruses may revert into a more virulent form within the subjects receiving the vaccine. This phenomenon occurs with regularity in the case of the Oral Polio Vaccine (OPV). The Sabin 2 strain, in particular, differs in only two nucleotide positions from wild type, and reverted virus can commonly be isolated from vaccinated patients. While this problem can be overcome by using inactivated viruses as vaccines, live vaccines frequently have advantages in terms of administration, storage, and conferred immunity. There is thus a critical need to identify and develop reliable, evolutionarily stable approaches to viral attenuation. Furthermore, synthetic biology now allows us to create arbitrary viral genomes, so any conceived attenuation strategy is no longer limiting. We merely need to know how to predict attenuation—and its evolutionary reversal.

Our long-term goal is to elucidate the biological mechanisms of viral attenuation, fitness, and adaptation. The objective for this proposal is to investigate viral attenuation and recovery in bacteriophage T7. Our central hypothesis is that engineered transcriptional and translational de-optimization yields reliable, evolutionarily stable attenuation of viruses. T7 provides a unique model system with which to probe generic principles of viral attenuation, because its genetic regulatory circuitry is well understood and a detailed, mechanistic computer model exists to interpret and predict transcription and (to a lesser extent) translation. We have assembled an experienced team of investigators to pursue this project, consisting of an experienced phage evolutionary biologist (Jim Bull), an expert in proteomics and molecular biology (Dan Boutz), and an expert in translational selection and codon-usage bias (Claus Wilke), all at The University of Texas at Austin. We have three Specific Aims:

Aim 1: Assess fitness effects and recovery suppression in different genetic recodings of bacteriophage T7. *Hypothesis:* Various regulatory modifications **(***codon de-optimization, gene rearrangements, RNA-structure modifications, promoter deletions) reduce fitness to varying degrees; the more that fitness reduction is caused by genetically irreversible modifications, the more resistant these recodings are to recovery.* Our prior work has shown that fitness in bacteriophage T7 can be reduced by several engineered designs, and that fitness recovery is often suppressed during even 1000 generations of subsequent adaptation. Here, we will extend those approaches, measuring the fitness effects of non-preferred codons throughout the genome, of genome rearrangements, and of recodings intended to modify RNA secondary structure or promoter knockouts. We will then observe fitness recovery and molecular evolution of recoded viruses for 1000 generations.

Aim 2: Dissect molecular mechanisms of fitness reduction and recovery. *Hypothesis: The various genetic recodings we apply here all result in dysregulated protein abundances, and several impede initiation or elongation rates of either transcription or translation.* Using the viruses from Aim 1, we will assess effects of protein dysregulation by measuring dynamics of viral mRNAs and protein abundances in infected cells. We will assess the impact on translation through ribosome profiling.

Aim 3: Develop a predictive, mechanistic model of how genome recoding affects T7 fitness. *Hypothesis: For phage T7, the experimental findings of Aims 1 and 2 can be integrated into a coherent, mechanistic model of the phage life cycle.* Gene regulation and life cycle of bacteriophage T7 are well understood, and a second-generation mechanistic model exists that describes all stages of a wild-type T7 infection. We will build on this model to develop a predictive model of T7 genome attenuation and evolution.

This project will result in a comprehensive, system's level understanding of T7 gene regulation, transcription, and translation. We will develop the capability to systematically engineer attenuated T7 variants that are resistant to evolutionary reversion. This work will be a first step towards rational design of live vaccines.

Research Strategy

A. SIGNIFICANCE

Viral attenuation has led to many of the most successful vaccines known to medicine (live virus vaccines), and many of them have realized profound success. Yet how to achieve attenuation has been a challenge. It is widely appreciated that reduced viral growth rate is the most common avenue to attenuation—a virus that no longer causes disease. How to get a virus with reduced growth rate has been less obvious [1]. For most of a century, the standard method of attenuation was hit-and-miss: a virus would be adapted to novel conditions, and the resulting adaptation would commonly reduce its ability to grow in the former host. Whether the attenuation would be sufficient to allow infection but avoid disease was unpredictable and required direct experimentation with live hosts.

The landscape for attenuation has vastly improved. Synthetic biology has provided new opportunities for engineering attenuated viruses. In one case, silent codon modification, the magnitude of attenuation is even generalizable across diverse viruses [2–12]. Furthermore, the level of growth rate reduction with silent codon modification has been shown to be quantitatively tied to the number of codon changes [12,13] (**Figure A1**). Another apparently generalizable approach uses rearrangement of the genes in viral genomes to disrupt fitness, but the degree of attenuation is far less predictable than with codon modification [14–17].

Attenuation is not the only goal in creating a safe and effective vaccine. A further issue is the evolutionary stability of an attenuated virus. Since attenuated viruses are live, they create ongoing infections in the patient. Disease is avoided because of reduced viral growth rate, but the live viruses can and do evolve during the infection and continue evolving when transmitted to other hosts. Some live vaccines are known to evolve to the point that attenuation is fully reversed (polio virus). Reversal of attenuation is especially serious when we are trying eradicate a virus or when trying to prevent its invasion into the human population—in either case, the vaccine runs the risk of evolving and creating the problem that it was engineered to solve. Thus the live Sabin or oral polio vaccine has brought us to the brink of global eradication, but the remaining areas of viral endemism have such poor vaccine coverage that they allow the vaccine to evolve and start new epidemics [18]. Wild-type type II polio virus has in fact been eradicated worldwide, but a vaccine-derived type II continues to circulate and cause disease.

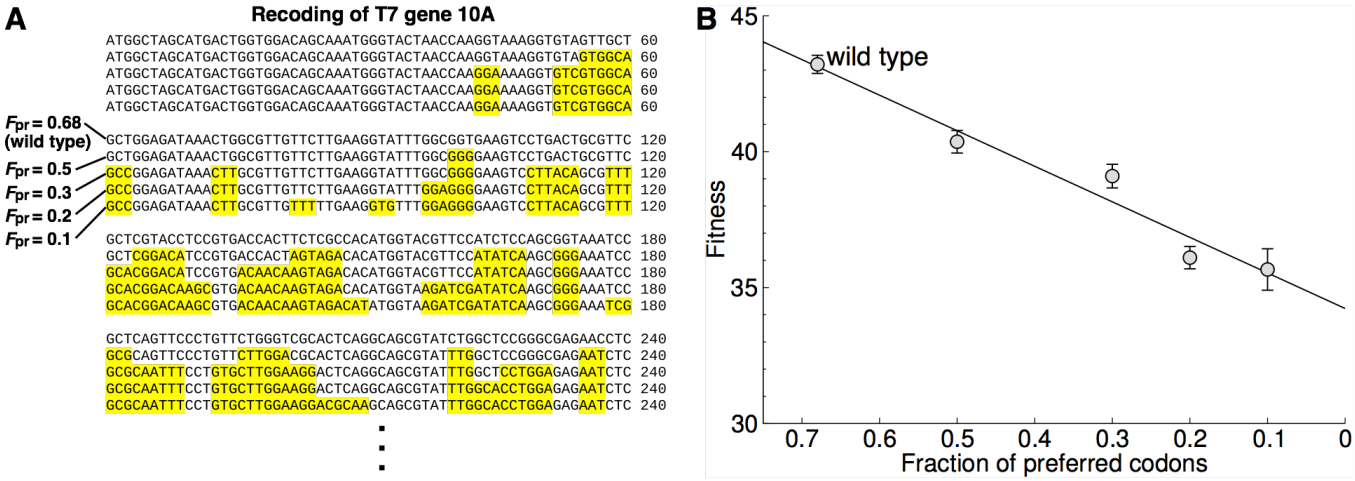

Figure A1: Recoding of T7's major capsid gene 10A with non-preferred codons causes fitness decline in proportion to the amount of codon de-optimization achieved. (A) We introduced non-preferred codons (highlighted in yellow) into the wild-type form of gene 10A until a given overall fraction of preferred codons (F_{pr}) was obtained. Note that the wt is predominantly encoded with preferred codons and has F_{pr} = 0.68. (B) We found that fitness of the recoded variants declined approximately linearly with the number of non-preferred codons introduced. Note that fitness is measured in units of log(2), so a reduction by 10 units corresponds to an approximately thousand-fold reduction in the number of offspring produced. From [13].

