nature biotechnology

REVIEW

Synthetic viruses: a new opportunity to understand and prevent viral disease

Eckard Wimmer[1], Steffen Mueller[1], Terrence M Tumpey[2] & Jeffery K Taubenberger[3]

Rapid progress in DNA synthesis and sequencing is spearheading the deliberate, large-scale genetic alteration of organisms. These new advances in DNA manipulation have been extended to the level of whole-genome synthesis, as evident from the synthesis of poliovirus, from the resurrection of the extinct 1918 strain of influenza virus and of human endogenous retroviruses and from the restructuring of the phage T7 genome. The largest DNA synthesized so far is the 582,970 base pair genome of *Mycoplasma genitalium*, although, as yet, this synthetic DNA has not been 'booted' to life. As genome synthesis is independent of a natural template, it allows modification of the structure and function of a virus's genetic information to an extent that was hitherto impossible. The common goal of this new strategy is to further our understanding of an organism's properties, particularly its pathogenic armory if it causes disease in humans, and to make use of this new information to protect from, or treat, human viral disease. Although only a few applications of virus synthesis have been described as yet, key recent findings have been the resurrection of the 1918 influenza virus and the generation of codon- and codon pair–deoptimized polioviruses.

Unprecedented progress in synthesis and sequence analysis of DNA lies at the heart of the recent transformation of molecular biology and the emergence of the field termed synthetic biology. Sequencing a DNA in the megabase (Mb) range is no longer a daunting undertaking and, applying the most advanced technology, can be accomplished within less than a week. DNA synthesis has not yet advanced to the efficiency of DNA sequencing, but synthesizing DNA of 8–30 kilobase pairs (kbp)—the genome size of most RNA viruses and many DNA viruses—can be accomplished easily and is largely a matter of available resources.

It is not surprising, therefore, that the *de novo* synthesis of viral genomes in the absence of a natural template has found its way into studies of viruses, although this branch of virology is still in its infancy. Chemical synthesis of viral genomes provides a new and powerful tool for studying the function and expression of viral genes, as well as their pathogenic potential. This method is particularly useful if the natural viral template is not available. It also allows the genetic modification of viral genomes on a scale that would be impossible to achieve by conventional molecular biology methods.

In this Review, we summarize briefly the recent advances in DNA synthesis and sequencing and their impact on virology. Specifically, we describe the *de novo* synthesis of viruses in the absence of a natural template with the aim of resurrecting viruses using archaevirology, identifying viruses that cause human diseases after zoonotic infections, reconstructing viral genomes to unravel complex biological systems ('refactoring' genomes) or recoding viral genetic information for the production of vaccine candidates.

Synthetic biology is the design and construction of new biological entities, such as enzymes, genetic circuits and cells, or the redesign of existing biological systems[1]. Such changes exceed those introduced previously into biological systems by methods of classic molecular engineering. Synthetic biology depends on the collaboration of specialists from different disciplines, as it requires knowledge in molecular biology, computer science, engineering, mathematics, physics and chemistry. Of the papers discussed in this short Review, it is only the research efforts to refactor the genome of a bacteriophage[2] or to recode RNA viruses[3–5] that belong to the category of synthetic biology. It can be predicted with certainty, however, that this will rapidly change in the coming years. In the future, large-scale changes will be introduced into numerous viruses, allowing the creation of redesigned particles that can provide new insights into biology or the design of new vectors that can prevent or cure infectious diseases, cure genetic deficiencies by delivering genes or treat cancer through oncolytic mechanisms, to name but a few applications.

Nucleic acid synthesis and sequencing

In 1828, Friedrich Wöhler synthesized urea from inorganic sources[6], striking a heavy blow to the doctrine of vitalism[7]. The chemical DNA was discovered in 1869 (ref. 8), but it took decades to solve the structural configuration of polynucleotides[9,10]. In keeping with their tradition, chemists began to synthesize DNA as soon as DNA structures had been published. The most ambitious of such early ventures was Khorana's synthesis of a 75-base-pair (bp) double-stranded DNA that encoded the nucleotide sequence of yeast $tRNA^{Ala}$, published in 1970 (ref. 11). This was followed by the chemical and enzymatic synthesis of the first man-made functional gene, the 207-bp DNA of *Escherichia coli* tyrosine suppressor tRNA[12].

These early landmarks consumed enormous resources, in the case of $tRNA^{Ala}$ some "20 man-years of effort"[13]. In the 1980s, however, DNA synthesis went through a rapid transformation, with the introduction of

[1]Department of Molecular Genetics and Microbiology, Stony Brook University, Stony Brook, New York, USA. [2]Influenza Division, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. [3]Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA. Correspondence should be addressed to E.W. (ewimmer@ms.cc.sunysb.edu).

Published online 9 December 2009; doi:10.1038/nbt.1593

novel activated nucleosides that allowed fully automated 3′-to-5′ synthesis of oligodeoxynucleotides (oligos) on solid supports[13,14]. In particular, phosphoramidites (that is, nucleotides that carry protective groups on the reactive hydroxyl and phosphate groups of the ribose and the amine of the base) have been the building blocks of choice. During the past 20 years, numerous DNA synthesis companies have been established in response to an exploding demand for oligos (~15–80 nucleotides (nt)) that are used for genetic analyses, PCR, diagnostic assays, sequence determination or other procedures. The turnaround time for an order of a 75-bp DNA, corresponding to yeast $tRNA^{Ala}$, with extra base pairs at each end encoding restriction sites for subcloning, is currently less than 1 week—a fraction of the time and effort expended originally in Khorana's laboratory.

The assembly of larger DNA segments representing genes or entire genomes, however, is still tedious and costly, even today. It requires many oligos that must be purified, because their chemical synthesis is error prone (none of the successive chemical reactions during 3′-to-5′ chain elongation proceeds at 100%). For this reason, the building blocks for the assembly of large polynucleotides are generally no longer than 40–80 nt. Different approaches have been used to assemble oligos into large polynucleotides, although all have in common the processes of enzymatic chain elongation and/or ligation of hybridized overlapping oligos[14,15]. Examples are the 2.7-kbp plasmid containing the β-lactamase gene[16] and the 4,917-kbp gene encoding the merozoite surface protein (MSP-1) of *Plasmodium falciparum*[17]. Currently, synthesizing genes or genomes is most cost efficient when done in part by commercial facilities, where the cost per base pair of finished and sequence-confirmed DNA is now as low $0.39 (E.W. and S.M., based on information obtained from an informal web survey).

Work in one of our groups (E.W. and colleagues)[18] led to the first chemical synthesis of a DNA (7,500 bp) corresponding to the entire genome of an infectious organism, poliovirus, published in 2002. At the time of its publication, the poliovirus-specific DNA was the largest DNA ever synthesized. This milestone was subsequently dwarfed in scale by the synthesis of the 582,970-bp genome of *Mycoplasma genitalium* in 2008 (ref. 19). Although this synthetic bacterial genome has not yet been 'booted' to life, the assembly of such a large DNA molecule bears witness to the vast possibilities that DNA synthesis will ultimately offer in engineering bacteria or viruses.

Although the mechanics of constructing genes or genomes from oligos is being refined, DNA synthesis is making rapid progress, so that it is likely to fundamentally change research in molecular biology[14]. In 2004, Tian *et al.*[20] published a massively parallel microchip-based DNA synthesis approach that they predicted "might increase yields in oligo synthesis from 9 bp per dollar to 20 kbp per dollar." Once this or related strategies have matured and reach commercialization, the synthesis of small viral DNA genomes (for example, the 3,215-bp genome of hepatitis B virus (HBV)) could be accomplished for less than $100. At so low a cost, who would then construct an HBV mutant by such classic methods as site-directed mutagenesis? All current gene synthesis methods, either practiced or just conceived, still depend on relatively short oligos as their basic building blocks. But further progress in synthetic biology will require accurate synthesis of long, continuous DNA sequences.

Synthesizing large DNA molecules would be of only limited value if new methods of DNA sequencing had not kept pace with the advances in synthesis. In fact, the advances in DNA sequencing have dwarfed current DNA synthesis technology.

The first sequence of a naturally occurring polynucleotide, yeast $tRNA^{Ala}$, was deciphered by R. Holley and colleagues[21] in 1965. Initiated in 1958, the most difficult task of this project was to isolate from 140 kg of bakers' yeast 1 g of highly purified $tRNA^{Ala}$, whose 76 ribonucleotides were then sequenced in 2.5 years (the sequence was later revised slightly)[22]. Since then, however, two phases of technological innovation in sequencing have led to rapid progress[23].

The first phase was based on generating radioactive, sequence-specific fragments of DNA and separating them by PAGE[24,25]. Sanger's method of producing fragments enzymatically by chain termination with dideoxynucleoside triphosphates proved to be more practical than the chemical method of Maxam and Gilbert. Subsequently, gels were replaced by capillaries, and radioactive labels by four-color fluorescence; the process was automated and streamlined, but the underlying principle of the dideoxy method remains, to this day, the most widely used platform of DNA sequencing.

