

Genome engineering

Peter A Carr¹ & George M Church²

For more than 50 years, those engineering genetic material have pursued increasingly challenging targets. During that time, the tools and resources available to the genetic engineer have grown to encompass new extremes of both scale and precision, opening up new opportunities in genome engineering. Today, our capacity to generate larger *de novo* assemblies of DNA is increasing at a rapid pace (with concomitant decreases in manufacturing cost). We are also witnessing potent demonstrations of the power of merging randomness and selection with engineering approaches targeting large numbers of specific sites within genomes. These developments promise genetic engineering with unprecedented levels of design originality and offer new avenues to expand both our understanding of the biological world and the diversity of applications for societal benefit.

Our capacity to understand and employ living systems has been intimately enmeshed with our ability to manipulate and test the instructive molecules. The ancient manipulation and testing of billion-base-pair DNA systems is evident in the diversity of dog breeds (spanning 3 logs in mass) and agricultural species relative to their wild ancestors. Moving in the direction of specific genetic control, the awesome power of merging chemistry with biology in the 1960s was evident in the use of synthetic oligonucleotides (oligos) to elucidate the fundamentals of the genetic code^{1,2} and in the 1970s to produce the first synthetic gene³ and first synthetic gene functionally tested *in vivo*⁴.

Since that time, benchmarks in the capacity to synthesize, manipulate and analyze DNA constructs have been achieved at exponential scales, in a manner reminiscent of Moore's Law⁵ for improvement in integrated circuit density. Figure 1 displays milestones in the *de novo* synthesis of DNA, from the first dinucleotide, dTdT, reported in 1955 by Michelson and Todd⁶, to the recent construction of a compact microbial genome (*Mycoplasma genitalium*; 582,970 bp)⁷. Figure 2 charts the improvement over time in the efficiency of DNA sequencing (in base pairs per dollar) as well as synthesis—both oligos and double-stranded DNA—trends that have also been noted for their exponential behavior^{8–10}. The tendency of some of these trends to increase in rate has been called the 'law of accelerating returns', emphasizing that this acceleration can go beyond even normal exponential growth¹⁰ (inflected upward on a log-linear graph).

Nevertheless, a paradoxical gap exists between our ability to synthesize and our ability to design valuable novel constructs. We can now produce oligos at 100 kbp/dollar and sequence DNA at 1 Mbp/dollar, but final gene-length DNA constructs are 2 bp/dollar. Getting a novel DNA construct to work as intended is a nontrivial process. Even modest deviations from natural genes cannot be taken for granted as functional, and must be tested thoroughly; much more so for *de novo* designs (in contrast, consider that combinatorial libraries of oligos can be constructed at 10¹⁴ bp/dollar¹¹, although the density of these pools for a given function can be low).

¹Massachusetts Institute of Technology, Media Lab, Center for Bits and Atoms, Massachusetts, Cambridge, USA. ²Harvard Medical School, Department of Genetics, Massachusetts, Cambridge, USA. Correspondence should be addressed to P.A.C. (carr@media.mit.edu)

Published online 9 December 2009; doi:10.1038/nbt1590

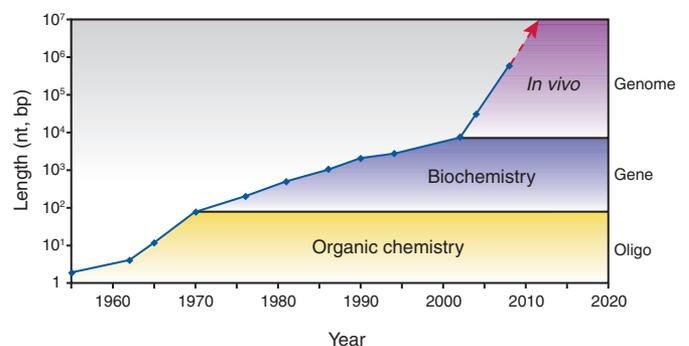


Figure 1 Milestones in the sizes of *de novo* synthesized DNA. Length is nucleotides (nt) for oligos before 1970, base pairs (bp) for double-stranded DNA from 1970 on. *In vitro* biochemical processing steps enabled the leap from oligos to genes, and *in vivo* processing steps (multiple cycles of cloning, sequencing and assembly) made possible the leap from genes to genomes. Future extensions of these tiers may include complex microbial communities or tissue organization. Data for this graph can be found in **Supplementary Table 1**.

Thus, although specific design techniques are improving, we expect that two of biology's special tools—variation and selection—will maintain an important role in the engineer's repertoire.

In discussing the idea of genome engineering, we apply this working definition: extensive and intentional genetic modification of a replicating system for a specific purpose. We leave the terms 'extensive' and 'replicating system' purposefully broad. On a practical level, we include work on viruses and bacteriophage, such as factoring a phage genome into many separate parts, and reorganizing and testing the new combinations¹². More examples are given in Figure 3. We exclude smaller replicative units, such as viroids (as small as 220 nucleotides (nt) in length). The tools and research interests of genome engineering also overlap those of genome-scale engineering. For this latter term, we refer to engineering of genetic systems on a similar scale (e.g., hundreds of genetic components or more) but not integrated into a single replicating (typically cellular) system. Examples of genome-scale engineering include producing and characterizing hundreds of different versions

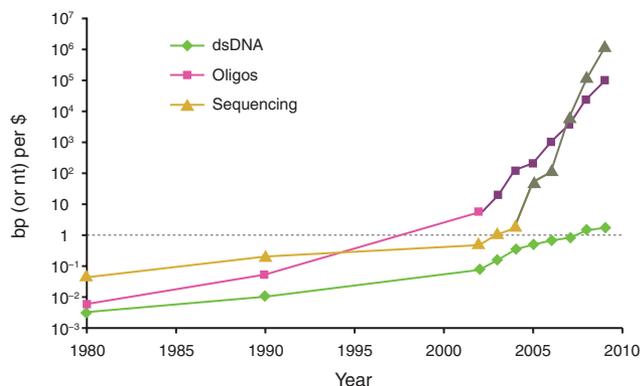


Figure 2 Efficiency trends in synthesis and sequencing over the past 30 years (base pairs per dollar). Double-stranded DNA synthesis ('gene synthesis') while improving rapidly, seems to lag behind the other two trends. The accelerated improvement in sequencing and oligo synthesis this past decade has been predicated on new miniaturization technologies (next-generation sequencing and microarray synthesis, respectively) to where the critical events take place on surface features measured in μm^2 . These transitions in technology are noted by a change to a darker line color. Commercial gene synthesis relies on both oligo synthesis (building blocks) and sequencing (verification and error control) but has yet to take effective advantage of these miniaturized formats. Some proofs of principle have been demonstrated^{41,51,113}. Data for this graph are detailed in **Supplementary Table 2**.

of a gene *in vitro*, or saturating a genome with single-gene knockouts (thousands of separate strains each with one modification)^{13,14}.

Genetic engineering as applied over the past several decades has most often employed small numbers of specific components (e.g., a single promoter and ribosome binding site coupled to a protein-encoding gene). Over the past decade, advanced designs have been engineered using larger numbers of components and with more complex interdependencies between them (see ref. 15). Several examples discussed below refer to current work at these scales (e.g., a dozen components), which in turn point the way toward future designs that may approach the genome scale.

Thus, genome engineering is genetic engineering applied to genomes (or at least large portions thereof). The tools used for this purpose are often those developed for smaller-scale genetic engineering, and applied in high-throughput fashion. In addition, genome engineering requires new technology specifically suited to that scale. For example, *de novo* construction of DNA molecules of up to a few thousand base pairs has relied on organic chemical and biochemical procedures. To generate an entire microbial genome, however, requires new methods for combining those smaller synthetic pieces (as detailed in ref. 7).

Genome engineering is in its infancy. The new techniques that have enabled initial work are modest compared with the needs for more tools at all stages: design, DNA construction and manipulation, implementation and testing, and debugging. Similarly, although potential applications are enticing, they are largely unproven at this point in time. As we discuss both these ideas and current progress, we begin with the motivations for expanding current gene and gene systems work to the genome scale, along with some goals that can only be achieved by dramatic engineering (or reengineering) of genomes.

Motivations for genome engineering

What are the factors that will continue driving DNA engineering toward increasingly larger and more complex designs? There is interplay between motivating applications and the technical advances, which enable larger scales while reducing costs. The pursuit of challenging goals leads to new

technology and the availability of that new technology encourages more ambitious pursuits. So why do we build genetic systems? Put another way, if you could design and build a genome, what would you want to make?