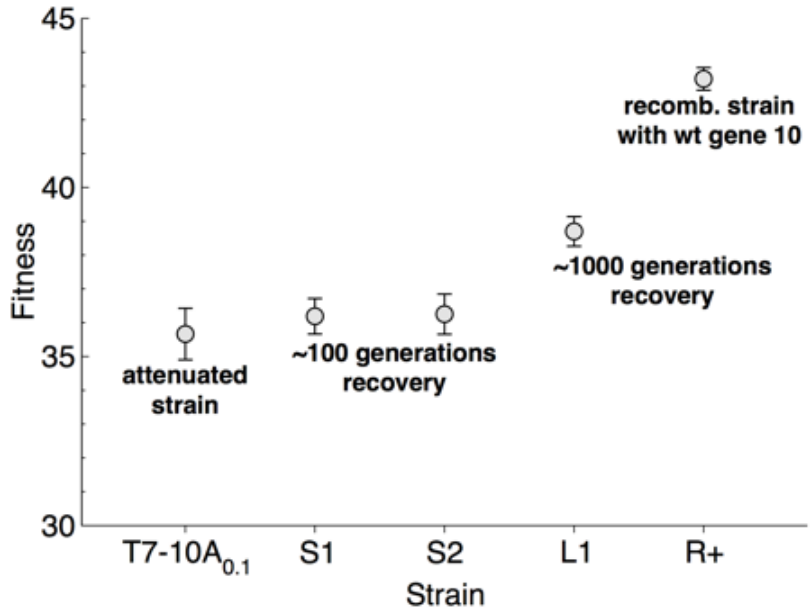

Figure A2: Limited fitness recovery even after ~1000 generations of evolution. We studied fitness recovery of the attenuated strain T7-10$A_{0.1}$, which was the most de-optimized strain we constructed, with F_{pr} = 0.1. Two replicates (S1 and S2) were allowed to adapt for ~100 generations, and showed nearly no fitness increase. One replicate (L1) was allowed to adapt for ~1000 generations, and showed moderate fitness increase. Replicate R+ adapted for ~30 generations, but in the presence of wt gene 10 on a plasmid. This gene was taken up by the phage, which subsequently attained wt fitness, thus demonstrating that complete recovery would be possible if all required mutations could occur simultaneously. From [13].

Evolutionary stability of attenuation is directly linked to the fitness landscape in which the attenuated viral strain resides. If much of the attenuation can be reversed with just a few mutations, then live vaccines are likely to revert. Therefore evolutionarily stable attenuation has to be created with an appreciation of how to block adaptation. One evolution-retarding approach uses a strategy of "death by a thousand cuts": if many mutations with individually small effects are used to attenuate, then reversion may take many hundreds or thousands of generations, since individual mutations of small effect are slow to fix in the viral population. We have demonstrated this mechanism in bacteriophage T7 by codon de-optimizing a single gene via the introduction of 182 synonymous mutations; we showed that we could achieve a hundred-fold reduction in fitness (number of descendants per hour) that was resistant to reversion (**Figure A2**). A second approach is to attenuate by 'irreversible' genetic changes, such as deletions and genome rearrangements. In all cases, however, the presumption of irreversibility rests on a necessarily imperfect knowledge of the viral intracellular dynamics and possible escape mechanisms. Improving that imperfect knowledge is a focus here.

The enterprise of biotechnology is on the brink of being able to engineer predictably attenuated viruses with ease; indeed we may have attained that ability with some methods already. Furthermore, we are poised to create vaccines that do not revert to high virulence. The latter challenge is the more difficult, however, because it requires an understanding of evolutionary mechanisms in response to engineering—which is a frontier in synthetic biology. The work proposed here will develop that frontier. A model virus (bacteriophage T7) will be engineered, evolved, analyzed at the sequence, transcript, and proteomics levels, all interpreted with a viral virtual model. T7 is unusual among viruses in that it encodes its own RNA polymerase, so its gene expression is amenable to quantitative understanding [19,20]. As was true of the first-generation T7 model, the second-generation model of the T7 life cycle (TABASCO) remains the only viral model parameterized empirically. Here, this model will both serve as a foundation for interpreting results and as a basis for further model development. The end product should be a widely generalizable understanding of viral attenuation and evolutionary stability.

B. INNOVATION

There are two ways in which this work is novel. First is the proteomic and RNA analysis of viral evolution and fitness recovery. To date, studies of recovery from codon-based attenuation have been limited to fitness and sequence analysis [12,13]. The inclusion of protein expression and RNA expression adds an essential component of the genotype-phenotype map that is perhaps most critical to fitness, and indeed, this type of analysis has been done to study the bases of attenuation [2,10–12]. The T7 life cycle is linear, with genome injection, ordered gene expression, and ultimately progeny production before dissolving the cell wall at lysis. The resolu-

tion offered by MassSpec proteomics enables a coupling of genome-wide protein expression with fitness and changes in DNA sequence. We can thus ask whether a favored sequence change moves the life cycle closer to wild-type balance of protein expression. If the usual outcome of fitness recovery is to restore wild-type protein levels, it then becomes possible to design attenuation strategies that permanently disrupt that balance (e.g., by removing critical promoters). Indeed, the very analysis of protein levels enables us to study the effects of different attenuation strategies at a mechanistic level.

Second, the proposed work is highly integrative, spanning proteomes and transcriptomes, computational modeling and prediction, ultimately tied to fitness and evolution. Varied engineered modifications of viral genomes will be analyzed at the level of fitness for their immediate effects and the ability of the virus to evolve recovery. Fitness will be tied to mRNA and protein expression levels as well as and ribosome occupancy on transcripts. The results will be integrated in a close handshake with a virtual model of the life cycle, a model that already includes transcription and translation. This work will lead to a 3rd generation virtual model, one that can be used to predict the effects of genome manipulations and evolution.

Why a bacterial virus? For a long term goal of improving attenuated vaccines, it may seem ill advised to use a non-pathogenic virus, especially a virus that infects bacteria. Indeed, if the goals of this work could be achieved in the same time frame with a virus that infects humans or even other mammals, we would readily accept that such work warrants a higher priority than ours. The reality is that no other virus – even another phage—has the foundation for understanding viral dynamics and viral evolution that T7 offers. The second-generation virtual model of the T7 life cycle (TABSCSO) offers unprecedented detail for interpreting and predicting viral attenuation and evolution. It remains the only viral virtual model parameterized empirically. Our platform of genome design, synthesis, evolution and prediction/interpretation is unparalleled, and the ease and safety of T7 work enables rapid progress on all fronts. Even the proteomic work is vastly simpler and more repeatable in our prokaryotic system than in a eukaryotic system. Furthermore, the fact that several attenuation methods tested here have been demonstrated to work for eukaryotic viruses (silent codon modification, genome rearrangement) suggests that our results will generalize. If T7 proves highly predictable, the work will inspire parallel attempts with other viruses.

C. APPROACH

Prior work. This project is a continuation of R01 GM088344, *The Biophysical Basis of Translational Selection*, 08/01/2009 – 05/31/2015. The prior project has resulted in 29 publications to date [13,21–48]. The primary aims of this project were to identify the selective forces that shape codon-usage bias, to test experimentally whether selection acts against protein misfolding, and to investigate how protein biophysics shape codon usage bias. Notable results from this project have been the discovery of a universal trend of reduced mRNA secondary-structure stability near the start codon, for both cellular organisms [21,37] and viruses [38], an experimental demonstration of the fitness cost of protein misfolding [34], the discovery that codon usage correlates with specific features in protein structure [44], and, most important for this application, a demonstration that codon de-optimization leads to substantial and evolutionarily stable attenuation in bacteriophage T7 [13] (see also **Figures A1** and **A2**).

Here, we will build on these prior efforts and will apply our insights into translational selection to the problem of evolutionarily stable viral attenuation.

Team. We have assembled a diverse team of experienced scientists with a history of collaboration. Jim Bull has over two decades of experience with experimental evolution of phages. He will lead Aim 1. Dan Boutz has over a decade of experience with molecular biology and proteomics. He will lead Aim 2. Claus Wilke has 15 years of experience in computational biology, and was PI on the previous grant. He will lead Aim 3 and direct the overall project.