The second phase, still in its infancy, falls under the rubric of a single paradigm, termed 'cyclic array sequencing'. "Cyclic array platforms achieve low cost by simultaneously decoding a two-dimensional array bearing millions (potentially billions) of distinct sequence features"[23]. Such instruments, with slightly different technologies, are already commercially available from companies. Other methods, such as single-molecule sequencing, sequencing by microelectrophoresis, sequencing by mass spectrometry or sequencing by squeezing DNA through tiny nanopores (reviewed in ref. 23) are being tested, but have yet to mature into commercially useful techniques.

The strong progress in sequencing technologies is evident in the reduction in time and costs of human genome projects. The sequence of the 'inaugural human genomes' (3×10^9 bp), published in 2001 (refs. 26–28), was determined over a period of roughly 10 years at a cost of $3 billion—and it was incomplete[27]. In contrast, the complete sequence of Jim Watson's genome was determined in 4 months at a cost of less than $1 million[29]. Currently, the price has dropped further to below $50,000 (ref. 30), and there is reason to believe that the number of solved human sequences will exceed 1,000 in the near future.

Virology in the era of gene synthesis

Viruses store their genetic information in DNA or RNA. Total-genome synthesis of a viral genome seemed likely to occur first with one of the small DNA viruses; the protocol seemed straightforward: simply transfect the synthetic DNA into suitable host cells and assay the emerging virus. In fact, the first chemical whole-genome synthesis was performed with poliovirus, an RNA virus.

How do RNA viruses fit into the world of DNA synthesis and DNA sequencing? The answer is 'reverse genetics'. In their landmark paper of 1978, Weissmann and colleagues[31] converted the RNA genome (4,127 nt) of the RNA phage Qβ into double-stranded DNA with the aid of reverse transcriptase, an enzyme (of retroviruses) that transcribes RNA into DNA. The virus-specific double-stranded DNA (cDNA), which was embedded into a plasmid, yielded authentic Qβ phage following transfection into bacteria. At the time, the authors concluded that the viral cDNA "would allow genetic manipulations that cannot be carried out at the RNA level"[31,32], an understatement that revolutionized molecular biology of RNA viruses. Three years later, Racaniello and Baltimore[33] repeated this experiment with poliovirus. Again, the virion RNA, embedded as cDNA into a plasmid, yielded authentic poliovirus in very poor yield when transfected into HeLa cells, a human cancer cell line that is optimal for poliovirus proliferation.

Depending on the nature of the RNA virus (either positive-strand viruses, whose genome is of the same polarity as mRNA, or negative-strand viruses, whose genome is of the opposite polarity to that of mRNA), virus-specific cDNAs can now be readily prepared and used by different strategies to regenerate the parental RNA virus in high yield. The utility of reverse genetics was quickly recognized and, not surprisingly, it has now been developed for member viruses of nearly every known RNA virus family (for example, rabies virus[34], respiratory

syncytial virus[35], influenza A virus[36,37], measles virus[38], Ebola virus[39] and bunyavirus[40]. Reverse genetics systems have also been recently achieved for members of the *Reoviridae* (viruses with a segmented double-stranded genome—for example, rotavirus[41]—using a helper virus–driven reverse genetics procedure).

Synthesis of viral cDNA with reverse transcriptase requires, of course, naturally occurring virion RNA as template. An alternative is the chemical synthesis of cDNA, which, of course, requires knowledge of the viral genome sequence. The first replicating structure that was synthesized from sequence information was a replicon of the hepatitis C virus that lacked the genes for the structural proteins[42], and the first synthesis of a complete viral genome was that of poliovirus[18]. Currently, 2,361 complete viral genome sequences have been deposited in the Viral Genome Resource (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239), ready to be downloaded and investigated further. Thus, there are huge resources of information available in the virus field, waiting to enter studies that we may broadly term synthetic virology.

There are, of course, cases in which no complete viral genome sequences are available for chemical synthesis. A notable recent example is the synthesis of the 1918 'Spanish' influenza pandemic virus, which caused the most severe influenza pandemic in history. Although the pandemic virus was not isolated at the time, work in one of our laboratories (J.K.T. and colleagues)[43–49] deciphered the genome sequence using influenza viral RNA fragments, <100 nucleotides in length, that were preserved in the tissues of victims of the 1918 pandemic. In addition, Hahn and colleagues successfully synthesized chimpanzee retrovirus simian immunodeficiency virus cpz (SIVcpz)[50], the natural reservoir of HIV-1 and another case in which chemical synthesis was the only means of obtaining cDNA to generate infectious virus. The resurrection of an infectious retrovirus by whole-genome synthesis[51] of a consensus sequence of ancient remnants endogenous to the human genome also illustrates the potential of this approach in archaevirology. As with the synthesis of SIVcpz, the total synthesis of an infectious recombinant bat severe acute respiratory syndrome (SARS)-like coronavirus cDNA (29.7 kbp) was also aimed at studying mechanisms of *trans*-species infection and zoonosis[52].

In the following sections, we discuss in more detail the complete synthesis of several different viruses, either in the absence of natural template or using reverse genetics. We then go on to describe several applications of synthetic virology, such as the large-scale recoding of the viral genomes for the production of attenuated vaccine candidates or the use of refactoring (that is, the synthesis of portions of a genome) to facilitate the elucidation of individual gene functions.

Whole-genome synthesis of poliovirus

In 2002, one of our groups (E.W. and colleagues)[18] published the cell-free chemical biochemical synthesis of poliovirus type 1, Mahoney (PV1(M) in the absence of a natural template. This work caught global attention, high praise, ridicule and fierce condemnation[53]. Apart from providing a 'proof of principle', the experiment signaled a new era in biology—that is, the chemical synthesis of organisms as an approach to investigating gene function and pathogenicity by allowing large-scale changes in a genome of interest. The extent of such alterations (for example, large changes in genome architecture and gene structure) cannot be achieved with the 'traditional' methods in molecular biology[54].

The synthetic polio cDNA contained 27 intentional nucleotide changes that were placed across the genome to serve as genetic markers (watermarks). When grown in HeLa cells, an efficient tissue culture system for poliovirus proliferation, the synthetic virus (denoted sPV1(M)) showed no phenotypic changes compared with the wild-type PV1(M). However, when injected intracerebrally into *CD155* tg mice, which are transgenic for the poliovirus receptor, CD155, the median lethal dose (LD_{50}) was five orders of magnitude higher than that of the wild-type virus (LD_{50} values of 10^2 and 10^7 for the wild-type virus and sPV, respectively; E.W. and colleagues)[18]. Unexpectedly, the genetic locus for the enormous attenuation of sPV(M) was a single A residue at position 102 of the genome, located in the 5′ nontranslated region (5′ NTR) of the genome at a site long thought to serve as simply a spacer between two highly structured regions[54]. This unexpected result led us[55] to develop a highly attenuated oncolytic poliovirus.

The synthesis of poliovirus did not require living cells. Subsequent to its chemical synthesis, the cDNA was transcribed *in vitro* into infectious viral RNA (E.W. and colleagues)[56] that, in turn, yielded infectious sPV1(M) upon incubation in an extract of non-infected HeLa cells (E.W. and colleagues)[57]. For the chemist, therefore, poliovirus is nothing more than a chemical. When the virus enters a cell, however, it has a program for survival. It will subvert cellular compartments and turn them into viral factories, in which it will proliferate subject to the evolutionary laws—heredity, genetic variation, selection toward fitness, evolution into different species and so on. That is, poliovirus obeys the same rules that apply to living entities[53]. One could even argue that poliovirus has sex in the infected cell, as it readily recombines with sibling progeny or with related viruses should they co-infect the same cell (E.W. and colleagues)[58]. This fascinating dual nature of viruses as nonliving and living entities[53,59–63]—that is, an existence as chemicals with a life cycle—has been largely ignored in response to the chemical and biochemical synthesis of poliovirus, which was published in 2002.

Finally, it should be noted that the synthesis of poliovirus also confirmed the accuracy of the genome sequence. This may be considered utterly superfluous, as the sequencing of PV1(M), the first of any lytic animal RNA virus, was originally carried out by two different methods[64,65] and confirmed subsequently in numerous genetic analyses. But chemical synthesis is clearly useful in providing confirmation of sequence and will prove useful going forward in the proofreading of larger genomic sequences[14,18,66].

Whole-genome synthesis of 1918 'Spanish' influenza virus

Unlike the strain of poliovirus type 1 whose synthesis was described above, the virus causing the 'Spanish' influenza pandemic in 1918–1919 was not isolated at the time of the outbreak, and thus its reconstruction using gene synthesis and reverse genetics technology first required characterization of the viral genome using archaevirology. The influenza pandemic of 1918–1919 caused up to 50 million deaths worldwide and remains an ominous warning to public health as to the possible impact that a future influenza pandemic could have (J.K.T. and colleagues)[67,68]. Many questions about its origins, its unusual epidemiological features and the basis of its pathogenicity remain unanswered, but interest in the 1918 virus has been prompted by the possible emergence of a future pandemic caused by the H5N1 virus. Understanding how the 1918 pandemic virus emerged and mapping the virulence factors may also help us in preparations for the current H1N1 influenza pandemic.