Build to understand. The Richard Feynman quote "what I cannot create I do not understand"¹⁶ is a favorite among synthetic biologists—and for good reason. Endy¹⁷ has pointed out that for some, synthetic biology is the pursuit of comprehending biological systems by trying to engineer them. (And we defer to that reference for greater exposition on the term 'synthetic biology'.) Much of the history of genetic engineering has been for the sake of understanding the molecular workings of life, frequently at the level of small numbers of parts (e.g., putting the coding sequence for a protein in a new genetic context such as a plasmid for easy manipulation and study). The complexity of such designs is increasing^{15,18}. For example, genetic circuits recently have been constructed to produce pattern formation in microbial communities¹⁹—a model system for studying the basic principles influencing developmental patterns in higher organisms. Furthermore, a central goal of the *M. genitalium* genome synthesis has been to produce a construction technology to examine minimal gene sets required for life²⁰.

Build for production. Living systems produce a staggering array of products tailored to human needs, including foodstuffs, materials and clothing. Recent years have seen substantial progress in metabolic engineering of microbes—combining, modifying and tuning many genes from different organisms for the sake of producing medicines²¹ and biofuels²². At the genome level, there is much interest in engineering a cellular 'chassis' for the optimal performance of such metabolic systems, involving large numbers of modifications to a microbial genome.

Build for protection. Genetic systems have also been designed to harness microbes as biosensors for various types of threats^{23,24} and bioremediation²⁵. Designs are currently in development for systems that allow microbes to hunt and destroy cancer cells^{26,27} and instruct one's own cells to minimize the risk of septic shock²⁸. An example of genome-wide engineering in this area would be the production of organisms with fundamentally altered codon usage—'orthogonal' genomes incapable of correctly translating genetic messages from other organisms and vice versa. At the scale of microbial genomes, this feature could prevent an engineered laboratory strain from using acquired genes to improve its fitness (e.g., antibiotic resistance genes) and from donating its specially engineered features to wild organisms. Plant genomes (e.g., crops) with this feature would be resistant to many wild pathogens (and uniquely susceptible to designed 'watchdog' pathogens). They would also be incapable of outcrossing with wild strains or conventional crops.

Build to creatively explore. An excellent array of explorations can be found at the website of the International Genetically Engineered Machines (iGEM) competition (<http://www.igem.org/>). These projects stand out as the accomplishments of interdisciplinary teams of undergraduates, operating in a time frame (months) conventionally considered brief for these types of efforts. There are too many intriguing applications to list here, but they include: (i) programming cells to communicate their growth state by emitting different odors; (ii) employing microbes as a photographic print medium; and (iii) many examples of genetically encoded logic and computation. Although the individual projects often fall into one or more of the above categories of understanding, production or protection, the entire undertaking serves as an experiment in the education and motivation of a new generation of synthetic biologists. In doing so, the students seize the opportunity to explore such questions as, How can I program a cell?

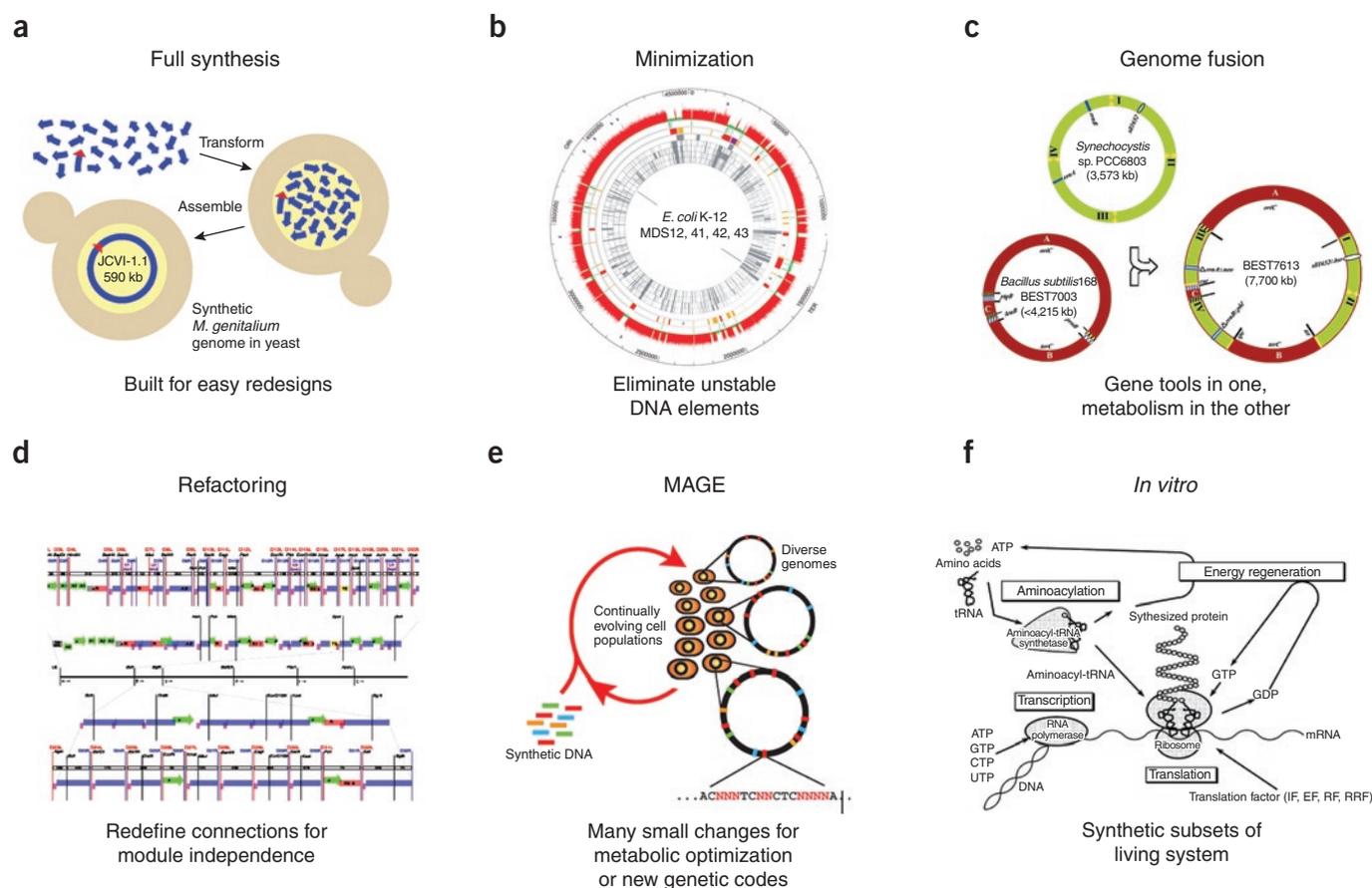


Figure 3 Examples of engineering at the genome scale. (a) Gibson *et al.*⁶¹ constructed the first all-synthetic microbial genome from commercially produced DNA cassettes. (b) Posfai *et al.*⁷⁶ deleted many large segments of the *E. coli* genome to eliminate unstable DNA elements. (c) Itaya *et al.*⁶⁶ transferred the majority of one (archaeal) genome into another (prokaryote) genome. (d) Chan *et al.*¹² decomposed the T7 bacteriophage genome into many reconfigurable modules. (e) Wang *et al.*³⁸ demonstrated a technique for making large numbers of targeted changes to a genome. (f) Shimizu *et al.*¹¹⁴ developed a purified translation system useful for *in vitro* prototyping of genetic functions without requiring moving genes into living cells. Note that manipulations of large DNA segments > 100 kilobase pairs in a, b and c relied heavily on *in vivo* recombination-based techniques.

In partnership with the Registry of Standard Biological Parts^{29,30} this work also helps tackle the question of how effectively biological systems can be engineered with composed, standardized and characterized genetic components. Wrestling with these questions is essential if we are to consider designing genetic systems the size of genomes.

Regardless of purpose, most projects in gene and genome engineering share a common set of tools and overall organization principles. In considering the accomplishments, challenges and opportunities of genome engineering, we examine four basic phases of an engineering project, applied here to genomes and other complex genetic systems: design, construction, implementation and/or testing and debugging (troubleshooting).