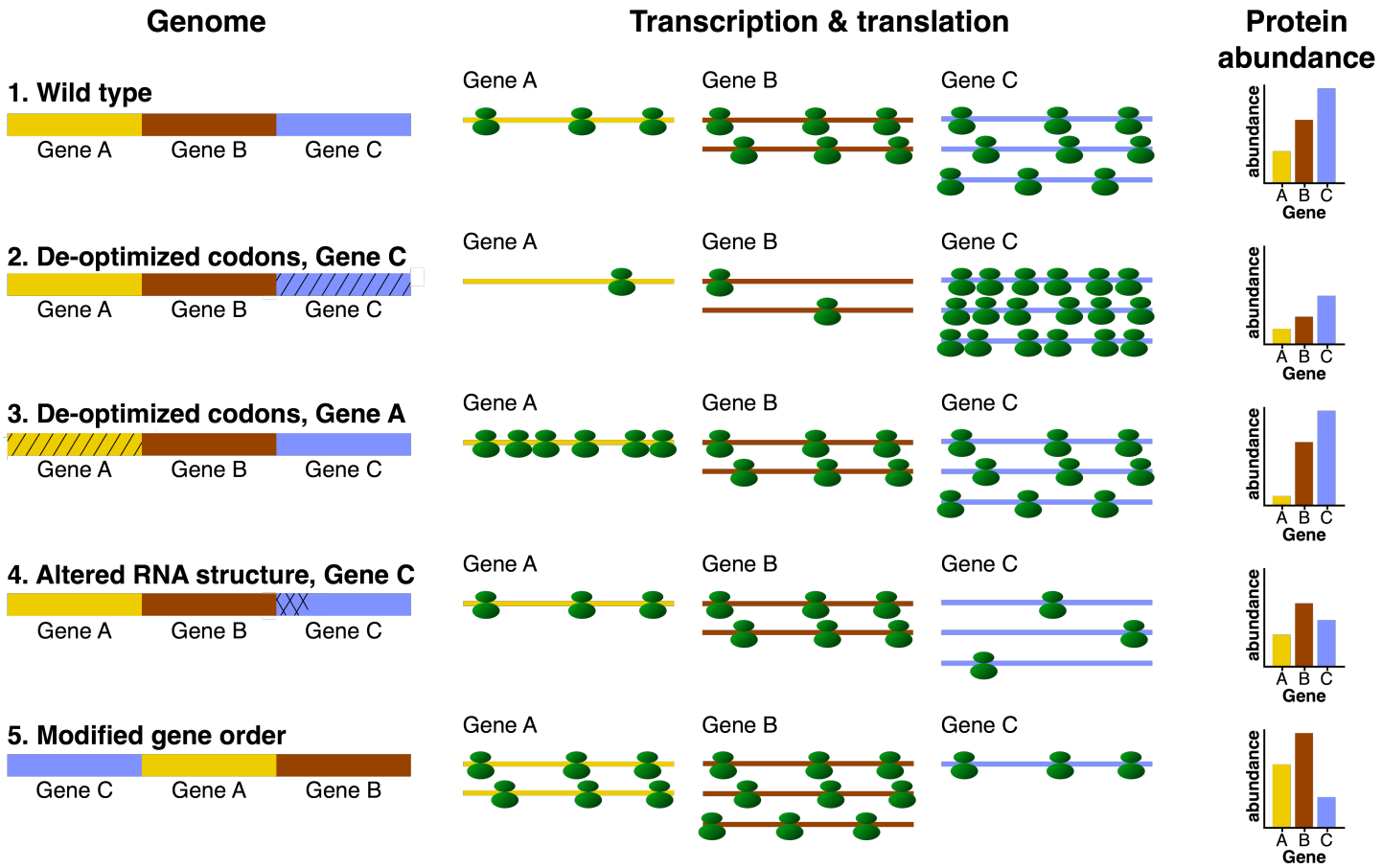

Figure C1: Schematic drawing of genome recodings considered and their likely effects on transcription, translation, and resulting protein abundance. 1. mRNA expression in T7 is approximately proportional to gene order, such that later genes show higher expression levels than earlier genes. 2. De-optimizing a late, highly expressed gene sequesters ribosomes and causes overall reduction in protein abundance, even for genes that weren't recoded. 3. De-optimizing an early gene of low expression level impedes translation of that gene but has little effect on other genes. 4. Changing the secondary structure of mRNA near the start codon reduces translation initiation for that gene, resulting in lowered protein abundance for that gene. 5. Changing the gene order affects mRNA expression levels and consequently protein abundances.

Conceptual framework. Our central hypothesis is that reliable, evolutionarily stable attenuation of viruses can be achieved through targeted genome recodings that interfere with efficient transcription and translation in ways that are not easily reversed by evolution. **Figure C1** illustrates several recodings we will consider, their expected effect on transcription and translation, and resulting protein abundances. In this project, we will construct these recoded genomes for bacteriophage T7 (**Aim 1**), measure their effects on transcription, translation, and protein abundance (**Aim 2**), and develop a calibrated model of gene expression and regulation for T7 (**Aim 3**).

Aim 1: Assess fitness effects and resistance to recovery of different genetic recodings of bacteriophage T7

The methodology used in this aim is straightforward and has been developed extensively in prior work: genomes are created, fitness is assessed (measured as viral growth rate in defined conditions), the genomes are evolved in a constant environment (hosts are replaced to avoid co-evolution), and fitness is measured as the evolution continues. Genome sequences of isolates and populations are now trivially obtained [17,49]. Although this aim is based heavily on methods used in prior work, it provides the foundation for the analyses in

Aims 2 and 3, which add fundamentally new dimensions to work that has gone before. Furthermore, the genomes created and evolved in Aim1 will provide the genomic material for the analyses in Aims 2 and 3. The motivation for particular designs is offered below.

Aim 1.1: Change synonymous codon usage throughout the genome. In our prior work, we have changed codon usage in gene *10A* only (**Figures A1** and **A2**). Here, we will expand this work to three other genes located throughout the T7 genome.

T7 is a dsDNA virus of 40kb with nearly 60 genes encoded on one strand of its linear genome; there are few gene overlaps. Approximately 20 of those genes are essential under standard laboratory conditions. Our prior work on codon de-optimization modified only a single gene in T7, the major capsid gene *10A*. This gene is the most highly expressed of all T7 genes. Using the codon usages of the *E. coli* host to classify codons as 'preferred' or not, the fraction of preferred codons in wild-type gene *10A* is 0.68 (of 345 codons, excluding the stop). We constructed genes in which that fraction was 0.5, 0.3, 0.2 and 0.1; the 5' and 3' ends of the gene were avoided. These modified genes were recombined into a common genomic backbone. The most extreme modification had 182 changed codons. Fitness, measured as doublings/hr, declined linearly with number of codon changes, from 43.2 for wild-type to 35.7 for the most extreme engineering. Thus the maximal attenuation was not extreme, but the effect was predictable. Evolution of the most extreme genome yielded no detectable fitness increase during approximately 100 generations, but nearly half the fitness was recovered in 1000 generations of adaptation. The population from this latter adaptation exhibited 9 nucleotide changes polymorphic, 7 of them outside the codon-modified region [13].

Plan. The chief questions arising from this work are: (i) How much attenuation is achieved when codons are modified in single genes whose wild-type expression is lower than that of *10A*? (ii) How will viral attenuation behave when changes are spread over multiple genes? (iii) Can evolutionary recovery be suppressed even further by spreading codon modifications across multiple genes? (iv) Which parts of genes offer the most effective paths to attenuation by recoding?

For comparison to our prior work on *10A*, we will de-optimize codons in 3 genes: RNA polymerase (gene *1*), DNA polymerase (gene *5*), and an interval virion protein gene (gene *16*), these genes representing early, middle, and late genes spread across the genome (see below for meaning of early, middle and late). De-optimization will be introduced to the same level as in the most extreme *10A* modification (10% preferred codons remaining). Levels of attenuation will be compared for T7 genomes with single-gene modifications, doubles, and all 4 (including gene *10*, from the previous study). Adaptations of the singles, doubles and the quadruple will be carried out as in the previous study to monitor fitness recovery and sequence changes. We particularly wish to know whether recovery rate shows any pattern with the number of genes modified.

We will adopt two distinct engineering strategies. First, we will de-optimize extensively the middle of genes and avoid modifying the 5' and 3' ends (as in our previous study). For this work, non-preferred codons will be introduced to the same level as in the most extreme previous *10A* modification (10% preferred codons remaining). Second, we will limit codon modification to the 5' ends, but introduce synonymous codons that are predicted to form stable RNA secondary structures. This approach is motivated by experimental work showing reduced expression of genes with more stable 5' mRNA secondary structure in *E. coli* [50,51], and by our computational work demonstrating selection for reduced 5' mRNA secondary structure in cellular organisms [21,37] as well as viruses [38]. The utility of the latter approach is simplicity (few codons need be changed), but it may also be prone to easy reversion.

Aim 1.2: Change gene order in the phage genome. As an alternative to extensive codon de-optimization, we will also change the order of the genes in the phage genome. Gene expression in T7 is well understood because the virus encodes its own RNA polymerase and has phage-specific promoters spread across much of its genome. The genome is linear, and the 5' end of the positive strand enters the cell first. The genome is divided

into early, middle, and late regions, these regions defined according to expression properties. The host RNA polymerase is responsible for expression of the early region, which spans several non-essential genes and the first essential gene, phage RNA polymerase (gene *1*). Once the phage RNA polymerase (RNAP) is made, it drives expression of middle and late regions from the 17 phage-specific promoters. Transcript overlap is extensive, however, with nearly all genes present on transcripts initiated at multiple promoters.