The effort to determine the complete genomic sequence of the 1918 influenza virus began in 1995, when one of us (J.K.T. and colleagues) initiated a project to recover viral RNA fragments of the 1918 virus from preserved tissues of victims of the pandemic using reverse transcription PCR (RT-PCR)[43]. The genome was completed in 2005 (refs. 44–49; reviewed in ref. 69). The development of reverse genetics technology for influenza viruses in 1999, which allowed the production of infectious virus entirely from plasmid-cloned influenza gene segments without helper virus[36,37,70], makes it possible to produce influenza viruses with specific sequences for research into pathogenesis and molecular virology, as well as for vaccine production. This technology also made possible experiments using infectious viruses that contain 1918 influenza genes. Once

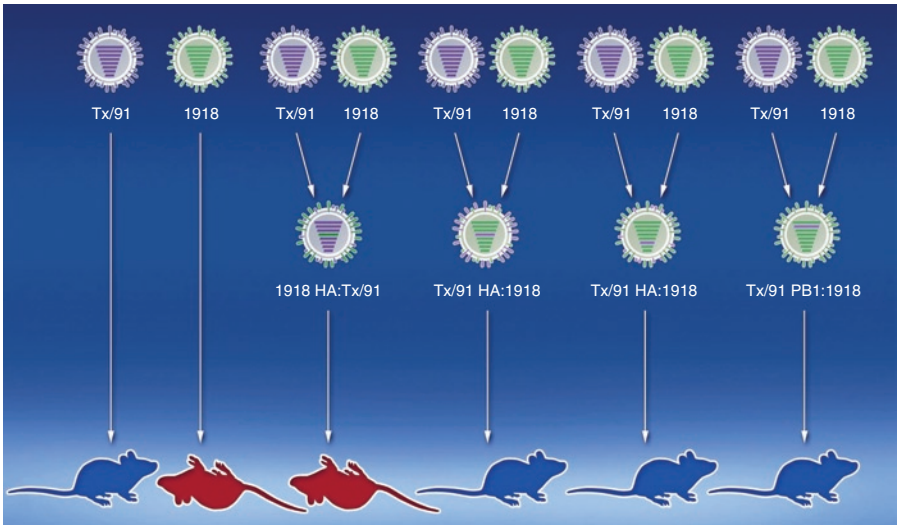

Figure 1 Comparison of lethality in mice infected with select 1918 and modern human H1N1 influenza A (Tx/91) reassortant viruses. BALB/c mice were inoculated intranasally with 10^5 PFU of virus to determine which virus genes of the 1918 virus contributed to virulence. Among all eight gene segments tested individually, the HA gene was the only 1918 virus gene able to confer a virulent phenotype when rescued on the genetic background of Tx/91 H1N1 virus. In the reciprocal experiments, the exchange of most of the individual 1918 influenza virus genes with seasonal influenza Tx/91 virus genes did not alter the virulence of the lethal 1918 virus; however, substitution of the HA, NA or PB1 genes substantially affected the ability of this virus to cause severe disease in mice. Illustration by J. Archer (Centers for Disease Control).

the sequence was determined, the 1918 influenza virus gene segments were synthesized using commercially obtained overlapping oligos and subcloned into plasmids for 'rescue' using reverse genetics[36,37,71]. This was crucial, because sequence analysis alone offered no direct clues to the pathogenicity of the 1918 virus.

Work in our laboratories (J.K.T., T.J.T. and collaborators)[72–78] has shown that, in mice, viral constructs bearing at least the 1918 hemagglutinin (HA) gene in a background of modern, non–mouse-adapted human influenza A virus are all highly pathogenic. Furthermore, expression microarray analysis performed on whole-lung tissue of mice infected with the reconstructed 1918 virus or viral constructs containing at least the 1918 HA and neuraminidase (NA) genes showed marked upregulation of mouse genes involved in apoptosis, tissue injury and oxidative damage[73,75]. Pathology in mice, although reminiscent of some of the acute viral pneumonia pathology seen in 1918 autopsy studies (J.K.T. and Morens, D.M.)[79], is nevertheless distinctive. These findings were unexpected, because the viruses with the 1918 HA gene had not been adapted to mice. Control experiments in which mice were infected with modern human viruses produced limited viral replication and little disease.

A recent study in which single gene segments of the 1918 virus were replaced with those from a recent human H1N1 influenza virus, work in one of our laboratories (T.M.T. and colleagues)[76] has revealed that the HA, NA and polymerase PB1 genes are important for virulence and replication in the mouse system; however, the fully virulent phenotype is observed only with the completely reconstructed virus (**Fig. 1**; T.M.T., J.K.T. and colleagues)[80,76]. The demonstrated role of the HA and PB1 genes in replication efficiency and virulence is particularly interesting because both genes were transferred by reassortment from an avian virus to the then-circulating human influenza virus, to generate the 1957 and 1968 pandemic strains. The acquisition of an avian influenza PB1 gene by reassortment might result in increased transcriptional activity of the RNA-dependent RNA polymerase and increased virus replication efficiency of a new pandemic strain[49,81].

The HA and its binding preference for particular sialic acid (SA)-terminated glycans has also been implicated in efficient transmission of the 1918 virus in ferrets (T.M.T. *et al.*)[82]. It has been generally suspected that a switch in receptor-binding preference that confers efficient transmission among humans would be a necessary step for avian influenza viruses in the generation of a pandemic virus. Notably, we (T.M.T. and colleagues) have found that mutation of two amino acid residues (D190E, D225G) in the HA, which was previously identified as sufficient to switch the receptor-binding preference of parental 1918 HA (α2,6 SA receptor preference) to the avian α2,3 SA receptor preference[44,83], prevented transmission among ferrets without affecting the replication efficiency of the rescued 1918 virus[82]. These findings suggest that changes in receptor binding of avian influenza viruses could potentially move them one step closer to a pandemic phenotype.

The viral genotypic basis of the 1918 pandemic virus's virulence and transmissibility has not yet been fully mapped; however, by making chimeric viruses containing at least one 1918 influenza virus gene segment, and by targeted mutagenesis or gene synthesis, future experiments should help us to determine how this pandemic virus killed and spread so efficiently. Such knowledge may help us to elucidate virulence factors for other influenza viruses such as the 2009 influenza pandemic and, thereby, help us to identify targets for future drug intervention.

Whole-genome syntheses of other RNA viruses

Apart from poliovirus and influenza virus, the complete genomes of several other RNA viruses have recently been chemically synthesized. These include human endogenous retrovirus, HIVcpz and SARS-like coronavirus.

Reconstitution of an infectious, human endogenous retrovirus. Of the 3×10^9 bp hat constitute the human genome, nearly 8% (that is, 2.8×10^8 bp), comprise sequences of retroviral origin[84,85]. After having invaded the chromosome of human germ cells, they were inherited for millennia in a mendelian manner; thus, they are viral fossils, but the function of these remnants in human evolution, physiology and disease remains unclear. Most of the genes or gene fragments are, however, inactive owing to various replication errors during proliferation of the host cells. An exception is the *env* gene, which seems to be conserved because it may have a crucial role in hominoid placental physiology[84,85]. Nevertheless, all of the ancient human retroviruses are degenerate, including the human mouse mammary tumor virus–like 2 provirus (HML-2) of the human endogenous retrovirus (HERV) K proviruses (HERV-K(HML-2)). The latter may have been added to the Old World primate genomes relatively recently in human evolution, but no functional proviruses able to produce infectious particles have been isolated.

To reconstruct a replicating retrovirus that may resemble the ancestor of HERV-K sequences, Lee and Bieniasz designed a consensus genome (9,472 nt) and, using whole-genome synthesis, generated the proviral clone HERV-K_{CON}, which "likely resembles the progenitor of HERV-K(HML-2) variants that entered the human genome within

the last few million years"[51]. In a parallel study, Dewannieux *et al.*[86] also reconstructed infectious HERV-K(HML-2) from a consensus sequence, but they applied site-directed mutagenesis to arrive at an infectious provirus, which they named *Phoenix*.

In both studies, the first human retrovirus of endogenous origin had all the properties of a C-type retrovirus. The infectivity of ancestral retrovirus in various cell types, however, was extremely low, which to some extent dispelled concerns that resuscitating an ancient human infectious virus is inherently risky[87]. Still, studying the pathogenic potential of a virus that probably circulated in the then-human population for millions of years may yield valuable clues as to its impact on human evolution.

Synthesis of HIVcpz—the origin of the HIV-1 pandemic. It was long suspected that chimpanzees provided the natural reservoir for the human immunodeficiency viruses that caused the zoonotic infections responsible for the AIDS pandemic. But because the simian immunodeficiency virus most closely related to HIV-1 (SIVcpz) was found only in animals (*Pan troglodytes troglodytes*) in captivity, direct proof was lacking.

In 2006, Hahn and colleagues[88] provided the first convincing evidence of SIVcpz antibodies and nucleic acid in fecal samples from wild *P. T. troglodytes* in a narrow area in south-eastern Cameroon. However, recovery of replication-competent virus from fecal samples had failed. These authors therefore analyzed virus-specific nucleic acids isolated from the fecal samples and obtained a consensus sequence that, when chemically synthesized, yielded infectious molecular clones of SIVcpz[50]. Analyses of these isolates yielded the important result that "naturally occurring SIVcpz strains already have many of the biological properties required for persistent infections of humans." The authors conclude that "medically important 'SIV isolates' that have thus far eluded investigation… are needed to identify viral determinants that contribute to cross-species transmission and host adaptation"[50].