The design of genetic systems

Although Figure 1 emphasizes benchmarks achieved in genetic construct size, an even more significant focus should be on engineered function. Figure 4 compares the scale of a genetic engineering project (*x*-axis, in base pairs) to the proportion of that scale that was designed *de novo* (Fig. 4a) and the number of ‘design units’ manipulated (Fig. 4b). No one or two metrics are expected to unify such a broad range of designs and investigations. And although a portion of them can be said to have maximized some metric as a goal (genes deleted, proportion designed, degree of reorganization or synthesis scale), many also have no such goal in mind. Degree of difficulty, the

nature of the specific functions, complexity of new configurations and number of steps in an assembly hierarchy are certainly among the terms worthy of consideration.

Nevertheless, we see these projects falling into three broad categories of genetic design:

1. Design of small protein folds (up to 100% new sequence) and design of enzymatic activity (modifying scaffolds to 10–20% new sequence).
2. Design of genetic devices using naturally derived parts. These tend to display little *de novo* designed sequence; instead, new functions are derived from new configurations of existing parts. These have been well reviewed recently^{15,18}.
3. Manipulation of genomes by constructing, deleting and to some extent reorganizing components. These tend to be proof-of-principle reports pushing the limits of scale—often asking, How much of this can the cell tolerate?—but not of design. This statement is not a criticism, but an observation that genome engineering is in its infancy.

Figures 1, 2 and 4 together also illustrate an underlying principle: just as current DNA sequencing capacity dwarfs DNA synthesis capacity,

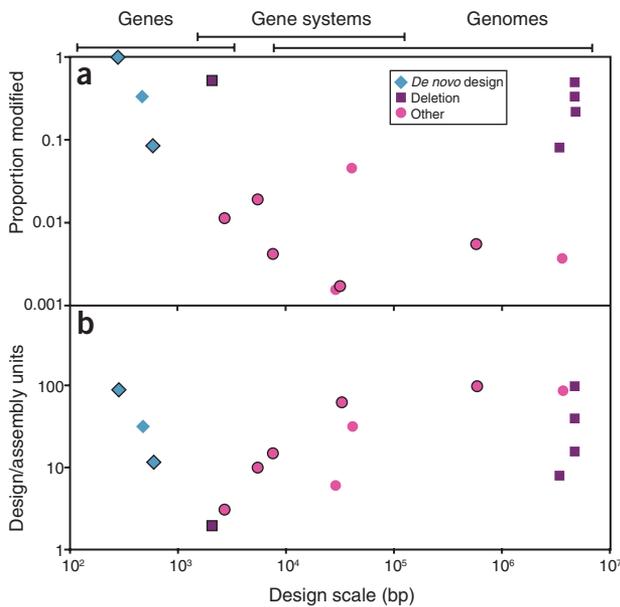


Figure 4 Current projects in genetic engineering exhibit a trade-off between the scale that can be manipulated and the scale at which one can effectively design. Projects are plotted according to (a) proportion modified or (b) number of design or assembly units. Symbols for constructs produced by *de novo* DNA synthesis are outlined in black. At the largest scales, the extent of modification has been essentially the manipulation of large natural DNA segments (synthesis, deletion, repositioning or cloning). However, the techniques developed to perform these operations set the stage for unprecedented new engineering efforts. Data for this graph are given in detail in **Supplementary Table 3**.

so DNA synthesis dwarfs current capacity for functional design and debugging. If the scale of available synthesis can be considered the size of the canvas on which we may paint, the available choices of brushes and colors are still rather modest.

The recent accomplishment by Gibson *et al.*⁷ at the J. Craig Venter Institute (JCVI; Rockville, MD, USA) illustrates the cutting edge of the field. The synthesis and assembly of a 582-kb pair *M. genitalium* genome exceeded by tenfold the size of any previously published *de novo* DNA construct (but did not reduce the cost per base pair). The extent to which this genome was reengineered, however, was small, primarily a handful of DNA watermarks—intended to show that the construct truly is synthetic. And even with these slight changes, getting the product to function proved challenging. Nevertheless, the choice of minimal modification seems especially prudent as the JCVI group seeks to ‘boot’ the genomic software in a fully operational cytoplasm and debug the ensuing design/assembly process. A failure to run the genetic operating system at this stage does not distinguish between problems with design and problems with the general production and ‘booting’ methods (see below). As this assembly technology becomes more robust, putting such synthetic capacity into the hands of genetic engineers will generate enticing new opportunities for design.

Standards, parts and design frameworks. In recent years, those in the synthetic biology community have championed the need for a standardized system of genetic parts, with the hopes of enabling sophisticated genetic systems design^{17,21,31–35}. A comparison is drawn with progress made last century in electronic design: standardization of parts, such as transistors and resistors, allowed mass production, generalized design and abstraction hierarchies. Such a hierarchy builds from the bottom

up, so that at each level of abstraction a specialist may take advantage of foundational work from more fundamental levels. One engineer may design single parts, the next a device based on such parts and a third ‘software’ using such devices. Integration of an advanced design framework based on this idea requires specialists at each level, as well as generalists broadly versed in the overall design system³⁶.

In the initial stages of synthetic biology, design has been closely linked to physical assembly. For example with BioBricks—the first major standard implemented—assembly is kept general and independent of specific parts through the use of a restriction-ligation scheme. Although this places some sequence limits on the part boundaries and requires keeping the restriction sites themselves out of the part sequences, the flexible framework has been employed to great effect. The value of the overall concept is underlined by the development of at least five alternative assembly standards³⁷. The long-term expectation in this area is that increasingly available DNA synthesis will make some of the current assembly restrictions unnecessary, and that new or modified standards will develop to take advantage of these resources.

Designs with standardized genetic parts may involve on the order of 10–20 parts—modest compared with the scale of a genome—but quite complex compared with most other genetic engineering. It is hoped that the use of such standards, coupled with vigorous characterization, will pave the way for new levels of design complexity. As this type of genetic programming approaches the scale of genomes, cloning contexts will of necessity shift from an emphasis on plasmids, to bacterial and yeast artificial chromosomes, to the primary chromosome(s) of the strain being engineered.

The interplay of design and randomness. Relative to most other fields of engineering, genome engineering has two huge potential advantages. One is the preexistence of highly evolved modules, which have some of the properties of careful design (albeit initially lacking specification sheets and without guarantees of interoperability or lack of cross-talk). The second advantage consists of the capacity to harness present-day (lab-scale) evolution and integrate the targeting of combinatorial changes genome-wide^{38,39}.

One general—and powerful—category of genetic engineering focuses on improving (or in some cases originating) function without a specific genetic design and instead takes a broader approach of directed evolution. A great body of successful metabolic engineering has benefited by applying this principle. Directed evolution has also been applied to the optimization of synthetic gene circuits⁴⁰. Future breakthroughs will probably focus on the ability to design and select from large collections of semi-synthetic DNA, with major challenges including the collecting and designing of biosensors⁴¹ and developing more complex selection criteria (e.g., involving cellular counters⁴²). Biosensing can be implemented using a second cell that requires the sensed molecule for growth (syntrophy)^{41,43}. Biosensors can also be obtained from allosteric regulatory proteins and RNA (riboswitches)⁴⁴. These can be evolved *in vitro* or *in vivo* to new specificities.

Computer-aided design tools (CAD). Once natural enzymatic and regulatory modules are adapted, refined and measured, they can be combined—at the drawing console—with a high degree of abstraction (ideally with intuitive graphics) while increasingly sophisticated computational methods handle ‘lower level’ steps. CAD is required at levels ranging from high-level design and simulation tools for synthetic biology⁴⁵ down to the detailed layout and sequences of oligos needed for multiplex assembly of genes or genomes^{46–48}. The need for CAD tools spans two extremes of design: first, combinatorial genetic

modifications that enable genome engineering with functional selection in metabolic engineering, where exploring all combinations is feasible (e.g., *cis* regulation of dozens of genes or more³⁸); and second, sequence-based screening, where the number of changes to be made is too large, selections are lacking or combinations are not needed (e.g., genome-wide codon conversion in *Escherichia coli*, where, for example, all TAG stop codons are to be converted to TAA).

CAD tools are also needed to generate metabolic and signaling pathways, including processes not yet found in nature. Looking forward, a key goal will be integrating and automating the various aspects from protein design⁴⁹ to compatibility of standards and intellectual property. Purnick and Weiss¹⁸ give a detailed listing of

current computational tools for design and analysis of genetic networks—many of these demonstrate features extensible to the genome scale, which will require handling hundreds to thousands of design components. The CLOTTHO software platform⁵⁰ is one example of an environment meant to be extensible to diverse design needs at different scales.