Gene order is important to expression both via juxtaposition to different promoters and because the phage RNAP becomes modified during the phage life cycle so that its activity shifts in favor of late promoters as the infection progresses. Disrupting gene order is thus expected to reduce fitness by creating imbalances in protein abundances. Furthermore, since the wild-type gene order is not easily re-evolved in a genome with altered gene order, this method of attenuation should be largely irreversible. These expectations have been borne out in two studies of T7 [17,52]. So far, only limited rearrangements have been generated, chiefly those with displaced RNAP genes (**Figure C2**).

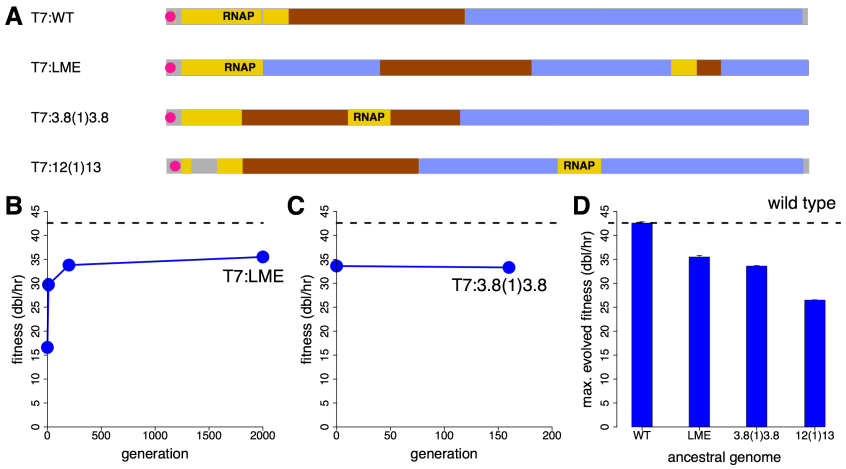

Figure C2: T7 variants with rearranged genomes show limited fitness recovery. (A) Several T7 variants considered. (B, C) Over 100 generations of adaptation are not sufficient to recover wt-level fitness. (D) Maximal attainable fitness for different variants. From [53].

Plan. The T7 phages with altered gene order that were studied in previous work (see [53,54] and **Figure C2**) will provide the initial material for the analyses in Aims 2 and 3. As most of those rearrangements were simple displacements of the phage RNAP gene, we expect the proteomic analyses to reveal the obvious: The phage life cycle is delayed because middle and late phage genes are slow to be expressed, but once the phage RNAP is expressed, the life cycle should be normal. However, of greater interest are the evolved phages that retained the engineered gene order but increased in fitness. What changes in gene expression are coupled with the fitness increase? Genome sequences of the evolved phages do not reveal any insight; we expect the proteomic analysis to be highly informative. In addition, there is one phage genome that exchanged early and late genes; the driving question here is whether the gene expression patterns match the (straightforward) expectations. Do expression patterns of translocated genes follow those of the region in which they are located?

Beyond existing phages with reordered genomes, we will construct new genomes with specific motivations: (i) We will move genes between early, middle, and late regions. (ii) We will reorder genes within functional modules (e.g., scaffolding and major capsid genes, which are typically juxtaposed in phage genomes, will be separated).

The motivation for these types of modifications is that gene order is often highly conserved in viruses (at least for major classes of genes). Thus a finding that these types of rearrangements usually attenuate and whether those attenuations also are evolutionarily stable should be broadly useful. The combination of proteomics, transcriptomics, ribosome profiling, and virtual model analysis done here should give insight to the mechanisms of attenuation and recovery that cannot be obtained by fitness measures and sequence analysis.

Aim 1.3: Introduce promoter knockouts. The foundation for this Aim lies in well-studied mechanisms of T7 transcription. Can we attenuate by relatively simple reductions in transcription? And are those mechanisms evolutionarily stable? This subaim is the simplest of the 3, and it is also the one most easily integrated into Aims 2 and 3. The data are especially amenable to interpretation in the virtual model.

T7 has 17 promoters, not all identical [55]. The T7 promoter sequence is considered to span 23 bases, and the bases important to recognition are worked out from bioinformatics and *in vitro* transcription [56,57]. Likewise, the importance of different T7 promoters to the transcriptome has been worked out for decades [55]. However, fitness effects of specific promoter knockouts have not been examined. We conducted one evolutionary study related to T7 gene regulation via wholesale promoter changes: a T7 phage deleted of its RNAP gene was forced to grow under control of the T3 RNAP gene [58]. At least *in vitro*, T3 RNAP activity on T7 promoters is about 1% that of T7 RNAP; this low activity stems from 1-2 bases in the promoter sequence, so evolution can readily improve expression by T3 RNAP. Initial fitness was low but recovered profoundly (from a fitness of ~5 doublings/hr to a final fitness of ~33; the presumed upper limit was 37, so the adaptation likely neared the maximum). Sequence evolution was seen in approximately half the T7 promoters. The approach used in that study gives some insight to T7 regulatory evolution but it does not directly address what is proposed here, which is to modify selected promoters so that they cannot re-evolve activity.

Plan. We will knock out single and multiple promoters, investigating the fitness effects and changes in transcription and translation (Aim 2), comparing both to virtual model predictions (Aim 3). Knock-outs will maintain sequence length but destroy recognition by the RNAP. Late promoters immediately upstream of the major capsid gene (phi-*10*) and scaffold protein gene (phi-*9*) are expected to have the largest effects. We will further ask whether promoter knockouts with the largest attenuating effects correspond to the promoters that evolved in the study of T7 grown on T3 RNAP [58]. This simple study will thus integrate the virtual model and a prior study in a test of whether the T7 regulatory network is well understood.

Modified viruses will be adapted for hundreds of generations. Given the high sequence-specificity of T7 RNAP for its promoters, is not expected that promoter activity will evolve in situ at ablated promoters. However, duplications of existing promoters may evolve to compensate [e.g., 59], although it is not expected that duplications of existing promoters will compensate for imbalances in gene expression. Thus, while there may be compensatory evolution at a molecular level, it may have only modest fitness effects.

Expected results, Aim 1. We have conducted enough studies of the types proposed in Aim 1 that we can anticipate the general classes of outcomes. First, most modifications of the T7 genome will reduce fitness but not eliminate the ability of T7 to grow; under ideal conditions T7 grows at 42-43 doublings/hr, which offers a huge dynamic range for fitness reduction without killing the virus. Second, many of the modifications proposed here will be robust to large magnitudes of compensatory evolution; we expect some fitness improvement but not much. This latter prediction may fail, but improving our ability to predict such outcomes is part of the motivation for our work. Third, it will be straightforward to observe sequence evolution of attenuated viruses but—in the absence of the proteomics and virtual model—it may be difficult or impossible to interpret how those changes improve fitness. Developing a mechanistic understanding of relationship between sequence changes and fitness is part of our goal and motivates Aims 2 and 3.

Potential problems and solutions, Aim 1. Given that the methods used in Aim 1 mirror those used successfully by us in the past, we do not anticipate difficulties with Aim 1 *per se*—constructs, adaptations and sequencing. Potential difficulties in integrating Aim 1 with Aims 2 and 3 will be discussed below.

Aim 2: Dissect mechanisms of fitness reduction

A major goal of this project is to develop a mechanistic understanding of attenuation mechanisms. Aim 1 develops the genomes and fitness measures. Aim 2 establishes the components and timing of the viral intracellular life cycle. Since T7 infects a bacterium (which is single-celled), and progeny production is essentially a matter of the virus producing a small number of protein (and DNA) components that self-assemble, establishing a mechanistic basis of fitness is feasible in this system. Our approach is two-pronged, going after the time course and abundance of proteins as well as the dynamics of viral RNA and ribosomal occupancy of that RNA.

Aim 2.1: Measurement of viral and cellular protein abundances. Analysis of phage proteomics rarely extends beyond the identification of structural proteins composing the virion [60,61], while the dynamics of phage protein expression during infection remain largely uncharacterized. A recent proteomic analysis of phage 2972 infection in *S. thermophilus* demonstrated the power of the shotgun proteomics approach by quantitatively profiling 37 of 40 predicted phage proteins and nearly 50% of host proteins [62] over the course of infection. While this prior work has clearly demonstrated that shotgun proteomics is a viable approach to dissecting phage biology, this technique has not previously been applied to T7 or any other well characterized phage that might be used for our goals.

We have begun carrying out shotgun proteomics, profiling the expression of phage proteins over the course of infection and for both wt and modified T7 genomes. At this preliminary stage, we hoped to identify the functional impact of codon de-optimization, so the assay included three published genomes from our study of the deoptimized major capsid gene: the wild-type control, the most severely deoptimized phage (182 silent codon changes) and the deoptimized genome that had recovered about half of the lost fitness [13].