Synthesis of infectious bat SARS-like coronavirus. In 2002, a new acute respiratory syndrome emerged in China, caused by an unknown infectious agent. By the summer of 2003, the agent had caused disease in 8,427 people, of whom 813 died, and fears of a deadly pandemic spread around the globe. As a result of unprecedented collaborative efforts, led by the World Health Organization (Geneva), the pathogenic agent was rapidly identified as a new coronavirus, named severe acute respiratory syndrome virus coronavirus, or SARS-CoV.

Coronaviruses are plus-strand RNA viruses with the largest-known RNA genome (~30 kb). The properties (genome sequence, cultivation and serology) and pathogenic potential of SARS-CoV were rapidly established, but, intriguingly, in July 2003 SARS-CoV disappeared (http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5228a4.htm) as quickly as it had emerged. This lucky break happened despite the fact that there were no drugs, let alone vaccines, available to treat or prevent SARS infection. Isolation of patients, an old medical practice, has been credited with the fading of the SARS-CoV epidemics.

Fear remained that SARS-CoV might reappear, perhaps more contagious than before. Thus, the source of SARS-CoV became an important issue, as it was suspected that the human agent may have evolved from a zoonotic infection, as true of influenza and HIV. Early evidence implicated the Chinese cat-like mammals known as civets, but overwhelming evidence now suggests that "bats are natural reservoirs of SARS-like coronaviruses"[89].

As yet, there is no known tissue culture system that supports the replication of bat SARS virus, suggesting that it is not infectious in humans[52]. However, this inability to culture the virus also prevents investigation of the mechanism of cross-species transmission from bats to civets to humans (or transmission directly from bats to humans). The nucleotide sequence of the bat SARS virus is known. To determine the possible steps by which the bat SARS-CoV may have adapted to human populations, Denison and colleagues[52] synthesized the 29.7-bp bat SARS virus cDNA. They subsequently succeeded in converting the bat SARS-CoV to an infectious clone by exchanging the region encoding its receptor-binding domain (RBD) with that of the human SARS-CoV. The result is the largest replicating genome to be synthesized so far[52]. The authors conclude that "rational design, synthesis, recovery of hypothetical recombinant virus can be used to investigate mechanisms of *trans* species movement of zoonoses and has great potential to aid in rapid public health responses…"[52].

Whole-genome synthesis of DNA viruses

To date, the complete genome of only one DNA virus—ΦX174—has been assembled by synthesis. Other work has applied DNA synthesis to understanding the structure and function of bacteriophage T7 DNA, but this involved the creation of portions, rather than the entire reconstitution, of the viral genome (see "Refactoring the bacteriophage T7 genome" below).

Eighteen months after the poliovirus synthesis, Smith *et al.*[66] described the *de novo* synthesis of the first DNA virus genome—the 5,386-bp genome of bacteriophage ΦX174[66]. Remarkably, existing methods of DNA synthesis were fine-tuned to complete the genome in 2 weeks. The unedited DNA was then transfected into bacteria, which sorted the good from the bad and produced viable bacteriophages[66]. This elegant work confirmed the general utility of DNA synthesis in assembling the genomes of viruses that was first shown with poliovirus, and it has subsequently spurred further work to make larger DNA assemblies, allowing the synthesis of whole bacterial chromosomes.

Virus attenuation by large-scale recoding

Synthetic biology strives to generate new biological systems that do not exist in nature, primarily for medical or commercial applications. The poliovirus synthesis described above was not intended to be an example and, indeed, hardly falls into the category of synthetic biology, because it resulted in a poliovirus with a nearly identical phenotype as the model wild-type virus. In the following section, we describe experiments that have led to the rational design of vaccine candidates from the poliovirus and influenza viruses.

Upon entry into a host cell, the poliovirus uses its genome as mRNA, the hallmark of all plus-strand RNA viruses[90]. Poliovirus belongs to a large family of human and animal pathogenic viruses, the *Picornaviridae*. These viruses express all of their proteins in the form of a single polypeptide of just over 2,000 amino acids—the polyprotein (**Fig. 2a**). This large precursor polyprotein is co- and post-translationally cleaved by the proteolytic action of two virus proteinases that are, remarkably, embedded in the polyprotein itself (E.W. *et al.*)[91]. A poliovirus polypeptide of 2,209 amino acids can be encoded in about $10^{1,100}$ ways, a number much larger than the number of atoms in the universe. This poses the question of how and why selection has led to one of these possible $10^{1,100}$ sequences that we consider 'wild-type' PV1(M).

It should be noted that poliovirus, like all RNA viruses, is a quasispecies that, in reality, exists in nature as a large swarm of different genotypes[92,93]. This is the consequence of the high error rate of viral RNA synthesis in the absence of proof reading and editing functions. The 'wild-type' sequence in this vast swarm is the genotype that can proliferate most efficiently under the prevailing conditions, where it out-competes all of its related genotypes[92,93].

One of our groups (E.W., S.M. and colleagues)[4,5] has been investigating the effect of genome-scale changes in poliovirus on codon usage. To reduce the complexity of our experiments, we have restricted our investigations to only one-third of the poliovirus genome. This is the

region encoding the P1 capsid precursor. The P1 polyprotein, which consists of 881 amino acids (2,643 nt; **Fig. 2**), can be encoded in 10^{442} ways—still a mind-boggling number.

Codon bias. A major reason for the nearly unlimited possibilities of encoding a protein is the degeneracy in the genetic code (for example, several synonymous codons can specify the same amino acid). However, the preference for a synonymous codon is not the same in *E. coli*, in jellyfish or in human cells; this phenomenon is termed 'codon bias'. For example, in humans, the alanine codon GCC is used four times more frequently than the synonymous codon GCG. The cell's preference of one synonymous codon over another to specify the same amino acid is thought to relate to the abundance of the corresponding cognate tRNAs in the cell. Consequently, rare codons are associated with a suboptimal translation of an mRNA. Codon bias, then, may contribute to the restriction of the abundance of sequences encoding the same protein. Codons used frequently in the jellyfish may be used rarely in human cells, and thus expression of the jellyfish green fluorescent protein (GFP) in human cells is poor unless the codons of the jellyfish gene have been changed to those frequently used in human cells; accordingly, the GFP gene has been 'humanized' to achieve good expression in human cells[94].

To exploit this phenomenon, we (S.M., E.W. and collaborators)[5] have 'dehumanized' the sequence encoding P1 of the poliovirus polyprotein. We chose this segment of the poliovirus genome because we have gathered abundant evidence that the P1 coding sequence does not harbor RNA signals essential for viral proliferation (E.W. and colleagues)[91,95]. For example, (i) the nucleotide sequence of the P1 region can be changed drastically, as long as the amino acid sequence that it encodes is preserved[3–5]; (ii) the P1 coding region can be exchanged with foreign genes (for example, firefly luciferase[96]); or (iii) the P1 coding region can be deleted altogether (in defective interfering particles[97]) without loss of efficient RNA replication. However, changing synonymous codons in the P1 region from frequently used to rarely used codons (that is, 'codon deoptimizing' this segment of viral mRNA) will unbalance the synthesis of the polyprotein without changing its amino acid sequence, resulting in attenuated viruses[3,5].

Because of the existence of a polyprotein in picornaviruses, codon deoptimizing the P1 region (the N-terminal third of the polyprotein) compromises viral replication: fewer ribosomes arrive at the coding region for the essential replication proteins (genomic regions P2 and P3; **Fig. 2a**), and genome replication is thus reduced or shut off altogether. It was not surprising, therefore, that extensive codon deoptimization in virus PV-AB (**Fig. 2b**), harboring 680 changes (out of 2,643 nt) without altering a single encoded amino acid, led to a 'dead' phenotype. However, subcloning individual segments of the recoded P1 segment revived the virus, albeit in attenuated form. Indeed, the subclones are not only inhibited in protein synthesis[5], but their neurovirulence is attenuated in *CD155* tg mice as well[5].