The construction of synthetic genetic material

At the simplest synthesis scale for DNA, single oligos are very affordable and available commercially on rapid time scales. For a pair of PCR primers, the time and cost of synthesis are more or less the same as the time and cost of shipping (frequently, next day shipping). Even so, for

Box 1 Joining DNA

DNA fragments can be joined in essentially one of four ways: chemical coupling, ligation, polymerization and recombination. These are summarized below.

Chemical coupling. Organic chemical synthesis of oligos proceeds by stepwise addition of single nucleotide bases to a growing chain (Fig. 5). The extensible end of this base (typically a 5' hydroxyl group) is protected from further reaction by a protecting group, which is removed for the next cycle. The majority of reaction failures are also terminated by addition of a capping group to halt further chain extension. This highly optimized chemistry can provide oligos with an average stepwise yield of 99% or higher, enabling the production of oligos up to 200 units in length (and on some occasions longer⁸³). Phosphoramidite chemistry dominates current synthetic methods, though alternative chemistries have also been used to great utility⁸⁴ and new developments have been recently reported⁸⁵. This stage of DNA synthesis is also distinct as the only one achieved without a template or complementary sequence (though sequence-independent ligation of larger segments for this purpose is conceivable). Instead, the single nucleotide building blocks are built into specific strings by choices designated at each step of the serial assembly.

Ligation. At the heart of nearly all synthetic gene-sized construction is self-assembly by means of programmed complementary base-pair interactions. After the specific association of two or more strands, the next step in producing larger pieces typically follows one of two enzymatic courses: ligation by a DNA ligase (Fig. 6), or oligo extension by a DNA polymerase (Fig. 7). The first gene syntheses employed ligation of oligos^{3,86} and some newer protocols employ ligases as well⁵⁴. Many protocols for assembling larger constructs also rely on ligation. Some of these have used short specific overhangs of 2–4 nt generated by restriction enzymes as the means of association⁶⁶, whereas Gibson *et al.*⁷ generated long overhangs using the 3' to 5' exonuclease activity of DNA polymerases.



Figure 5 Chemical synthesis of DNA. Nucleotide bases (purple circles) are added sequentially to the 5' end of the growing chain. Yellow arrowhead indicates the 3' end.



Figure 6 Ligation. DNA ligase makes backbone phosphate bonds (purple) connect strands of DNA (yellow).

Polymerization. Although polymerases had been well-studied long before, the introduction of the polymerase chain reaction, PCR⁸⁷, paved the way for a new set of gene synthesis protocols^{88,89} (Fig. 7). Polymerase-based protocols employ pairs of oligos which anneal and are extended, each oligo serving as both primer and template. The typical reaction is set up to employ a pool of oligos with several of these pairings occurring simultaneously in a thermocycled reaction, essentially growing progressively longer intermediates until the full-length product is obtained. The many variations on this theme have been well reviewed elsewhere^{90–92}. PCR-based overlap-extension methods can be used to generate fairly large constructs (e.g., 15 kbp by Tian *et al.*⁴⁶), but because the upper limits of long PCR may be ~50 kbp, these approaches seem unlikely to yield larger genomes by themselves. This does not exclude the possibility of alternative methods for genome assembly employing highly processive strand-displacing polymerases in a nonthermocycled *in vitro* context.

Recombination. Recombination methods have been employed both *in vitro* and *in vivo* for the assembly of DNA constructs (Fig. 8). A well-known *in vitro* example is the Gateway system (Invitrogen; Carlsbad, CA, USA), which uses phage λ site-specific recombination enzymes for both cloning and higher order assemblies. The other common site-specific recombination system is Cre-*loxP*. Homologous recombination systems have been used for manipulating quite large pieces of DNA, including double-stranded linear replacement⁹³, double-stranded circle-c integration⁹⁴ in *E. coli* and *Bacillus subtilis*⁶⁶, and single-stranded-oligo invasion of replication⁹⁵. Although generally used to manipulate one piece of DNA at a time, Gibson *et al.*⁶¹ recently demonstrated the simultaneous recombination of 25 linear DNAs ~22 kbp each in yeast. An advantage of homologous recombination approaches is that no exogenous sequences are required for targeting, giving the possibility for scar-free assemblies.



Figure 7 DNA joining by polymerization. Overlapping pairs of oligos (yellow) that anneal serve as both primer and template for extension (purple) by DNA polymerase (in direction of arrows).

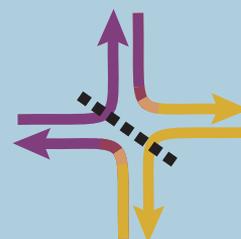


Figure 8 DNA joining by recombination. Two DNA duplexes (yellow, purple) are brought together to form a four-stranded junction. When resolved across the dotted line, new hybrid DNA duplexes result.

large synthesis projects, these costs can be considerable (e.g., 1 million base pairs of double-stranded DNA would currently cost \$200,000 or more for the oligos alone before assembly, assuming synthesis at the 10 or 25 nanomole scale).

Gene-scale synthesis is also becoming highly commoditized—commercial synthetic gene providers crossed the ‘buck a base’ threshold (\$1 per base pair) some time ago, and are currently near half that price. The price of such synthetic DNA continues to drop, with trends noted as comparable to Moore’s Law, dropping a factor of 1.5 per year^{8–10} (Fig. 2). It seems plausible that in a few (perhaps 3–5) years, commercial gene synthesis could reach the same level of convenience as for synthetic oligos: a cost and time on par with overnight shipping. When this condition is met, much of the work currently done to manipulate DNA in research labs will be outsourced. Instead of cloning into vectors stored in those labs, custom or standard vectors could simply be resynthesized on demand. To enable this flexible design structure, the synthesis community may employ intellectual property distributions comparable to VLSI (very large-scale integration) library licensing in microchip manufacture²¹.

Reaching this tipping point will likely depend on emerging technologies for highly parallelized and miniaturized synthesis^{46,51,52}. Bypassing or dramatically modifying current time- and cost-intensive steps, such as cloning, could also be required. High-quality error correction methods^{46,53,54} may also allow some applications to proceed without conventional cloning and sequencing, or *in vitro* single-molecule cloning may be adopted⁵⁵.

The technology of synthesis and assembly. A great variety of specific protocols exist for generating DNA constructs of different sizes. Although there are too many individual techniques to discuss in detail here, they are most easily presented by factoring their particulars as combinations of a few common core elements, regardless of synthesis scale. The broad categories of ‘DNA joining’, ‘assembly organization’ and ‘error control’ are detailed in Box 1 (Figs. 5–8), Box 2 (Figs. 9–12) and Box 3 (Fig. 13), respectively. Certain combinations of these elements are especially popular, but others represent untapped potential, such as chemical coupling of large constructs.

In addition, the environmental context in which these procedures are implemented has in many ways defined the limits of synthesis scale. Figure 1 denotes the three tiers of contexts that enable synthesis of increasingly larger DNA targets. Each builds on the one below: oligo synthesis is performed via organic chemical reactions, oligos are assembled into genes via biochemical reactions and genomes are produced by manipulating gene length constructs taking advantage of one or more *in vivo* processing steps. It is worth discussing each of these steps in more detail.

The first enabling technology for the assembly of genes was that of oligo synthesis^{6,56}. An excellent short history of this type of organic chemistry has been written by Hogrefe⁵⁷. Synthetic oligos are the building blocks for larger pieces of genetic material. Although a large proportion of oligos are short (~20 nt) and used for processes such as PCR and DNA sequencing, those used for gene synthesis are longer, typically 30 nt or more.

The addition of an *in vitro* biochemical step was a crucial advance for the report of the first synthetic gene in 1970 by Khorana and coworkers³. The authors used the newly characterized enzyme T4 ligase to link oligos of 8 to 20 nt in length, generating the structural gene for a 77-bp yeast alanine tRNA. Such processes have been used to assemble products up to several thousand base pairs in length^{58,59}. The synthetic product at this stage may variably be referred to as a gene, a synthon, a chunk, a cluster, a cassette, a segment or a part. Though these definitions overlap, they are not all equivalent. In this context, the term ‘gene’ often refers to a protein reading frame, possibly with additions relating to transcription

and translation. A part connotes function, and may be as small as, for example, a promoter, or as large as a complex genetic device assembled from several smaller parts (as per discussion of standardized parts above). The other terms are more general and may include multiple genes or a fraction of one. Regardless of name, these constructs are generally pieces of double-stranded DNA, assembled from two or more oligos. The construction process is very often referred to as gene synthesis.