Figure C3 shows abundances of T7 proteins as well as of *E. coli* proteins at 3 times after infection (see also **Figures A1** and **A2** for corresponding fitnesses of the 3 viruses). While we have only modified codon usage in gene *10A* for this study, we see systematic reductions in the expression of most phage proteins in the deoptimized genome (**Figure C3** top two rows) but not host proteins (**Figure C3** bottom two rows). This finding suggests ribosome sequestration as the main cause of fitness attenuation—the ribosomes get stuck translating gene *10A* and thus are less available for other transcripts. We can also see the temporal progression of T7 protein abundance. Production of gene *10A* starts immediately upon infection, but the gene's abundance rises throughout the entire infection cycle. By contrast, the tail A and tail fiber proteins are produced only late in the infection cycle. Finally, the DNA and RNA polymerases reach their maximum abundances around 5 min, before most other T7 proteins are maximally expressed.

Plan. We will systematically measure viral protein abundances from the T7 phage genomes generated in Aim 1. Samples of phage-infected *E. coli* will be collected in triplicate at three time-points (1min, 5min, 9min) post infection, so that the analysis of a single T7 variant requires 9 samples. Samples will be analyzed by nanoLC-MS/MS on an in-house LTQ-Oribtrap mass spectrometer (standard bottom-up, shotgun proteomics experiment). We will analyze MS data using Proteome Discoverer v1.4 (Thermo Scientific) to obtain spectral counts, which represent protein abundances.

Aim 2.2: Measurement of mRNA abundance by RNA-seq. While codon de-optimization of genes is not expected to affect transcript abundance, promoter ablation or gene rearrangements likely result in either increased or decreased transcript levels relative to wt. We will measure these changes in transcript levels using whole-genome RNA-seq. In addition to revealing any changes in T7 transcription, as in the case of the proteomics, our approach will also reveal if phage infection has measurable effects on an *E. coli* transcripts.

Plan. For all samples collected for proteomics in Aim 2.1, we will collect duplicate samples (in triplicate at each time point) for profiling mRNA abundances by RNA-seq. RNAseq of *E. coli* samples is routinely done in our laboratories, and the entire process is outsourced to the Genome Sequencing and Analysis Facility (GSAF) at UT Austin. We submit pelleted cells to the GSAF and receive in return BAM files containing raw reads. These reads are then processed with FLEXBAR [63], aligned to the *E. coli* and T7 reference genomes with Bowtie 2 [64], and analyzed for differential expression among samples with DESeq [65].

Aim 2.3: Measurement of translational efficiency by ribosome profiling. Ribosome profiling is a relatively new method which generates high-resolution maps of ribosome density on active transcripts, thus providing a detailed view of translational efficiency [66]. Although this method is only a few years old, it has already been

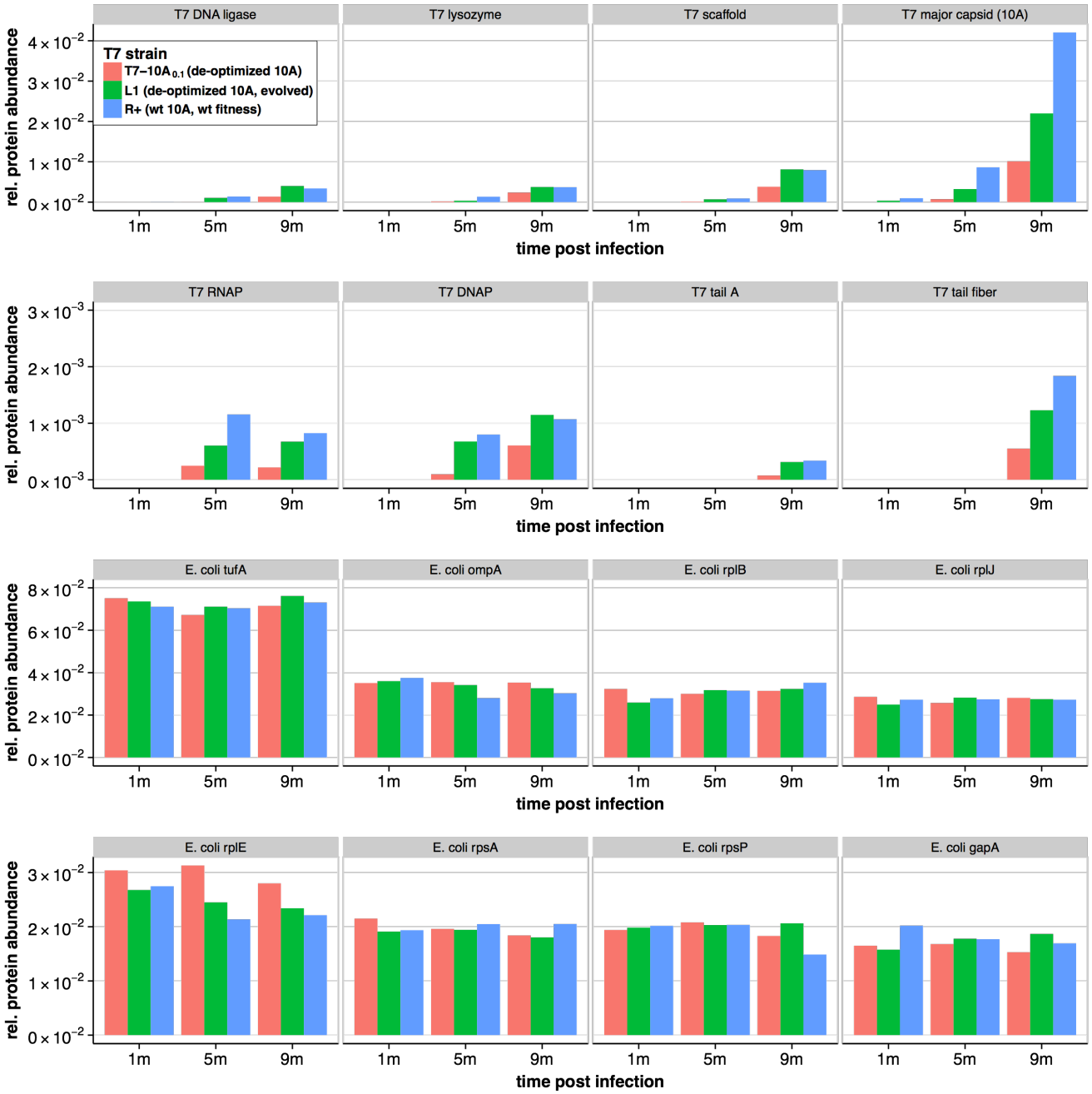

Figure C3: Relative protein abundances for selected T7 and *E. coli* genes, measured at 3 time points post infection and for 3 different infecting T7 strains each. Top 2 rows: T7 protein abundances generally increase over time, and protein abundances are commensurate with strain fitness (see **Figure A2** for strain fitnesses). Bottom 2 rows: Shown are the 8 most-highly expressed proteins in *E. coli*. Relative abundances are generally of comparable magnitude regardless of time point and infecting T7 strain. This pattern holds throughout the *E. coli* proteome.

utilized to study translation of bacteriophage lambda [67] and cytomegalovirus [68] genes during infection, along with multiple studies of *E. coli* [69,70].

Plan. We will follow the standard protocol for ribosomal profiling in *E. coli* [69,70]. In brief, we will infect *E. coli* bacteria with wt or modified T7 phage, harvest after ~9min, filter, and flash freeze. We will then add GMPPNP and chloramphenicol, pulverize cells in liquid nitrogen, digest RNA for 1h with micrococcal nuclease (MNase), and isolate ribosome-protected fragments by sucrose gradient and phenol extraction. From these fragments we will generate a cDNA library and sequence. The reads are then mapped onto the *E. coli* and T7 genomes as under Aim 2.2, and the mapped profile indicates ribosome occupancy.

Expected results, Aim 2. This Aim will generate the data used to parameterize and improve the mechanistic model studied in Aim 3. However, the RNA and proteomics data will first be used to answer fundamental questions about the mechanistic bases of attenuation. We have clear expectations for some of the genomic modifications (see also **Figure C1)**. (1) Promoter ablation should have its effect on protein levels through decreases in transcript abundance. (2) Genome rearrangement should cause imbalances in the relative abundances of proteins. (3) We should be able to discriminate the many models for the effects of codon de-optimization. For example, codon de-optimized highly expressed genes may sequester ribosomes and indirectly affect expression of all T7 genes. By contrast, de-optimized genes with low expression level will not have such global effects. Altered RNA secondary structure will affect only translation initiation. By measuring protein and transcript abundances as well as ribosome profiles we will be able to dissect the various possibilities and identify the specific mechanisms that caused dysregulation for specific T7 variants. These insights will be useful independent of the work done in Aim 3.