A notable property of the subclones of PV-AB is a marked reduction of their specific infectivity, also observed by Burns and colleagues[3]. Wild-type poliovirus (PV1(M)) has a specific infectivity of one plaque-forming unit (PFU) per 115–130 particles[4,5]; in one of the subclones (PV-AB$^{2470–2954}$),

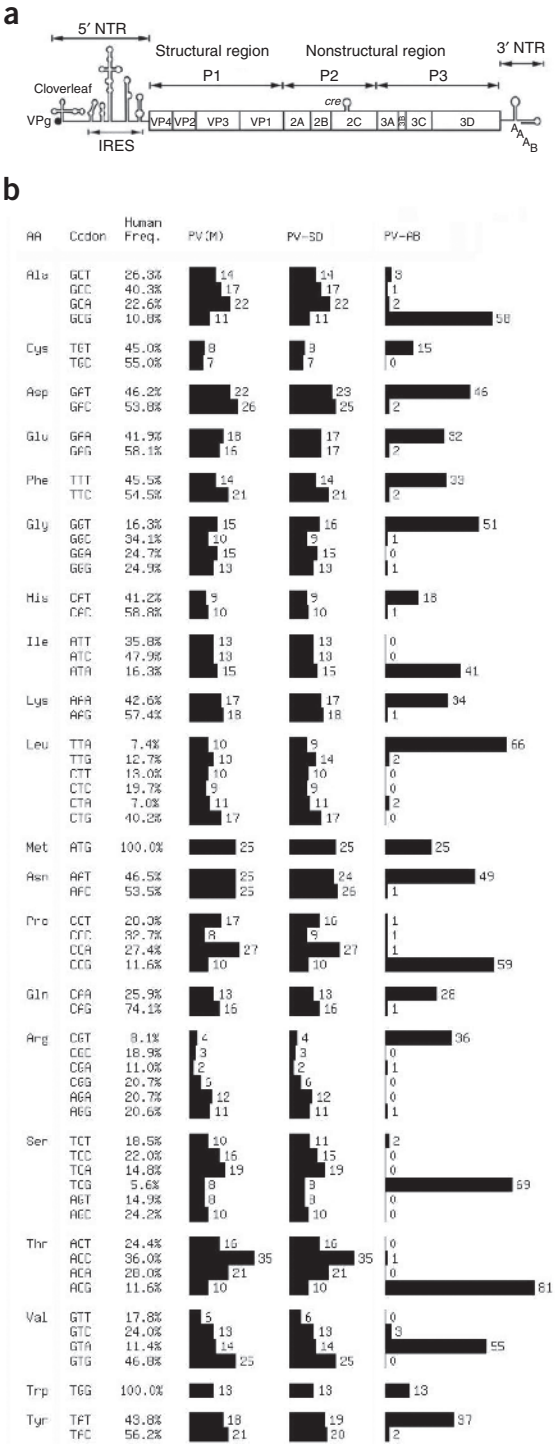

Figure 2 The poliovirus genome and the effect of codon bias. (**a**) Poliovirus genomic RNA[56,91] is of plus-strand polarity (that is, it functions as mRNA in viral replication). It is covalently linked at the 5′ end to the small viral protein VPg (3B of the polyprotein), followed by a long 5′ nontranslated region (5′ NTR), a continuous open reading frame (ORF), a 3′ NTR and poly(A). The 5′ NTR consists of structural elements that control RNA replication (cloverleaf) and translation (the internal ribosomal entry site (IRES)). The ORF encodes the polyprotein, the single translation product of the viral mRNA. The polyprotein is proteolytically processed by viral proteinases 2A^{pro} and 3C/3CD^{pro} into functional proteins, which have been divided into the structural region (P1 and the capsid precursors) and the nonstructural regions P2 and P3 (replication proteins). The ORF is followed by a 3′ NTR, which contributes to the control of RNA synthesis, and poly(A). The P1 coding region has been a target for codon and codon pair deoptimization. (**b**) Codon use statistics in synthetic P1 capsid designs. PV-SD maintains nearly identical codon frequencies compared to wild-type PV1(M), while maximizing codon positional changes within the sequence[5]. In PV-AB capsids, the use of nonpreferred codons was maximized. The length of the bars and the numbers behind each bar indicate the occurrence of each codon in the sequence. As a reference, the normal human synonymous codon frequencies for each amino acid are given. Adapted from S.M. *et al.*[5].

this ratio was reduced to about 1 PFU per 100,000[5] particles. That is, only one plaque can be expected to emerge if 10^5 particles are plated onto a dish of 10^6–10^7 HeLa cells. Once this one virus has succeeded in overcoming the host cell, however, its burst size will be only one order of magnitude lower than that of the wild-type poliovirus. That is, although the dehumanized virus can replicate in HeLa cells, once released it has enormous problems in spreading to other cells. It should be noted that we have analyzed the sequences of the codon-deoptimized viruses for the emergence of higher-order structures that could have impeded replication. No such structures have been found. The properties of the virus with a 'scrambled' P1 region (PV-SD; **Fig. 2b**) are discussed below.

Codon pair bias. It has been known since 1989 that in addition to, and independently of, codon usage, pairs of synonymous codons do not exist in the genome at the frequency that one might expect on the basis of the frequency of the two individual codons that make up the pair. This phenomenon, called 'codon pair bias', was discovered in prokaryotic cells[98] but has since been seen in all other examined species, including humans[99]. For example, given the known codon frequencies in humans, the amino acid pair Ala-Glu is expected to be encoded by GCC GAA and GCA GAG about equally often. In fact, the codon pair GCC GAA is strongly underrepresented, despite containing the most frequent alanine codon, such that it is used only one-seventh as often as GCA GAG[4]. The functional significance of codon pair bias is a mystery, but it can be studied in systems, such as poliovirus, in which large-scale changes of codon pairing are likely to present with phenotypes in viral proliferation.

On the basis of 14,795 annotated (known) human genes, the Wimmer group has calculated a codon pair score, specific for each of the possible 3,721 codon pair combinations, as well the codon pair bias (CPB) for each gene, taking into consideration the codon frequency for each of the paired codons and the frequency of the encoded amino acid pair. In **Figure 3**, the calculated CPB of a human gene is plotted against its amino acid length. Underrepresented codon pairs yield negative scores. Wild-type PV1(M) shows a slightly negative score (CPB = –0.02), but, not surprisingly, it uses codon-pairing corresponding to human genes (**Fig. 3**). Using a custom-made computer optimization algorithm, we have constructed polioviruses whose P1 coding region had either a substantially negative codon pair bias, containing many underrepresented codon pairs (PV-Min, CPB = –0.474), or a substantially positive codon pair bias, containing many overrepresented codon pairs (PV-Max, CPB = +0.246; **Fig. 2**), while retaining the exact set of codons present in the wild-type virus[4] and the same amino acid sequence of P1.

Unexpectedly, transcripts of PV-Min, using underrepresented codon pairs, did not yield virus upon transfection and blind passages. Apparently, the accumulation of hundreds of unfavorable codon pairs led to a dead phenotype ('death by a thousand cuts'). Conversely, various subclones carrying segments of the P1 region of PV-Min cloned into wild-type poliovirus (for example, PV-MinXY or PV-MinZ, with CPB scores of –0.32 and –0.19, respectively; **Fig. 3**) were viable, albeit severely debilitated, as revealed by plaque assays and single-step growth kinetics experiments, and their neurovirulence in *CD155* tg mice was reduced 1,000-fold[4]. Similar to the reduced specific infectivity of subclones of PV-AB, that of the PV-Min subclones, PV-Min XY and PV-MinZ, was also reduced to roughly 1 PFU per 10,000 particles[4]. Moreover, the translational activities of the subclone RNAs were impaired, an observation suggesting a relationship between viability and viral protein synthesis[4,5].

This prompted us to investigate whether PV-Max, the variant in which the P1 coding region carried many overrepresented codon pairs, would grow to titers substantially higher than those of wild-type poliovirus or be more neurovirulent than wild-type virus in *CD155* tg mice. It had neither of these phenotypes. PV-Max, therefore, was not a highly virulent variant virus, an observation that suggests that the end product of evolution of poliovirus had already optimized the encoding of polypeptide P1 (ref. 4) and cannot be 'improved'. For an RNA virus, which exists as a quasi-species, this is not unexpected. The signature of poliovirus is the efficient replication of its small genome, which, under optimal conditions, is driven by an irresistible desire to maintain optimal genome structure; that is, during replication, important replication signals are constantly rebuilt, and unnecessary nucleotide sequences (for example, foreign genes and duplications) are deleted. Indeed, the virus has an inexhaustible arsenal to achieve these goals—by exploiting point mutations, by homologous or illegitimate recombination and even by the acquisition of foreign RNA sequences. These considerations do not mean that the sequence of PV-Max is the only other possible sequence that can express wild-type phenotypes in tissue culture and in *CD155* tg mice. On the contrary, there are probably a huge number of sequences with wild-type phenotypes.

Regardless of the changes in usage of rare codons or underrepresented codon pairs, the product of the translational machinery remains the same; however, the efficiency of protein synthesis may be vastly altered. Thus, no matter how many synonymous changes have been introduced into the genome, a virus synthesized in the infected cell will have the same structure and will encode the same replication proteins as the wild-type virus, but it may be substantially disadvantaged in terms of proliferation. Such a variant of a human pathogenic virus may enter the host by its normal route and replicate poorly, but still allow the host to mount an immune response strong enough to induce lasting protective immunity. In other words, a human virus with altered codon usage or altered codon pair usage could possibly serve as a vaccine. Recoding viral genomes, a process that we call 'synthetic attenuated virus engineering' (SAVE), may be a new and rapid route to discover vaccine candidates and prevent viral disease. Indeed, polioviruses harboring underrepresented codons or underrepresented codon pairs are attenuated in *CD155* tg mice. Infection of these mice with a sublethal dose of codon- or codon pair–deoptimized viruses induced an immune response that protected the animals against a lethal dose of the wild-type virus[4,5].