The first synthetic genome reported was that of poliovirus, published in 2002 (ref. 60). An important aspect of that synthesis was the use of processing in living systems. Though the final destination of synthetic genes before that time was also typically an *in vivo* context, the distinction made in the top tier of Figure 1 is for processes for which *in vivo* handling was a crucial and fundamental intermediate assembly step. Thus Cello *et al.*⁶⁰ first synthesized subsets of the full-length viral genome and cloned these separately into plasmids, which were subsequently used to transform living cells. The resulting clones were sequenced and perfect clones were selected, where possible. In the absence of a perfect clone for a given segment, site-directed mutagenesis was performed to repair the DNA. Once error-free clones were obtained, propagating these and larger assemblies *in vivo* ensured a minimal rate of introduction of new errors, a strategy used in all the large assemblies discussed here.

More recent methods have relied on DNA recombination to perform assembly of very large segments of DNA *in vivo*, with yeast proving especially apt for this purpose. The JCVI team assembling the *M. genitalium* genome employed native recombination mechanisms to produce their full-length product, demonstrating that more laborious *in vitro* handling was unnecessary at some earlier steps⁶¹. There also exists potential to take advantage of organisms with more extensive recombination capacity, such as *Deinococcus radiodurans*, which can reassemble its own genome after extreme fragmentation⁶².

As increasingly larger syntheses are attempted, the fragility of long (genome-length) DNA strands is expected to become a more challenging issue. Using cells to perform not only assembly and amplification but also DNA transfer is likely to become routine for assemblies a million base pairs and larger. Immediate opportunities are apparent in conjugative transfer of DNA between bacterial cells⁶³ and yeast mating and recombination^{64,65}.

Figure 14 displays two recent examples of large-scale DNA construction, characterized by choices of assembly technology. In addition to specific choices for joining, organization and error control (Boxes 1–3), a degree of parallelization and tuning are inherent in most of these processes, although emphasized more in some than in others. The different stages can be considered essentially modular—methods applied at one stage (e.g., oligo synthesis) need not be tightly coupled to the next (e.g., gene synthesis). Some approaches are better suited to specific stages, but not necessarily limited to them. For example, oligo synthesis chemistry is fairly standardized around serial condensation of phosphoramidite monomers in organic solvents. Gibson *et al.*⁷ have noted that the large-scale assembly method they pioneered need not be limited to applications of *de novo* synthesis but should perform equally well for DNA extracted directly from natural sources. This is true for many of the methods detailed here. Similarly, techniques that have been applied to the large-scale manipulation of extracted natural DNA^{12,64,66,67} are also worth considering for genome-scale DNA construction. The largest such construct so far is the 10 Mbp minichromosome of Kuroiwa and coworkers⁶⁸.

Automation. A central feature of efficient synthetic DNA production is automation and scale-up. DNA synthesis companies generally employ fluid-handling robots and moderately high density (96-, 384- or 1,536-well) plate formats common to the biotech industry.

These approaches are sufficient for high-throughput production on the order of megabase pairs per month. A detailed example of a production pipeline has been published by Hellinga and coworkers⁶⁹.

Another approach to high-throughput DNA production is microfluidic processing. Specific advantages to this approach include the following: first, minimization of reagent and consumable use; second, less dependence on expensive robotics; and third, direct coupling to high-density microarray-fabricated sources of oligo building blocks. Regarding this last advantage, we note the potential represented by high-density arrays that contain more genetic information—as oligos—in a few square centimeters (many megabase pairs) than any commercial gene synthesis provider currently assembles in one month. Early reports using such oligos for gene synthesis have removed oligos from array surfaces and manipulated them in macroscopic volumes (e.g., 10–20 μ l), frequently requiring parallel amplification of all oligos in a pool^{46,51}. Microfluidic devices present a unique opportunity to instead confine these oligos in small volumes, obtaining useful nanomolar concentrations from less than

femtomolar oligo yields⁵². Spatially separating the large numbers of array oligos into many small compartments (nanoliters) for parallel syntheses is also expected to reduce the complexity of diverse oligo pools (several thousand sequences or more) to manageable levels. Applications for this purpose are currently still in their early stages⁵². Such devices have been forecast as a principal enabling technology for dramatically pushing down future consumer costs⁷⁰. Other avenues will exploit growing libraries of prefabricated parts or genomes requiring merely hundreds of changes as enabled by the multiplex automated genome engineering (MAGE) approach from our groups³⁸.

Implementation—‘booting’ a designed genetic system

After construction or extreme modification of a genome, another special challenge remains: the DNA software must be ‘booted’. Booting refers to the pulling oneself up by one’s bootstraps, or more recently ‘booting’ a computer, wherein the software is loaded into ‘blank’ computer hardware⁷¹. How is one to get an entire genome running? One

Box 2 Assembly organization

Several different organizational schemes are available for assembling pieces of DNA into larger fragments. These can be categorized in assembly in series, by hierarchy, in parallel or by pooling. These are described below.

Serial. This simply refers to adding one unit at each stage of a synthetic process (Fig. 9). Organic chemical syntheses of oligos intrinsically use this approach: the choice of nucleotide added (A, C, G or T) at each stage determines how specific sequences are constructed in the absence of any DNA template. Serial assembly has been employed at every scale of DNA assembly, including the large-scale cloning of one genome (3.5 Mbp) into another species of cell⁶⁶. One advantage of serial processes is control⁹⁶. Even in cases where higher throughput methods may be preferred, serial construction can prove the fall-back option for difficult projects, such as the synthesis of low-complexity or repeat-intensive sequences.

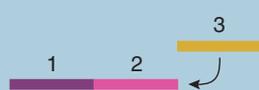


Figure 9 DNA assembly in series. Single subunits (whether single nucleotides or large DNA cassettes) are added one at a time sequentially.

Hierarchical. These schemes provide a potent balance of throughput and control. Pieces of DNA are joined in multiple stages, frequently combining in pairs at each stage (Fig. 10). For example, eight pieces joined as pairs produces four larger assemblies; joining these as pairs produces two even larger assemblies, and combining these yields the final construct. Employing this hierarchy requires three stages of assembly compared with seven stages for a serial approach. The advantage grows considerably at larger scales (more pieces) as N pieces can be combined with on the order of $\log_2 N$ stages (versus $N-1$ for a serial approach). An increase in size by a factor of ten requires only a few more stages. A version of this strategy was employed in the initial *M. genitalium* genome synthesis, with combinations of two to four DNA segments at each stage⁷. (See below, however for a pooled version.) Another advantage of hierarchical assembly strategies is that the intermediates produced can be helpful for debugging problems in complex designs. This challenge is also expected to become increasingly difficult at the longer synthesis scales.

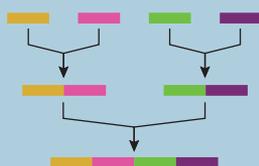


Figure 10 Hierarchical DNA assembly. Segments are joined together in subsets, producing successively larger constructs at each stage.

Parallel. It is important to recognize a degree of parallelization inherent to some steps of a large-scale DNA construction effort (Fig. 11). All the oligos synthesized for producing a gene or a genome are expected to be produced in parallel. The same is true for assembling several DNA cassettes on the way to a multigene construct. Microarrays are an exceptional example of parallelizing the serial process of oligo synthesis. Oligo microarrays can be synthesized *in situ*, typically on a few cm^2 of surface, with complexities as high as over one million different specific oligo sequences.



Figure 11 DNA assembly in parallel. Most large synthesis projects require some degree of parallel processing.

Pooling. Performing several joining reactions in the same mixture has proven extremely advantageous to improving the efficiency of DNA construction (Fig. 12). At the level of gene synthesis, both ligase-based and polymerase-based assemblies are often performed with pools of oligos. See Khorana *et al.*⁸⁶ as well as Dillon and Rosen⁸⁹ for early examples. One gene synthesis protocol, thermodynamically balanced inside-out (TBIO), combines advantages of both serial and pooled strategies⁹⁷: oligos to make a DNA segment are combined in one pool and extended via a thermocycled polymerase reaction, but the arrangement of oligos allows only incremental growth of the product at each step. Pooled assembly reactions have been performed with groups of >200 oligos to produce a 5.3-kbp phage genome⁵⁹, and a pool of ~600 oligos has been used to assemble 21 separate genes that were later hierarchically combined to yield a 15-kbp product⁴⁶. Pools can also be used *in vivo*^{38,61}.