Finally, we do not expect to see changes in protein or transcript abundances among the *E. coli* transcripts/genes, but we will analyze all samples for this possibility. Finding differentially expressed *E. coli* genes would likely indicate novel T7 biology that has not previously been recognized.

Potential problems and solutions, Aim 2. We do not expect to encounter any major issues with the proteomics, since we have already successfully carried out protein abundance measurements on T7-infected cells (**Figure C3**). Importantly, the vast majority of *E. coli* proteins showed identical or near identical abundances among conditions, highlighting the reproducibility of our measurements **Figure C3**, bottom two rows.) We similarly do not expect to encounter major issues with RNAseq. Our sequencing facility routinely processes *E. coli* samples. Moreover, Wilke is involved in other collaborations doing RNAseq and has developed extensive experience handling the data analysis side of these samples. Ribosome profiling, by contrast, is new to us, even though some collaborators at UT have carried it out successfully. We consider this part the high-risk aspect of the proposal, with the potential to produce a wealth of new insight. Importantly, our overall project success does not crucially depend on ribosome profiling data. Should these data prove to be unreliable or difficult to obtain, we can proceed with Aim 3 on the basis of protein and transcript abundances alone.

Aim 3: Develop a predictive, mechanistic model of how genome recoding affects T7 fitness

There has been extensive interest in the recent literature in disentangling how exactly transcription and translation are affected by different gene encodings, and in particular by codon usage (see e.g. [21,50,51,71–81]). This literature has considered a bewildering array of different hypotheses, including reduced translation speed or accuracy, depletion of cellular pools of polymerases, ribosomes, or tRNAs, traffic jams among polymerases or ribosomes, and modified translation initiation due to RNA secondary structure. What is missing from the current literature is a systems-level approach to the problem, where all these different possibilities are evaluated in a single, coherent framework.

In this context, T7 provides a unique model system for studying transcription and translation. T7's biology is extremely well characterized [55,82–85], and several generations of models describing T7's gene regulation and life cycle have been developed [20,84,85]. While the first models of T7 were simple ODE models of gene expression [84,85], the most recent and most sophisticated computational model for T7, called TABASCO, provides a stochastic simulation of T7 gene expression and translation inside the *E. coli* cell [20]. We will build on this model to develop a comprehensive, system's level mechanistic model of how genome recodings affect T7 replication and fitness.

Aim 3.1: Calibrate TABASCO simulation for wild-type and recoded T7 genomes. The TABASCO simulator [20] provides a nucleotide-level simulation of polymerase movements along the T7 genome, and it simulates subsequent translation via a simplified translation model. The source code for TABASCO is freely available, and the Wilke lab has successfully run the model for T7 wt as well as for several simulated genome modi-

fications **(Figure C4).** Importantly, the 2007 TABASCO paper [20] focused on introducing the novel simulation techniques in Tabasco, not on parameter optimization to obtain the best possible T7 simulator. And since publication of this paper, no further research on this topic has been carried out. Therefore, we will begin our modeling work for this Aim with a careful calibration of the Tabasco simulation for wild-type and recoded T7 genomes.

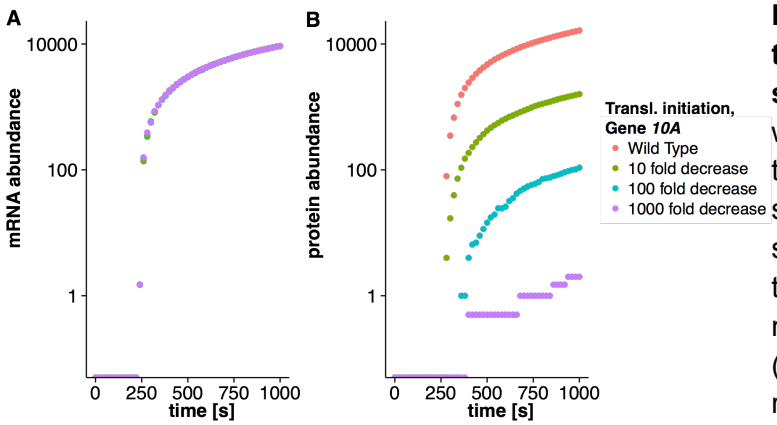

Figure C4: Predicted effect of modified translation initiation for gene *10A*, using the TABASCO simulation. We simulated four different scenarios, wt and three different levels of reduction for translation initiation. This simulation mimics expected results for recoded RNA secondary structure near the start codon (Aim 1.1). The simulation corresponds to the first 16 min (~1000s) of a T7 infection. (A) mRNA abundances are unchanged, as expected. (B) Protein abundances decline in proportion to the reduced efficiency of translation initiation.

Plan. We will initially consider T7 wt only. We will fit the TABASCO simulation to all measured RNA and protein abundances for wt T7. This fitting procedure entails choosing the optimal rate constants for transcription initiation, transcription elongation, translation initiation, translation elongation, etc., such that simulated RNA and protein abundances fall as closely as possible to the values experimentally measured under Aim 2. We will perform the fit using Approximate Bayesian Computation [86,87], an iterative strategy that is particularly well-suited to fit complex simulation models to measured data.

Once we have obtained satisfactory results for T7 wt, we will repeat the fitting procedure for recoded genomes. Depending on the type of genome recoding carried out for a particular strain (codon de-optimization, genome rearrangement, RNA secondary-structure modification, promoter ablation), we will follow two distinct strategies. For some modifications, such RNA secondary-structure modifications and promoter ablations, which should alter only translation-initiation rates, we will simply re-fit the model to the new data and obtain new rate constants. For other modifications, such as genome rearrangements, we will have to first modify the model to reflect the rearrangements, e.g. place all genes under the control of the correct promoters. Finally, for codon de-optimizations, we may have to develop a more sophisticated version of TABASCO to capture all measurable aspects of the biology (see also Aim 3.2).

Aim 3.2: Incorporate an explicit model of translation into TABASCO simulation. While TABASCO simulates the transcription process with nucleotide-level accuracy, the translation model in TABASCO is not similarly sophisticated. Instead of simulating the movement of individual ribosomes along individual mRNAs, TABASCO treats translation as a single step and assumes that translation proceeds at a given mean elongation rate; the exact translation time is then drawn from a gamma distribution [88]. This model has three important drawbacks that we will overcome here: First, it does not model differential translation speed due to codon usage bias. Second, it does not model ribosome-ribosome interactions, such as traffic jams. Third, it does not consider effects that happen on the scale of the cell, such as ribosome depletion due to traffic jams on some transcripts.

Plan. We will pursue two separate approaches to developing a more sophisticated translation model. First, we will continue treating translation as a single step but will calculate a gene-specific elongation rate as a function of the codon usage in the gene and the ribosome pool in the cell. This approach will cause virtually no slowdown in the simulation yet already provide much improved model realism. Second, we will simulate individual ribosome movements along individual transcripts, using the same techniques Tabasco currently uses to simu-

late RNA polymerases moving along DNA. This approach will be much slower, but it will allow us to incorporate more complex ribosome dynamics, such as ribosome traffic jams or ribosome sequestration, where slow translation of one gene may inhibit translation of other genes as well.

Aim 3.3: Use TABASCO to predict new attenuations to be tested under Aims 1 and 2. The first-generation computational model of T7 had previously been used to make predictions about the effects of altered gene order in T7 [85], but there was no further work beyond these initial tests. Here, we will use our calibrated and/or modified models from Aims 3.1 and 3.2 to predict novel recodings, and we will subsequently build and test them under Aims 1 and 2.

Plan. We will use the simulation model to design at least one recoded genome each, corresponding to the goals of Aims 1.1 (altered codon usage), 1.2 (changed gene order), and 1.3 (promoter knockout). In each case, we will strive to design a modified phage that shows broad changes in gene expression over many T7 genes, such that we can expect a major fitness reduction that is difficult to undo in just a few subsequent mutations. At the same time, we will only consider recodings in which all essential T7 genes, and in particular the major capsid gene *10A*, are predicted to have non-zero protein abundance, such that the phage is still viable. Our general aim is to produce attenuated but viable T7 variants. Any design that yields an inviable T7 genome does not meet that aim.