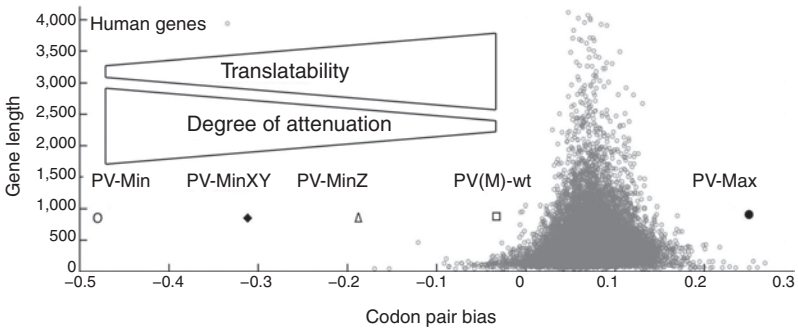

Figure 3 The codon pair bias (CPB) score for each of the 14,795 annotated human genes was calculated[4]. Each dot represents the calculated CPB score of one gene plotted against its amino acid (aa) length. Predominant use of underrepresented codon pairs yields negative CPB scores. Various poliovirus constructs are plotted according to the CPB score of their P1 capsid precursor protein. As the CPB decreases, translatability decreases and the attenuation effect on the virus increases. PV-Min is nonviable; PV-Max expresses replication and virulence phenotypes similar to those of wild-type PV. Adapted from ref. 4.

Box 1 'Dual use' concerns and total synthesis of viruses

As discussed elsewhere[53], the publication of the *de novo* synthesis of poliovirus in the absence of natural template aroused unusually strong and sometimes conflicting responses from different quarters of society. Many of the negative reactions were tainted by fear that the poliovirus synthesis could aid bioterrorism. This was not surprising, as the news of the synthesis reached a public (particularly in the United States) that was highly sensitized to the threat of bioterrorism in the months following the attack on the World Trade Center on September 11, 2001 and the anthrax attacks in Washington and elsewhere in 2001 and 2002. However, these concerns, led to numerous useful public debates about complex issues of biological research, scientific publication and national security[105–109]. Notably, the synthesis of the bacteriophage ΦX174 genome in 2003, which stood out because of its speed of merely 2 weeks[67], or the resurrection of the 1918 'Spanish' influenza virus in 2005 (ref. 80), did not set off the shock waves experienced in 2002, when the poliovirus genome synthesis was published[18]. There may be two reasons for this. First, these later papers were embedded in numerous editorials or, as in the case of the ΦX174 synthesis, in a carefully orchestrated press conference called by the then US Secretary of Energy. These activities were aimed at explaining to the public the significance, particularly the benefit, of the research involving *de novo* virus synthesis. Second, compared to in 2002, the general public was probably better prepared and better educated to accept the new reality of synthetic viruses and their possible consequences[53]. Meanwhile, several publications, of which only a few can be cited[105–109], attest to the serious efforts by the scientific community to define dual-use research and to limit its possible disastrous consequences.

The Wimmer group is now extending the SAVE strategy from poliovirus to the influenza virus because of its enormous importance as a major human pathogen, both in its pandemic and epidemic forms. We also want to know to what extent computer-aided rational design can lead rapidly to vaccine candidates of a virus whose genetic and pathogenic properties are completely different from that of poliovirus. First, influenza virus is a negative-stranded RNA virus with a segmented genome; on entry into the host cell the virion must activate its virus-associated RNA polymerase, which will synthesize genome-complementary RNA for translation and replication. Second, many of the important steps in influenza virus replication occur in the cell nucleus, an environment that is avoided by nearly all other RNA viruses (for example, poliovirus can replicate even in enucleated cells). Third, influenza virus is an enveloped virus that replicates in the respiratory tract. Although this work is still in progress, our results show that, by codon pair deoptimization, it is possible to rapidly construct highly attenuated influenza viruses that, after a single cycle of immunization, protect mice against a lethal dose of wild-type influenza virus (S.M., D. Papamichail, J.R. Coleman, S. Skiena and E.W., unpublished data). Most importantly, the best of the constructed vaccine candidate strains offer a wide margin of safety at a relatively low concentration of inoculating virus (S.M., D. Papamichail, J.R. Coleman, S. Skiena and E.W., unpublished data).

RNA sequences with shuffled codons

So far, we have discussed changing the codon bias (**Fig. 2b**) or codon pair bias (**Fig. 3**) in the P1 region while retaining its amino acid sequence. In a separate approach, the Wimmer group (S.M., E.W. and colleagues)[5] has shuffled synonymous codons of the P1 region according to a specifically designed computer algorithm to maximize the number of nucleotide changes while retaining the existing codons (for example, without changing the codon bias). The resulting P1 sequence was synthesized and cloned into the backbone of poliovirus, yielding PV-SD[5] (**Fig. 2b**). The P1 coding region of PV-SD contained 934 changes out of 2,643 nucleotides; that is, on average, every third nucleotide was different from that of the wild-type sequence. Notably, PV-SD replicated in HeLa cells with wild-type kinetics[5]. Apparently, the positioning of the synonymous codons in PV-SD, after the extensive codon shuffle, did not influence viral protein synthesis, confirming that synonymous mutations have an effect on the virus only when they are specifically directed to lower the codon bias[5] or codon pair bias[4]. An intriguing conclusion from PV-SD is that, at the genome level, RNA viruses are promiscuous with respect to nucleotide changes, as long as these changes do not affect protein function or the rate of protein synthesis. In other words, RNA structures throughout much of the viral coding regions, with notable exceptions, such as *cis*-acting replication elements or encapsidation signals, are probably rather inconsequential.

As pointed out above, a protein of 881 amino acids (the P1 capsid precursor of poliovirus) can be encoded in about 10^{442} different ways. How many of these sequences, if 'cloned' into the present-day poliovirus, would express a wild-type phenotype that would be stable if passaged in HeLa cells, the preferred substrate for poliovirus in the laboratory? We have not passaged PV-SD in HeLa cells for numerous generations in an attempt to observe genetic variation toward the sequence of wild-type PV1(M). Such an experiment may not yield relevant results, because PV1(M) may also express a genetic drift. After all, the natural human cells for poliovirus proliferation are not known, but they reside in the gastrointestinal tract and are obviously very different from HeLa cells, which have been derived from a cervical cancer.

We have used shuffled sequences to test for unknown *cis*-acting RNA elements in the poliovirus genome. The fact that PV-SD replicates with wild-type kinetics provided proof that the P1 region is void of essential *cis*-acting replication elements, which would probably have been destroyed by the large-scale shuffling. Note that other human picornaviruses, such as rhinovirus type 14 (HRV14), do contain such an essential element (*cre* in HRV14—a stem-loop structure of ~50 nt) in P1. Hence, HRV14-SD would have probably been nonviable[100]. The poliovirus equivalent to the HRV14 *cre* maps to the P2 coding region[101]. We expected that scrambling the P2 coding sequence of the poliovirus genome would destroy the *cre* element and kill the virus, which is indeed what we have found (Y. Song, C. Ward, D. Futcher, S. Skiena, E.W. and S.M., unpublished data). Re-establishing the *cre* sequence in P2 by molecular engineering rescues the virus, which indicates that the P2 region does not contain any other important RNA sequences in addition to *cre*. Synonymous scrambling of RNA virus sequences may be an excellent tool in searching for RNA structures that are essential for viral replication.

Refactoring the bacteriophage T7 genome

"Refactoring [is] a process that is typically used to improve the design of legacy computer software"[102]. Endy and colleagues[2] have used this definition to describe their efforts to redesign the DNA bacteriophage T7 genome (39,937 bp[103]) with an aim to test the functions of a set of T7 genes once they are untangled from each other (by removing overlapping gene segments). To this end, they replaced the left 11,515 bp of the wild-type genome with 12,179 bp of synthetic redesigned DNA (available in cassettes α and β) and tested the biological properties of the synthetic DNA when combined with the remainder of the wild-type (WT) genome. Notably, three chimera, α-WT, WT-β-WT and α-β-WT, were viable, but, perhaps not surprisingly to the

investigators, none grew as well as wild-type T7 phage (although no burst size was included in the report)[2].

Phage α-β-WT, called T7.1, represents a redesign of >30% of the T7 genome and, by removing overlaps, the genes in the α-β region could be studied independently, an enormous advantage for genetic analyses. T. F. Knight described the work as "the most compelling example of work in synthetic biology to date"[104]. From a scientific perspective, Endy's work demonstrated that overlapping genetic elements in the T7 genome were, in aggregate, non-essential for phage replication. Until these experiments, the community had been stressing the importance of these features, given that they are conserved across evolutionary distance, but the synthetic approach provided a way of clarifying this issue. From an engineering perspective, this work provided confidence that up to 5% of the DNA sequence of an organism can be changed while still maintaining its viability.

Conclusions

The ability to manipulate the genomes of viruses has long been important as a model for molecular systems, for investigating viral pathogenesis and for the production of viral vaccines. The methods of molecular biology and the utility of reverse genetics allow the rapid production of altered viruses from cloned viral genes, including those that are important for public health. Until recently, these methods have relied on PCR or RT-PCR amplification of templates from the pre-existing virus, followed by subcloning into the appropriate plasmid vectors, with or without mutagenesis, and such techniques will continue to be invaluable for virology and vaccinology. However, the advances in gene synthesis, coupled with the ability to use the techniques described in this Review, have allowed the production of viruses in the absence of available infectious virus. This has implications not only in terms of dual use (**Box 1**), but also for our understanding of evolution and the properties of important pathogens. Moreover, genome synthesis of both DNA and RNA viruses will lead to unprecedented possibilities in modifying naturally occurring genomes, thereby allowing new studies of viral genome architecture, viral gene expression and gene function. The examples presented in this Review are only the beginning of a new era in which genome synthesis is likely to dominate genetic experiments with viruses.