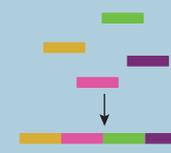


Figure 12 Pooling approaches. Multiple DNA segments are joined in a single reaction.

Box 3 Error control

Errors in synthetic DNA at any length scale need to be considered carefully. Even an error rate as low as 1 in 10,000 bp⁵³ can be a major concern if the product of interest is of that scale (10⁴) or larger. Two major types of mistakes are worth considering: failure to assemble (global error) and mutations in an assembled product (local error). However, far more detailed information is available on the latter. Although there are many options, error control is not explicitly required at every stage. For example, some single-gene syntheses may simply sequence a small number of clones of the gene product to find one that has no errors.

Select. When it can be arranged, selection for function or viability can dramatically reduce errors in the surviving clones. Several examples have shown utility for single genes⁹⁸, gene systems (e.g., plasmids^{58,99}) and a small phage genome⁵⁹. A more general form of this concept involves fusing synthetic open-reading frames (protein-encoding or not) in-frame to a downstream selectable gene^{69,100}. Most deletions in the synthetic genes should then give rise to frameshifts so that the host cell does not produce the downstream gene it needs to survive. As deletions (especially point deletions) can be the most common defect in chemically synthesized DNA^{53,69,101} and are generally the most deleterious to function, this approach can improve the quality of a construct substantially when the desired product is a single protein reading frame. Beyond this are selections for proper folding (**Fig. 13**, 1) and solubility^{102,103}.

Tune. Most stages of DNA construction rely on some degree of optimization to minimize the opportunities for flawed pieces to occur. Examples of tuning include the extensive optimization that has accompanied commercial oligo synthesis (average stepwise yields in excess of 99%), use of stringent annealing temperatures to favor joining of oligos without mismatches⁵⁴ and selecting the most high-fidelity polymerases for amplification. Because commercial oligo manufacture has generally been optimized for other applications, tuning this organic synthesis specifically for the purpose of gene synthesis is desirable. For example, oligos from some providers lead to single-base deletions as the primary error^{53,101}, whereas others lead mainly to point substitutions¹⁰⁴. Selection of such parameters as reagent concentrations and reaction times are likely to lead to these differences. All oligo syntheses are influenced by the degree to which undesired trace water is present during coupling reactions, as well as the age of the phosphoramidite reagents. Cerrina and coworkers¹⁰⁵ have demonstrated the utility of optimizing *in situ* oligo synthesis in microarrays specifically for the purpose of gene synthesis.

Repair. This category includes all manner of approaches that modify a DNA site containing an error (**Fig. 13**, 2). One such example used enzymes that cut at the site of an error (in the form of a DNA mismatch) coupled to exonuclease activity to degrade the defective sequence, and subsequent resynthesis by polymerases¹⁰⁶. Many applications have used information from DNA sequencing to fix flawed clones through site-directed mutagenesis^{60,107}. Note that in the former case no specific knowledge of the errors is involved, whereas in the latter it is an absolute requirement. There are a number of *in vivo* repair pathways that have not yet been adapted for synthetic DNA production. Thus, we might expect to see more applications of this type in the near future.

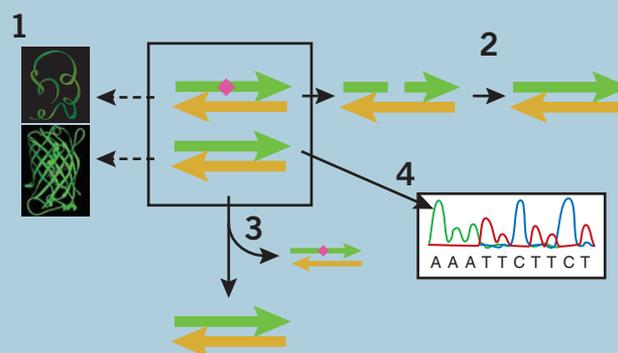


Figure 13 Error control. Two copies of a DNA assembly (green and yellow strands) are shown. One copy contains an error, such as a single-base substitution or indel (shown as a red diamond). Approaches to reduce the prevalence of such error-containing DNA include (1) selecting for clones that encode properly folded proteins that, for example, fluoresce, (2) DNA repair via nuclease excision and polymerase resynthesis or site-directed mutagenesis, (3) purification to remove defective DNA or (4) sequencing to identify error-free clones.

Purify. This category refers to methods for removing undesired species from the set of DNA molecules (**Fig. 13**, 3). Purification of oligos can be performed before gene synthesis to improve either the reliability of the process. Purification to remove defective oligos also has the potential to reduce mutations in the final assembled gene. Because one of the most common errors observed is a single-base deletion, purification of an oligo of length N must be stringent enough to remove defective oligos of length $N - 1$. Hybridization-based purification of oligos has also demonstrated a dramatic improvement in error rate⁴⁶. At the level of gene-sized pieces of DNA, the use of mismatch binding proteins has proved effective for separation of mismatched (error-containing) species^{53,108}. Cleavage of mismatch duplexes has also been accomplished with endonucleases, followed by electrophoretic separation¹⁰⁹ or selective degradation⁵⁴. This latter method possesses the additional advantage of all-fluid handling steps (*in vitro* biochemistry) without the need for additional separations.

Sequence. DNA sequencing is the gold standard for ascertaining the quality of a synthetic construct (**Fig. 13**, 4). For single-gene assemblies, sequencing is often the final stage of picking a winning clone. For most of the larger syntheses reported^{7,46,60,104}, it has been expedient to clone and sequence intermediate fragments, often of length 400–600 bp. One advantage of this size is that with error rates typical for commercial oligos, an error-free specimen can be identified after sequencing only a few clones. Also, typical read-lengths for conventional Sanger sequencing are slightly longer than this range. Thus, sequencing cloned constructs of this size can be performed using primers generic to the vector instead of specific to the construct. Going forward, the integration of second-generation (high-density microarray) synthesis⁴⁶ and sequencing¹¹⁰ may require multiplex tagging and/or selective release from oligo microarrays, as well as *in vitro* molecular cloning^{55,111,112} where single molecules are amplified by PCR to produce clone-like isolates, which are then sequenced.

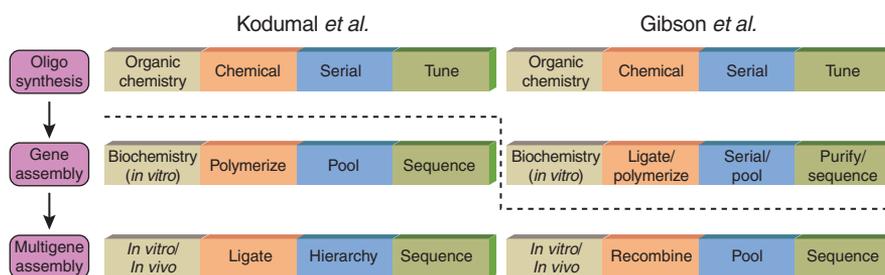


Figure 14 The general process of assembling large genetic constructs. Two recent examples^{61,104} are diagrammed. Three major tiers are shown, indicating the different stages for 1, the synthesis of oligos; 2, the assembly of oligos into larger double-stranded synthons (usually in the 0.5–5 kbp size range and frequently a gene); and 3, assembly of these units into larger constructs. Colored boxes correspond to choices made regarding the assembly environment (gold), joining mode (orange, see **Box 1**), assembly organization (blue, see **Box 2**) and error control (green, see **Box 3**). A dotted line indicates when the project flow crosses over from commercially provided services—oligos for Kodumal *et al.*¹⁰⁴ and 5–7 kbp cassettes for Gibson *et al.*⁶¹. Thus, for the latter report, some gene assembly elements likely vary between the three vendors employed. ‘*In vitro/vivo*’ refers to the toggling back and forth between cellular and test tube (aqueous) environments at stages in the processing.

may consider the possibility of creating cellular ‘ghost’ cells, with transcription and translation machinery, but no genome of their own. These could be generated through cell division mutants, by internal digestion of the host genomic DNA or by reassembly of membrane and cytoplasmic fractions. The synthetic genome could then be transformed into these cell-like compartments. The JCVI team has reported a related technique of transforming one type of cell with an extracted donor genome (with the host genome originally intact, but later selected against or possibly digested)⁷². Nevertheless, because the hardware environment of living systems is frequently redefined by the resident genome, the degree of designed modification (Fig. 3a) raises compatibility issues—including codon usage, restriction and/or modification systems⁷³, chromosome stability⁶⁸ and regulatory incompatibilities. In contrast, booting a synthetic bacteriophage or virus genome poses much less difficulty—though it is by no means trivial—typically using the type of cell host the wild-type virus is compatible with, ensuring fairly optimal compatibility^{12,67,74}.