Expected results, Aim 3. We expect that we can calibrate TABASCO for the wild-type T7 so that it makes accurate predictions of RNA and protein abundances throughout the T7 life cycle. We further expect that the calibrated model will accurately predict some of the modified genomes but not others. For example, as indicated in **Figure C1**, introduction of non-preferred codons into a highly expressed gene may cause ribosome sequestration, which indirectly affects protein abundances of all viral genes. To capture such effect, we will have to introduce appropriate modifications into TABASCO, as developed under Aim 3.2. In general, we expect that our modeling approach will allow us to develop a system-level understanding of T7 gene regulation, transcription, and translation, and that it will shed light on unexpected gene interactions, such as recoding of one gene affecting protein abundances of others (**Figure C3**). Finally, we expect that our calibrated model has predictive power, i.e., that it allows us to computationally design novel, attenuated genomes (Aim 3.3).

Potential problems and solutions, Aim 3. We do not expect to encounter any prohibitive problems in this Aim. Mathematical models of the T7 life cycle have been used successfully for over 15 years, and they seem to generally work well. However, it is possible that we will encounter specific modified genomes whose gene expression patterns will not be explainable with the current and/or improved models. Such incidences would indicate major missing biology in the model and would require further study and model development.

Time Line

We will work on all three aims in parallel. For Aim 1, we will carry out additional codon de-optimizations (Aim 1.1) in Year 1, and genome rearrangements (Aim 1.2) and promoter knock-outs (Aim 1.3) in Year 2. In Years 3 and 4, we will construct genomes predicted by our computational model (Aim 3.3), and we will pursue additional modifications as prompted by our findings from Aims 1-3 up to that time. For Aim 2, we will measure protein abundances, RNA, and ribosomal occupancy for all genomes as they become available. We will begin with the genomes we currently have, including codon-modified genomes and genomes with rearranged gene order. For Aim 3, we will begin model calibration on the data we have already collected (**Figure C3**), and will continue to improve our models as more data becomes available. We expect to begin working on improved translation models in TABASCO (Aim 3.2) in Year 2 and to be able to predict new attenuations (Aim 3.3) beginning with Year 3.

Bibliography

1. Hanley KA (2011) The double-edged sword: How evolution can make or break a live-attenuated virus vaccine. Evolution 4: 635–643. doi:10.1007/s12052-011-0365-y.

2. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, et al. (2008) Virus attenuation by genome-scale changes in codon pair bias. Science 320: 1784–1787. doi:10.1126/science.1155761.

3. Martrus G, Nevot M, Andres C, Clotet B, Martinez MA (2013) Changes in codon-pair bias of human immunodeficiency virus type 1 have profound effects on virus replication in cell culture. Retrovirology 10: 78. doi:10.1186/1742-4690-10-78.

4. Meng J, Lee S, Hotard AL, Moore ML (2014) Refining the balance of attenuation and immunogenicity of respiratory syncytial virus by targeted codon deoptimization of virulence genes. mBio 5: e01704–e01714. doi:10.1128/mBio.01704-14.

5. Nougairede A, De Fabritus L, Aubry F, Gould EA, Holmes EC, et al. (2013) Random codon re-encoding induces stable reduction of replicative fitness of Chikungunya virus in primate and mosquito cells. PLoS Pathog 9: e1003172. doi:10.1371/journal.ppat.1003172.

6. Pena L, Sutton T, Chockalingam A, Kumar S, Angel M, et al. (2013) Influenza viruses with rearranged genomes as live-attenuated vaccines. J Virol. doi:10.1128/JVI.02490-12.

7. Le Nouën C, Brock LG, Luongo C, McCarty T, Yang L, et al. (2014) Attenuation of human respiratory syncytial virus by genome-scale codon-pair deoptimization. Proc Natl Acad Sci U S A 111: 13169–13174. doi:10.1073/pnas.1411290111.

8. Nogales A, Baker SF, Ortiz-Riaño E, Dewhurst S, Topham DJ, et al. (2014) Influenza A virus attenuation by codon deoptimization of the NS gene for vaccine development. J Virol 88: 10525–10540. doi:10.1128/JVI.01565-14.

9. Ni Y-Y, Zhao Z, Opriessnig T, Subramaniam S, Zhou L, et al. (2014) Computer-aided codon-pairs deoptimization of the major envelope GP5 gene attenuates porcine reproductive and respiratory syndrome virus. Virology 450-451: 132–139. doi:10.1016/j.virol.2013.12.009.

10. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E (2006) Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J Virol 80: 9687–9696. doi:10.1128/JVI.00738-06.

11. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, et al. (2009) Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons. J Virol 83: 9957–9969. doi:10.1128/JVI.00508-09.

12. Burns CC, Shaw J, Campagnoli R, Jorba J, Vincent A, et al. (2006) Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. J Virol 80: 3259–3272. doi:10.1128/JVI.80.7.3259-3272.2006.

13. Bull JJ, Molineux IJ, Wilke CO (2012) Slow fitness recovery in a codon-modified viral genome. Mol Biol Evol 29: 2997–3004. doi:10.1093/molbev/mss119.

14. Ball LA, Pringle CR, Flanagan EB, Perepelitsa VP, Wertz GW (1999) Phenotypic consequences of rearranging the P, M, and G genes of vesicular stomatitis virus. J Virol 73: 4705–4712.

15. Novella IS, Ball LA, Wertz GW (2004) Fitness analyses of vesicular stomatitis strains with rearranged genomes reveal replicative disadvantages. J Virol 78: 9837–9841. doi:10.1128/JVI.78.18.9837-9841.2004.

16. Flanagan EB, Zamparo JM, Ball LA, Rodriguez LL, Wertz GW (2001) Rearrangement of the genes of vesicular stomatitis virus eliminates clinical disease in the natural host: new strategy for vaccine development. J Virol 75: 6107–6114. doi:10.1128/JVI.75.13.6107-6114.2001.

17. Cecchini N, Schmerer M, Molineux IJ, Springman R, Bull JJ (2013) Evolutionarily stable attenuation by genome rearrangement in a virus. G3 Bethesda Md 3: 1389–1397. doi:10.1534/g3.113.006403.

18. Burns CC, Diop OM, Sutter RW, Kew OM (2014) Vaccine-derived polioviruses. J Infect Dis 210 Suppl 1: S283–S293. doi:10.1093/infdis/jiu295.

19. Endy D, You L, Yin J, Molineux IJ (2000) Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. Proc Natl Acad Sci U S A 97: 5375–5380. doi:10.1073/pnas.090101397.

20. Kosuri S, Kelly JR, Endy D (2007) TABASCO: A single molecule, base-pair resolved gene expression simulator. BMC Bioinformatics 8: 480. doi:10.1186/1471-2105-8-480.

21. Gu W, Zhou T, Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol 6: e1000664. doi:10.1371/journal.pcbi.1000664.

22. Wilke CO (2012) Bringing molecules back into molecular evolution. PLoS Comput Biol 8: e1002572. doi:10.1371/journal.pcbi.1002572.

23. Meyer AG, Dawson ET, Wilke CO (2013) Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. Philos Trans R Soc Lond B Biol Sci 368: 20120334. doi:10.1098/rstb.2012.0334.

24. Zhou T, Gu W, Wilke CO (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27: 1912–1922. doi:10.1093/molbev/msq077.

25. Wallace EWJ, Airoldi EM, Drummond DA (2013) Estimating selection on synonymous codon usage from noisy experimental data. Mol Biol Evol 30: 1438–1453. doi:10.1093/molbev/mst051.

26. O'Dea EB, Pepin KM, Lopman BA, Wilke CO (2014) Fitting outbreak models to data from many small norovirus outbreaks. Epidemics 6: 18–29. doi:10.1016/j.epidem.2013.12.002.

27. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ (2013) Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol 30: 549–560. doi:10.1093/molbev/mss273.

28. Drummond DA (2012) How infidelity creates a sticky situation. Mol Cell 48: 663–664. doi:10.1016/j.molcel.2012.11.024.

29. Meyer AG, Wilke CO (2013) Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30: 36–44. doi:10.1093/molbev/mss217.

30. Sedaghat AR, Wilke CO (2011) Kinetics of the viral cycle influence pharmacodynamics of antiretroviral therapy. Biol Direct 6: 42. doi:10.1186/1745-6150-6-42.

31. Spielman SJ, Dawson ET, Wilke CO (2014) Limited Utility of Residue Masking for Positive-Selection Inference. Mol Biol Evol. doi:10.1093/molbev/msu183.

32. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO (2013) Maximum allowed solvent accessibilites of residues in proteins. PloS One 8: e80635. doi:10.1371/journal.pone.0080635.

33. Spielman SJ, Wilke CO (2013) Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. J Mol Evol 76: 172–182. doi:10.1007/s00239-012-9538-8.

34. Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL, et al. (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. Proc Natl Acad Sci U S A 108: 680–685. doi:10.1073/pnas.1017570108.