ACKNOWLEDGMENTS

We are indebted to our colleagues who have participated in the work described here and who have in part edited the manuscript, particularly A. Paul and B. Futcher, and we thank J. Shendure, L. Steward and A.B. Burgin for information provided. The work described here was supported partially by US National Institutes of Health (NIH) grants AI075219 and AI15122 and contract N65236 from the US Defense Advanced Research Project Agency to E.W.; and partially by the intramural research program of the NIH and the National Institute of Allergies and Infectious Diseases (NIAID). The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Keasling, J. The promise of synthetic biology. *The Bridge* **35**, 18–21 (2005).
2. Chan, L.Y., Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Mol. Syst. Biol.* **1**, 2005.0018 (2005).
3. Burns, C.C. *et al.* Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *J. Virol.* **80**, 3259–3272 (2006).
4. Coleman, J.R. *et al.* Virus attenuation by genome-scale changes in codon pair bias. *Science* **320**, 1784–1787 (2008).
5. Mueller, S., Papamichail, D., Coleman, J.R., Skiena, S. & Wimmer, E. 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 (2006).
6. Wöhler, F. Ueber die künstliche Bildung des Harnstoffs. *Ann. Phys.* **12**, 253–256 (1828) (in German).
7. Kinne-Saffran, E. & Kinne, R.K. Vitalism and synthesis of urea. From Friedrich Wohler to Hans A. Krebs. *Am. J. Nephrol.* **19**, 290–294 (1999).
8. Miescher, F. Ueber der chemische Zusammensetzung der Eiterzellen. *Med.-Chem. Unters.* **4**, 441–460 (1871) (in German).
9. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
10. Brown, D.M. & Todd, A.R. in *The Nucleic Acids* Vol. 1 (eds. Chargaff, E. & Davidson, J.N.) 409–430 (Academic, New York, 1955).
11. Agarwal, K.L. *et al.* Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature* **227**, 27–34 (1970).
12. Khorana, H.G. Total synthesis of a gene. *Science* **203**, 614–625 (1979).
13. Caruthers, M.H. Gene synthesis machines: DNA chemistry and its uses. *Science* **230**, 281–285 (1985).
14. Stewart, L. & Burgin, A.B. Whole gene synthesis: a gene-o-matic future. *Front. Drug Des. Disc.* **1**, 297–341 (2005).
15. Sanghvi, Y. A roadmap to the assembly of synthetic DNA from raw materials. in *Working Papers for Synthetic Genomics: Risks and Benefits for Science and Society* (eds. Garfinkel, M.S., Endy, D., Epstein, G.L. & Friedman, R.M.) 17–33 (2007).
16. Stemmer, W.P., Crameri, A., Ha, K.D., Brennan, T.M. & Heyneker, H.L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49–53 (1995).
17. Pan, W. *et al.* Vaccine candidate MSP-1 from *Plasmodium falciparum*: a redesigned 4917 bp polynucleotide enables synthesis and isolation of full-length protein from *Escherichia coli* and mammalian cells. *Nucleic Acids Res.* **27**, 1094–1103 (1999).
18. Cello, J., Paul, A.V. & Wimmer, E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* **297**, 1016–1018 (2002).
19. Gibson, D.G. *et al.* Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* **319**, 1215–1220 (2008).
20. Tian, J. *et al.* Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
21. Holley, R.W. *et al.* Structure of a ribonucleic acid. *Science* **147**, 1462–1465 (1965).
22. Penswick, J.R., Martin, R. & Dirheimer, G. Evidence supporting a revised sequence for yeast alanine tRNA. *FEBS Lett.* **50**, 28–31 (1975).
23. Shendure, J.A., Porreca, G.J. & Church, G.M. Overview of DNA sequencing strategies. in *Current Protocols in Molecular Biology.* (ed. Ausubel, F.M. *et al.*) Unit 7.1 (John Wiley and Sons, Hoboken, NJ, USA; 2008).
24. Maxam, A.M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74**, 560–564 (1977).
25. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
26. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
27. The International Human Genome Mapping Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
28. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
29. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
30. Pennisi, E. Personal genomics. Number of sequenced human genomes doubles. *Science* **322**, 838 (2008).
31. Taniguchi, T., Palmieri, M. & Weissmann, C. Qβ DNA-containing hybrid plasmids giving rise to Qβ phage formation in the bacterial host. *Nature* **274**, 223–228 (1978).
32. Weissmann, C., Weber, H., Taniguchi, T., Muller, W. & Meyer, F. Reversed genetics: a new approach to the elucidation of structure–function relationship. *Ciba Found. Symp.* **66**, 47–61 (1979).
33. Racaniello, V.R. & Baltimore, D. Cloned poliovirus complementary DNA is infectious in mammalian cells. *Science* **214**, 916–919 (1981).
34. Schnell, M.J., Mebatsion, T. & Conzelmann, K.K. Infectious rabies viruses from cloned cDNA. *EMBO J.* **13**, 4195–4203 (1994).
35. Collins, P.L. *et al.* Production of infectious human respiratory syncytial virus from cloned cDNA confirms an essential role for the transcription elongation factor from the 5′ proximal open reading frame of the M2 mRNA in gene expression and provides a capability for vaccine development. *Proc. Natl. Acad. Sci. USA* **92**, 11563–11567 (1995).
36. Neumann, G. *et al.* Generation of influenza A viruses entirely from cloned cDNAs. *Proc. Natl. Acad. Sci. USA* **96**, 9345–9350 (1999).
37. Fodor, E. *et al.* Rescue of influenza A virus from recombinant DNA. *J. Virol.* **73**, 9679–9682 (1999).
38. Takeuchi, K., Takeda, M. & Miyajima, N. Toward understanding the pathogenicity of wild-type measles virus by reverse genetics. *Jpn. J. Infect. Dis.* **55**, 143–149 (2002).
39. Neumann, G., Feldmann, H., Watanabe, S., Lukashevich, I. & Kawaoka, Y. Reverse genetics demonstrates that proteolytic processing of the Ebola virus glycoprotein is not essential for replication in cell culture. *J. Virol.* **76**, 406–410 (2002).
40. Överby, A.K., Popov, V., Neve, E.P. & Pettersson, R.F. Generation and analysis of infectious virus-like particles of uukuniemi virus (bunyaviridae): a useful system for studying bunyaviral packaging and budding. *J. Virol.* **80**, 10428–10435 (2006).
41. Komoto, S., Sasaki, J. & Taniguchi, K. Reverse genetics system for introduction of site-specific mutations into the double-stranded RNA genome of infectious rotavirus. *Proc. Natl. Acad. Sci. USA* **103**, 4646–4651 (2006).
42. Blight, K.J., Kolykhalov, A.A. & Rice, C.M. Efficient initiation of HCV RNA replication in cell culture. *Science* **290**, 1972–1974 (2000).