A second approach is to incrementally alter an existing genome while the cell continues to operate. (This requires genome compatibility from one stage to the next, though not necessarily between the original and final genome states.) This strategy is a continuum of ‘traditional’ genetic engineering and much akin to altering an operating system while a computer is running; it has proven useful in *E. coli* genome-scale deletion studies^{75,76}. In some cases, the incoming genome could stay largely silent during *in vivo* transfers and assembly. Having the core set of new transcription, translation and replication functions under dual, inducible control would allow them to be switched on and produced first with the host machinery, later assuming the dominant role in the cell. The two-genome fusion of Itaya and coworkers⁶⁶ employed the first part of this strategy, with the incoming genome segments largely dormant. Such ‘running patch’ methods are also proving efficient in attempts to reprogram the genetic code of *E. coli*, with the goal of enabling nonnatural amino acid applications and blocking effective horizontal gene transfer (F.J. Isaacs & P.A.C. *et al.*, unpublished data).

A third approach would be complete breakdown into *in vitro* modules, which also permits a radical degree of redesign and debugging⁷⁷. Taking such rebooting to an extreme in terms of degree of modification would be a mirror-image genome⁷⁸, where every stereocenter in every biomolecule would be inverted relative to life as we know it. Such a system would be incompatible with any existing cytoplasm, and would require true bootstrapping from a minimal set of biochemical functions (replication, transcription and translation). Nevertheless, even this grand challenge for

vitalism and the accidental nature of life’s chirality leans heavily on our knowledge of working living systems. Constructing a genome entirely from *de novo* designs or selections from true random-polymers (a la polynucleotide aptamer libraries) is considerably further off.

Troubleshooting—debugging the bugs

The various approaches shown in Figure 3 also illustrate the dual robustness and fragility of living systems. Vast amounts of a genome can be completely deleted without apparent harm to the organism and even yield improved performance^{75,76}. At the same time, very modest changes expected to be functionally invisible can cause reduced fitness⁶⁰ and single-point mutations can easily be fatal. Whereas a remaining challenge in genome engineering is to improve our ability to design more robustly, designs at these scales should also place an emphasis on planned troubleshooting.

Biological complexity represents a special challenge for genome engineers. Across the different fields of engineering, many kinds of design may have conflicts between the working parts—components that do not connect as intended, or which in combination display unexpected behaviors. But for biological systems especially, the background environment is still very incompletely understood when contrasted with other disciplines, such as electronics design. Though a given genome sequence may be known, the functions of many predicted proteins typically remain unknown and the relationships between known functions incompletely mapped. The interactions between a given designed genetic system and its cellular environment may display both general components (e.g., drains on cell resources, such as ATP and ribosome translation capacity) and specific components (e.g., undesired action of a designed DNA-binding protein on host genes). Various technologies in development have the potential to reduce this complexity, such as routing protein synthesis for the engineered system via an orthogonal ribosome⁷⁹ or running a genetic circuit in an existing organelle, such as a mitochondrion or engineered cellular vesicle. Nevertheless, many types of designs will not necessarily be amenable to such isolation.

Just as the value of design and assembly hierarchies has been emphasized above, hierarchical debugging strategies will greatly facilitate successful implementation of designed genetic systems. There are two relevant hierarchies to consider. The first mirrors that used for design. All the separate genetic parts of a designed system should be tested singly in parallel, or in as simple a representation as possible. Where possible, combinations of simple parts into larger units should be performed along lines of linked function, so that these combinations can also be tested en route to the final assembly. Parts assembly strategies such as for BioBricks are intended

to facilitate this process. In contrast, an all-or-nothing assembly and test strategy leaves the designer with little hint as to what went wrong.

The second hierarchy is that of the testing environment. *In silico* (virtual, simulation), *in vitro* and *in vivo* contexts all have a role to play. Numerical simulations involving ordinary differential equations have been used to test many genetic circuit designs. *In vitro* tests, such as with transcription/translation mixes, have the potential to quickly profile simple parts and test devices in the absence of complicating factors from the whole cell. Thus, failures at this testing level are more likely to reveal fundamental flaws in the parts themselves. Neither virtual nor *in vitro* simulation can replace *in vivo* testing, as they cannot effectively represent the complexity of the biological environment. Even so, problems revealed at these earlier testing stages are likely to indicate real concerns for the *in vivo* context—and with the right resources in place will typically be much faster to test. Large-scale, rapid and cost-effective DNA synthesis will be an enabling technology for any troubleshooting that involves modifying or making new parts for a design. Thus, transitions back and forth between the drawing table and the laboratory can be kept minimal.

Perspectives

We have commented on technological advances that will enhance molecular engineers' capacity to design and build at increasingly larger genetic scales. There are a number of research goals in this area likely to be achieved in the next several years that merit additional comment.

One of these is all *in vitro* processing of large-scale syntheses, particularly when coupled to protein synthesis and functional assays—a sharp departure from the noted trend of increasingly biological processing. These syntheses may be on scales up to megabase pairs, probably exploiting parallelism rather than stitching DNA fragments together into genome lengths. One genome-scale application would be synthesizing and testing hundreds of versions of proteins designed around a specific function, a scale-up in complexity (and scale-down in time and cost per gene) of recent work such as that from the Baker group⁴⁹. Another use would be the profiling of many genes for evaluation in constructing *in vivo* biosynthetic pathways. Starting from sequence data, cellulases from 100 or more organisms could be constructed and compared for performance in biofuel feedstock production, or enzyme components mixed and matched to optimize terpenoid production for pharmaceutical biomanufacture. A third version of these *in vitro* applications would be to start from DNA sequences obtained as clinical data (e.g., the entire genetic diversity of a specific HIV patient's viral load) and resynthesize the corresponding genes and proteins to test for compatibility with choices of drug regimens (or evaluate a new drug). This degree of personalized medicine, coupled with expected advances in DNA sequencing will be facilitated by microfluidic integration of oligo synthesis, gene synthesis⁵², translation of genes to proteins and assaying⁸⁰. The level of integration possible (sequence data in, assay data out) will also serve to decouple physical sample acquisition from the experimental molecules, which in turn will be decoupled from the data's final destination (that is, the clinician and patient). Centralized high-performance resources could thus serve a world-wide community with rapid response on the order of a day.

These forms of *in vitro* rapid prototyping have potential for evaluating not just single-gene designs, but also more complex systems—genetic circuits, metabolic pathways and even genomes. Such approaches will not represent the complexity such designs will face in their final *in vivo* settings. Rather, they provide the opportunity to characterize performance in a defined and adjustable setting (e.g., chemical environment) for single components and specified combinations of those components.

As the scale of synthesis and assembly continues to grow, new methods will also be developed to deal with the challenges of large, more fragile, genome-lengths of DNA. 'All-biological' handling of DNA transfer and

recombination events seems especially likely, such as with bacterial conjugation, yeast mating and mammalian cell fusions. Furthermore, although a majority of efforts in genome engineering have focused on single-celled organisms, emerging tools for large-scale genetic manipulation in higher organisms^{68,81} are also expected to prove of great use. It would not be surprising if the next tier extending the trends of Figure 1 relies on using cellular communities to execute designs that exceed the complexity of a single genome.

By way of summary, consider the utility of a microorganism with a reengineered genome that combines the following features: (i) removal of DNA elements that contribute to genetic instability, such as insertion sequences⁷⁶ and phase variation systems; (ii) restructuring of the genetic code to ensure no cross-compatibility with other organisms⁸²; (iii) simplification of the genetic code to allow easy inclusion of new, nonnatural amino acids (F.J. Isaacs & P.A.C. *et al.*, unpublished data); (iv) removal of metabolic pathways that drain the resources of the cell and are not needed in a laboratory or biomanufacturing capacity⁷⁵; and (v) incorporation of systems that allow rapid and efficient tuning of genetic components, taking advantage of directed evolution³⁸. These properties (and others) would contribute to the production of a cellular chassis that would be the starting point for a wide range of genetic programming applications. The 'genetic isolation' of a unique genetic code would give a safer context to perform advanced bioengineering—unable to make use of exogenous genes that encode toxicity factors or antibiotic resistance, and unable to effectively donate its own special genetic features to wild strains. Increased genetic stability would provide a more consistent engineering environment, with genetic variation occurring primarily only where directed. Removal of unneeded components that use up cellular resources provides the opportunity to direct more of these resources for producing useful compounds, leading to higher yields. We expect many different versions (and species) of such reengineered strains to be of great utility, with some in highly specialized application roles, and others serving for broad general use.