35. Scherrer MP, Meyer AG, Wilke CO (2012) Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol 12: 179. doi:10.1186/1471-2148-12-179.

36. Geiler-Samerotte KA, Hashimoto T, Dion MF, Budnik BA, Airoldi EM, et al. (2013) Quantifying condition-dependent intracellular protein levels enables high-precision fitness estimates. PloS One 8: e75320. doi:10.1371/journal.pone.0075320.

37. Keller TE, Mis SD, Jia KE, Wilke CO (2012) Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes. Genome Biol Evol 4: 80–88. doi:10.1093/gbe/evr129.

38. Zhou T, Wilke CO (2011) Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses. BMC Evol Biol 11: 59. doi:10.1186/1471-2148-11-59.

39. Wilke CO, Drummond DA (2010) Signatures of protein biophysics in coding sequence evolution. Curr Opin Struct Biol 20: 385–389. doi:10.1016/j.sbi.2010.03.004.

40. Drummond DA, Wilke CO (2009) The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet 10: 715–724. doi:10.1038/nrg2662.

41. Bull JJ, Heineman RH, Wilke CO (2011) The phenotype-fitness map in experimental evolution of phages. PloS One 6: e27796. doi:10.1371/journal.pone.0027796.

42. Ramsey DC, Scherrer MP, Zhou T, Wilke CO (2011) The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188: 479–488. doi:10.1534/genetics.111.128025.

43. Wilke CO (2011) Transcriptional robustness complements nonsense-mediated decay in humans. PLoS Genet 7: e1002296. doi:10.1371/journal.pgen.1002296.

44. Lee Y, Zhou T, Tartaglia GG, Vendruscolo M, Wilke CO (2010) Translationally optimal codons associate with aggregation-prone sites in proteins. Proteomics 10: 4163–4171. doi:10.1002/pmic.201000229.

45. Meyer AG, Sawyer SL, Ellington AD, Wilke CO (2014) Analyzing machupo virus-receptor binding by molecular dynamics simulations. PeerJ 2: e266. doi:10.7717/peerj.266.

46. Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, et al. (2014) Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. J Mol Evol 79: 130–142. doi:10.1007/s00239-014-9644-x.

47. Tien MZ, Sydykova DK, Meyer AG, Wilke CO (2013) PeptideBuilder: A simple Python library to generate model peptides. PeerJ 1: e80. doi:10.7717/peerj.80.

48. Jackson EL, Ollikainen N, Covert AW, Kortemme T, Wilke CO (2013) Amino-acid site variability among natural and designed proteins. PeerJ 1: e211. doi:10.7717/peerj.211.

49. Paff ML, Stolte SP, Bull JJ (2014) Lethal mutagenesis failure may augment viral adaptation. Mol Biol Evol 31: 96–105. doi:10.1093/molbev/mst173.

50. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. Science 324: 255–258. doi:10.1126/science.1170160.

51. Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. Science 342: 475–479. doi:10.1126/science.1241934.

52. Springman R, Kapadia-Desai DS, Molineux IJ, Bull JJ (2012) Evolutionary recovery of a recombinant viral genome. G3 Bethesda Md 2: 825–830. doi:10.1534/g3.112.002758.

53. Cecchini N, Schmerer M, Molineux IJ, Springman R, Bull JJ (2013) Evolutionarily Stable Attenuation by Genome Rearrangement in a Virus. G3 GenesGenomesGenetics 3: 1389–1397. doi:10.1534/g3.113.006403.

54. Springman R, Badgett MR, Molineux IJ, Bull JJ (2005) Gene order constrains adaptation in bacteriophage T7. Virology 341: 141–152. doi:10.1016/j.virol.2005.07.008.

55. Dunn JJ, Studier FW (1983) Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol 166: 477–535.

56. Imburgio D, Rong M, Ma K, McAllister WT (2000) Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. Biochemistry (Mosc) 39: 10419–10430.

57. Rong M, He B, McAllister WT, Durbin RK (1998) Promoter specificity determinants of T7 RNA polymerase. Proc Natl Acad Sci U S A 95: 515–519.

58. Bull JJ, Springman R, Molineux IJ (2007) Compensatory evolution in response to a novel RNA polymerase: orthologous replacement of a central network gene. Mol Biol Evol 24: 900–908. doi:10.1093/molbev/msm006.

59. Springman R, Molineux IJ, Duong C, Bull RJ, Bull JJ (2012) Evolutionary stability of a refactored phage genome. ACS Synth Biol 1: 425–430. doi:10.1021/sb300040v.

60. Maxwell KL, Frappier L (2007) Viral proteomics. Microbiol Mol Biol Rev MMBR 71: 398–411. doi:10.1128/MMBR.00042-06.

61. Lavigne R, Ceyssens P-J, Robben J (2009) Phage proteomics: applications of mass spectrometry. Methods Mol Biol Clifton NJ 502: 239–251. doi:10.1007/978-1-60327-565-1_14.

62. Young JC, Dill BD, Pan C, Hettich RL, Banfield JF, et al. (2012) Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in Streptococcus thermophilus. PloS One 7: e38077. doi:10.1371/journal.pone.0038077.

63. Dodt M, Roehr JT, Ahmed R, Dieterich C (2012) FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology 1: 895–905. doi:10.3390/biology1030895.

64. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. doi:10.1038/nmeth.1923.

65. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106. doi:10.1186/gb-2010-11-10-r106.

66. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218–223. doi:10.1126/science.1168978.

67. Liu X, Jiang H, Gu Z, Roberts JW (2013) High-resolution view of bacteriophage lambda gene expression by ribosome profiling. Proc Natl Acad Sci U S A 110: 11928–11933. doi:10.1073/pnas.1309739110.

68. Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, et al. (2012) Decoding human cytomegalovirus. Science 338: 1088–1093. doi:10.1126/science.1227919.

69. Li G-W, Oh E, Weissman JS (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484: 538–541. doi:10.1038/nature10965.

70. Oh E, Becker AH, Sandikci A, Huber D, Chaba R, et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell 147: 1295–1308. doi:10.1016/j.cell.2011.10.044.

71. Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. Mol Syst Biol 7: 481. doi:10.1038/msb.2011.14.

72. Pechmann S, Frydman J (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol 20: 237–243. doi:10.1038/nsmb.2466.

73. Shah P, Gilchrist MA (2011) Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc Natl Acad Sci U S A 108: 10231–10236. doi:10.1073/pnas.1016719108.

74. Aragonès L, Guix S, Ribes E, Bosch A, Pintó RM (2010) Fine-Tuning Translation Kinetics Selection as the Driving Force of Codon Usage Bias in the Hepatitis A Virus Capsid. PLoS Pathog 6: e1000797. doi:10.1371/journal.ppat.1000797.

75. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, et al. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141: 344–354. doi:10.1016/j.cell.2010.03.031.

76. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, et al. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12: R110. doi:10.1186/gb-2011-12-11-r110.

77. Reuveni S, Meilijson I, Kupiec M, Ruppin E, Tuller T (2011) Genome-scale analysis of translation elongation with a ribosome flow model. PLoS Comput Biol 7: e1002127. doi:10.1371/journal.pcbi.1002127.

78. Li G-W, Burkhardt D, Gross C, Weissman JS (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624–635. doi:10.1016/j.cell.2014.02.033.

79. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352. doi:10.1016/j.cell.2008.05.042.

80. Ran W, Higgs PG (2012) Contributions of speed and accuracy to translational selection in bacteria. PloS One 7: e51652. doi:10.1371/journal.pone.0051652.

81. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 25: 117–124. doi:10.1038/nbt1270.

82. Studier FW, Dunn JJ (1983) Organization and expression of bacteriophage T7 DNA. Cold Spring Harb Symp Quant Biol 47 Pt 2: 999–1007.

83. Garcia LR, Molineux IJ (1995) Rate of translocation of bacteriophage T7 DNA across the membranes of Escherichia coli. J Bacteriol 177: 4066–4076.

84. Endy D, Kong D, Yin J (1997) Intracellular kinetics of a growing virus: A genetically structured simulation for bacteriophage T7. Biotechnol Bioeng 55: 375–389. doi:10.1002/(SICI)1097-0290(19970720)55:2<375::AID-BIT15>3.0.CO;2-G.

85. Endy D, You L, Yin J, Molineux IJ (2000) Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. Proc Natl Acad Sci 97: 5375–5380. doi:10.1073/pnas.090101397.

86. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, et al. (2013) Approximate Bayesian Computation. PLoS Comput Biol 9: e1002803. doi:10.1371/journal.pcbi.1002803.

87. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface 6: 187–202. doi:10.1098/rsif.2008.0172.

88. Gibson MA, Bruck J (2000) Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. J Phys Chem A 104: 1876–1889. doi:10.1021/jp993732q.