43. Taubenberger, J.K., Reid, A.H., Krafft, A.E., Bijwaard, K.E. & Fanning, T.G. Initial genetic characterization of the 1918 "Spanish" influenza virus. *Science* **275**, 1793–1796 (1997).
44. Reid, A.H., Fanning, T.G., Hultin, J.V. & Taubenberger, J.K. Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. *Proc. Natl. Acad. Sci. USA* **96**, 1651–1656 (1999).
45. Reid, A.H., Fanning, T.G., Janczewski, T.A. & Taubenberger, J.K. Characterization of the 1918 "Spanish" influenza virus neuraminidase gene. *Proc. Natl. Acad. Sci. USA* **97**, 6785–6790 (2000).
46. Basler, C.F. *et al.* Sequence of the 1918 pandemic influenza virus nonstructural gene (NS) segment and characterization of recombinant viruses bearing the 1918 NS genes. *Proc. Natl. Acad. Sci. USA* **98**, 2746–2751 (2001).
47. Reid, A.H., Fanning, T.G., Janczewski, T.A., McCall, S. & Taubenberger, J.K. Characterization of the 1918 "Spanish" influenza virus matrix gene segment. *J. Virol.* **76**, 10717–10723 (2002).
48. Reid, A.H., Fanning, T.G., Janczewski, T.A., Lourens, R.M. & Taubenberger, J.K. Novel origin of the 1918 pandemic influenza virus nucleoprotein gene. *J. Virol.* **78**, 12462–12470 (2004).
49. Taubenberger, J.K. *et al.* Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**, 889–893 (2005).
50. Takehisa, J. *et al.* Generation of infectious molecular clones of simian immunodeficiency virus from fecal consensus sequences of wild chimpanzees. *J. Virol.* **81**, 7463–7475 (2007).
51. Lee, Y.N. & Bieniasz, P.D. Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* **3**, e10 (2007).
52. Becker, M.M. *et al.* Synthetic recombinant bat SARS-like coronavirus is infectious in cultured cells and in mice. *Proc. Natl. Acad. Sci. USA* **105**, 19944–19949 (2008).
53. Wimmer, E. The test-tube synthesis of a chemical called poliovirus. The simple synthesis of a virus has far-reaching societal implications. *EMBO Rep.* **7**, S3–S9 (2006).
54. De Jesus, N., Franco, D., Paul, A., Wimmer, E. & Cello, J. Mutation of a single conserved nucleotide between the cloverleaf and internal ribosome entry site attenuates poliovirus neurovirulence. *J. Virol.* **79**, 14235–14243 (2005).
55. Toyoda, H., Yin, J., Mueller, S., Wimmer, E. & Cello, J. Oncolytic treatment and cure of neuroblastoma by a novel attenuated poliovirus in a novel poliovirus-susceptible animal model. *Cancer Res.* **67**, 2857–2864 (2007).
56. van der Werf, S., Bradley, J., Wimmer, E., Studier, F.W. & Dunn, J.J. Synthesis of infectious poliovirus RNA by purified T7 RNA polymerase. *Proc. Natl. Acad. Sci. USA* **83**, 2330–2334 (1986).
57. Molla, A., Paul, A.V. & Wimmer, E. Cell-free, *de novo* synthesis of poliovirus. *Science* **254**, 1647–1651 (1991).
58. Jiang, P. *et al.* Evidence for emergence of diverse polioviruses from C-cluster coxsackie A viruses and implications for global poliovirus eradication. *Proc. Natl. Acad. Sci. USA* **104**, 9457–9462 (2007).
59. Claverie, J.M. Viruses take center stage in cellular evolution. *Genome Biol.* **7**, 110 (2006).
60. Villarreal, L.P. Are viruses alive? *Sci. Am.* **291**, 100–105 (2004).
61. Ryan, F.P. Viruses as symbionts. *Symbiosis* **44**, 11–21 (2007).
62. Koonin, E.V., Senkevich, T.G. & Dolja, V.V. Compelling reasons why viruses are relevant for the origin of cells. *Nat. Rev. Microbiol.* **7**, 615 (2009).
63. Claverie, J.M. & Ogata, H. Ten good reasons not to exclude viruses from the evolutionary picture. *Nat. Rev. Microbiol.* **7**, 615 (2009).
64. Kitamura, N. *et al.* Primary structure, gene organization and polypeptide expression of poliovirus RNA. *Nature* **291**, 547–553 (1981).
65. Racaniello, V.R. & Baltimore, D. Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome. *Proc. Natl. Acad. Sci. USA* **78**, 4887–4891 (1981).
66. Smith, H.O., Hutchison, C.A., III, Pfannkoch, C. & Venter, J.C. Generating a synthetic genome by whole genome assembly: ΦX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. USA* **100**, 15440–15445 (2003).
67. Taubenberger, J.K. & Morens, D.M. 1918 Influenza: the mother of all pandemics. *Emerg. Infect. Dis.* **12**, 15–22 (2006).
68. Taubenberger, J.K., Morens, D.M. & Fauci, A.S. The next influenza pandemic: can it be predicted? *J. Am. Med. Assoc.* **297**, 2025–2027 (2007).
69. Taubenberger, J.K., Hultin, J.V. & Morens, D.M. Discovery and characterization of the 1918 pandemic influenza virus in historical context. *Antivir. Ther.* **12**, 581–591 (2007).
70. Pekosz, A., He, B. & Lamb, R.A. Reverse genetics of negative-strand RNA viruses: closing the circle. *Proc. Natl. Acad. Sci. USA* **96**, 8804–8806 (1999).
71. Hoffmann, E., Neumann, G., Hobom, G., Webster, R.G. & Kawaoka, Y. "Ambisense" approach for the generation of influenza A virus: vRNA and mRNA synthesis from one template. *Virology* **267**, 310–317 (2000).
72. Tumpey, T.M. *et al.* Existing antivirals are effective against influenza viruses with genes from the 1918 pandemic virus. *Proc. Natl. Acad. Sci. USA* **99**, 13849–13854 (2002).
73. Kash, J.C. *et al.* Global host immune response: pathogenesis and transcriptional profiling of type A influenza viruses expressing the hemagglutinin and neuraminidase genes from the 1918 pandemic virus. *J. Virol.* **78**, 9499–9511 (2004).
74. Kobasa, D. *et al.* Enhanced virulence of influenza A viruses with the haemagglutinin of the 1918 pandemic virus. *Nature* **431**, 703–707 (2004).
75. Kash, J.C. *et al.* Genomic analysis of increased host immune and cell death responses induced by 1918 influenza virus. *Nature* **443**, 578–581 (2006).
76. Pappas, C. *et al.* Single gene reassortants identify a critical role for PB1, HA, and NA in the high virulence of the 1918 pandemic influenza virus. *Proc. Natl. Acad. Sci. USA* **105**, 3064–3069 (2008).
77. Tumpey, T.M. *et al.* Pathogenicity and immunogenicity of influenza viruses with genes from the 1918 pandemic virus. *Proc. Natl. Acad. Sci. USA* **101**, 3166–3171 (2004).
78. Tumpey, T.M. *et al.* Pathogenicity of influenza viruses with genes from the 1918 pandemic virus: functional roles of alveolar macrophages and neutrophils in limiting virus replication and mortality in mice. *J. Virol.* **79**, 14933–14944 (2005).
79. Taubenberger, J.K. & Morens, D.M. The pathology of influenza virus infections. *Annu. Rev. Pathol.* **3**, 499–522 (2008).
80. Tumpey, T.M. *et al.* Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* **310**, 77–80 (2005).
81. Naffakh, N., Massin, P., Escriou, N., Crescenzo-Chaigne, B. & van der Werf, S. Genetic analysis of the compatibility between polymerase proteins from human and avian strains of influenza A viruses. *J. Gen. Virol.* **81**, 1283–1291 (2000).
82. Tumpey, T.M. *et al.* A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission. *Science* **315**, 655–659 (2007).
83. Stevens, J. *et al.* Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. *J. Mol. Biol.* **355**, 1143–1155 (2006).
84. Ryan, F.P. Human endogenous retroviruses in health and disease: a symbiotic perspective. *J. R. Soc. Med.* **97**, 560–565 (2004).
85. Bannert, N. & Kurth, R. Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl. Acad. Sci. USA* **101**, S14572–S14579 (2004).
86. Dewannieux, M. *et al.* Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* **16**, 1548–1556 (2006).
87. Enserink, M. Viral Fossil brought back to life. *Science NOW* **1101**, 4 (2006).
88. Keele, B.F. *et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**, 523–526 (2006).
89. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).
90. Baltimore, D. Expression of animal virus genomes. *Bacteriol. Rev.* **35**, 235–241 (1971).
91. Wimmer, E., Hellen, C.U. & Cao, X. Genetics of poliovirus. *Annu. Rev. Genet.* **27**, 353–436 (1993).
92. Holland, J. *et al.* Rapid evolution of RNA genomes. *Science* **215**, 1577–1585 (1982).
93. Eigen, M. Viral quasispecies. *Sci. Am.* **269**, 42–49 (1993).
94. Zolotukhin, S., Potter, M., Hauswirth, W.W., Guy, J. & Muzyczka, N. A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J. Virol.* **70**, 4646–4654 (1996).
95. Gromeier, M., Wimmer, E. & Gorbalenya, A.E. Genetics, pathogenesis and evolution of picornaviruses. in *Origin and Evolution of Viruses* (eds. Domingo, E., Webster, R.G. & Holland, J.J.) 287–343 (Academic, New York, 1999).
96. Porter, D.C. *et al.* Demonstration of the specificity of poliovirus encapsidation using a novel replicon which encodes enzymatically active firefly luciferase. *Virology* **243**, 1–11 (1998).
97. Kuge, S., Saito, I. & Nomoto, A. Primary structure of poliovirus defective-interfering particle genomes and possible generation mechanisms of the particles. *J. Mol. Biol.* **192**, 473–487 (1986).
98. Gutman, G.A. & Hatfield, G.W. Nonrandom utilization of codon pairs in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **86**, 3699–3703 (1989).
99. Moura, G. *et al.* Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS ONE* **2**, e847 (2007).
100. McKnight, K.L. & Lemon, S.M. Capsid coding sequence is required for efficient replication of human rhinovirus 14 RNA. *J. Virol.* **70**, 1941–1952 (1996).
101. Goodfellow, I. *et al.* Identification of a *cis*-acting replication element within the poliovirus coding region. *J. Virol.* **74**, 4590–4600 (2000).
102. Fowler, M., Beck, K., Brant, J., Opdyke, W. & Roberts, D. *Refactoring: Improving the Design of Existing Code* (Addison-Wesley, Boston, 1999).
103. Dunn, J.J. & Studier, F.W. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **166**, 477–535 (1983).
104. Knight, T.F. Engineering novel life. *Mol. Syst. Biol.* **1**, 2005 0020 (2005).
105. Atlas, R. *et al.* Statement on the consideration of biodefence and biosecurity. *Nature* **421**, 771 (2003).
106. Atlas, R. *et al.* Statement on scientific publication and security. *Science* **299**, 1149 (2003).
107. National Research Council. *Biotechnology in an Age of Terrorism* (The National Academies Press, Washington, DC, 2004).
108. Bügl, H. *et al.* DNA synthesis and biological security. *Nat. Biotechnol.* **25**, 627–629 (2007).
109. Garfinkel, M., Endy, D., Epstein, G. & Friedman, R. Synthetic genomics: options for governance. *Biosecur. Bioterror.* **5**, 359–362 (2007).