We now find ourselves at an intriguing turning point. The current scale of *de novo* synthesis and reuse of engineered genetic parts seems to be leading directly to new modes of design and exploration. At the gene level, many simple gene modifications, such as cloning and mutagenesis, are being replaced with automated synthesis, assembly and even characterization. This transition will also allow entry into the field of designers who need not be experts in traditional DNA manipulation techniques. Some goals will be accessible only by genome-wide methods, such as the reformatting of the genetic code by altering tens of thousands of native codon assignments, chirality or pH/thermal stabilities genome-wide. In the longer term, the scaling of genome engineering will lead us toward engineering synthetic ecosystems, multicellular developmental systems (including human) and general programmable matter.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors wish to acknowledge helpful discussions with J. Jacobson, A. Forster, B. Chow, members of the Church and Jacobson laboratories and the comments of anonymous reviewers. Grant support has been provided by Department of Energy-GTL, National Science Foundation (SynBERC, Center for Bits and Atoms, and Genes and Genomes Systems Cluster) and Defense Advanced Research Projects Agency programmable matter programs.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at <http://www.nature.com/naturebiotechnology/>.

1. Khorana, H.G. Nucleic acid synthesis in the study of the genetic code. in *Nobel Lectures: Physiology or Medicine (1963–1970)* 341–369 (Elsevier Science, New York; 1972).

2. Nirenberg, M. The genetic code. in *Nobel Lectures: Physiology or Medicine (1963–1970)* 372–395 (Elsevier Science, New York; 1968).
3. Agarwal, K.L. *et al.* Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature* **227**, 27–34 (1968).
4. Ryan, M.J., Brown, E.L., Sekiya, T., Kupper, H. & Khorana, H.G. Total synthesis of a tyrosine suppressor tRNA gene. XVIII. Biological activity and transcription, in vitro, of the cloned gene. *J. Biol. Chem.* **254**, 5817–5826 (1979).
5. Moore, G.E.P. Cramming more components onto integrated circuits. *Electronics* **38**, 114–117 (1965).
6. Michelson, A.M. & Todd, A.R. Nucleotides Part XXXII. Synthesis of a dithymidine dinucleotide containing a 3':5'-internucleotidic linkage. *J. Chem. Soc.* 2632–2638 (1955).
7. Gibson, D. *et al.* Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* **319**, 1215–1220 (2008).
8. Carlson, R. The pace and proliferation of biological technologies. *Bio Secur. Bioterror.* **1**, 203–214 (2003).
9. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
10. Kurzweil, R. The 21st century: a confluence of accelerating revolutions. Keynote address at 8th Annual Foresight Conference, Bethesda, Maryland, November 3, 2000. <http://www.kurzweilai.net/meme/frame.html?main=/articles/art0184.html> (2000).
11. Wilson, D.S. & Szostak, J.W. In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611–647 (1999).
12. Chan, L.Y., Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Mol. Syst. Biol.* **1**, 2005.0018 (2005).
13. Hutchison, C.A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169 (1999).
14. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
15. Lu, T.K., Khalil, A.S. & Collins, J.J. *Nat. Biotechnol.* **27**, 1139–1150 (2009).
16. Hawking, S.W. *The Universe in a Nutshell* (Bantam Books, New York, 2001).
17. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).
18. Purnick, P.E. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.* **10**, 410–422 (2009).
19. Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H. & Weiss, R. A synthetic multicellular system for programmed pattern formation. *Nature* **434**, 1130–1134 (2005).
20. Glass, J.I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. USA* **103**, 425–430 (2006).
21. Baker, D. *et al.* Engineering life: building a fab for biology. *Sci. Am.* **294**, 44–51 (2006).
22. Rude, M.A. & Schirmer, A. New microbial fuels: a biotech perspective. *Curr. Opin. Microbiol.* **12**, 274–281 (2009).
23. Rider, T.H. *et al.* A B cell-based sensor for rapid identification of pathogens. *Science* **301**, 213–215 (2003).
24. Kobayashi, H. *et al.* Programmable cells: interfacing natural and engineered gene networks. *Proc. Natl. Acad. Sci. USA* **101**, 8414–8419 (2004).
25. Cases, I. & de Lorenzo, V. Genetically modified organisms for the environment: stories of success and failure and what we have learned from them. *Int. Microbiol.* **8**, 213–222 (2005).
26. Anderson, J.C., Clarke, E.J., Arkin, A.P. & Voigt, C.A. Environmentally controlled invasion of cancer cells by engineered bacteria. *J. Mol. Biol.* **355**, 619–627 (2006).
27. Cheong, I. *et al.* A bacterial protein enhances the release and efficacy of liposomal cancer drugs. *Science* **314**, 1308–1311 (2006).
28. Ciglic, M. *et al.* Engineered human cells: say no to sepsis. *IET Synth. Biol.* **1**, 13–16 (2007).
29. Kelly, J.R. *et al.* Measuring the activity of BioBrick promoters using an in vivo reference standard. *J. Biol. Eng.* **3**, 4 (2009).
30. Kosuri, S., Kelly, J.R. & Endy, D. TABASCO: A single molecule, base-pair resolved gene expression simulator. *BMC Bioinformatics* **8**, 480 (2007).
31. McDaniel, R. & Weiss, R. Advances in synthetic biology: on the path from prototypes to applications. *Curr. Opin. Biotechnol.* **16**, 476–483 (2005).
32. Canton, B., Labno, A. & Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.* **26**, 787–793 (2008).
33. Lucks, J.B., Qi, L., Whitaker, W.R. & Arkin, A.P. Toward scalable parts families for predictable design of biological circuits. *Curr. Opin. Microbiol.* **11**, 567–573 (2008).
34. Heinemann, M. & Panke, S. Synthetic biology—putting engineering into biology. *Bioinformatics* **22**, 2790–2799 (2006).
35. Ellis, T., Wang, X. & Collins, J.J. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.* **27**, 465–471 (2009).
36. Knight, T. Idempotent vector design for standard assembly of Biobricks. *Dspace@MIT* <http://hdl.handle.net/1721.1/21168> (2003).
37. Biobrick Formats. <http://bbf.openwetware.org/Standards/Technical/Formats.html>.
38. Wang, H.H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).
39. Santos, C.N. & Stephanopoulos, G. Combinatorial engineering of microbes for optimizing cellular phenotype. *Curr. Opin. Chem. Biol.* **12**, 168–176 (2008).
40. Haseltine, E.L. & Arnold, F.H. Implications of rewiring bacterial quorum sensing. *Appl. Environ. Microbiol.* **74**, 437–445 (2008).
41. Pfeleger, B.F., Pitera, D.J., Newman, J.D., Martin, V.J. & Keasling, J.D. Microbial sensors for small molecules: development of a mevalonate biosensor. *Metab. Eng.* **9**, 30–38 (2007).
42. Friedland, A.E. *et al.* Synthetic gene networks that count. *Science* **324**, 1199–1202 (2009).
43. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
44. Link, K.H. & Breaker, R.R. Engineering ligand-responsive gene-control elements: lessons learned from natural riboswitches. *Gene Ther.* **16**, 1189–1201 (2009).
45. Computational Tools. http://openwetware.org/wiki/Computational_Tools.
46. Tian, J. *et al.* Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
47. Python Synthetic Biology libraries. <http://bitbucket.org/chapmanb/synbio/>.
48. Czar, M.J., Cai, Y. & Peccoud, J. Writing DNA with GenoCAD. *Nucleic Acids Res.* **37**, W40–47 (2009).
49. Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
50. CLOTHO. http://2008.igem.org/Team:UC_Berkeley_Tools (2008).
51. Richmond, K.E. *et al.* Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.* **32**, 5011–5018 (2004).
52. Kong, D.S., Carr, P.A., Chen, L., Zhang, S. & Jacobson, J.M. Parallel gene synthesis in a microfluidic device. *Nucleic Acids Res.* **35**, e61 (2007).
53. Carr, P.A. *et al.* Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.* **32**, e162 (2004).
54. Bang, D. & Church, G.M. Gene synthesis by circular assembly amplification. *Nat. Methods* **5**, 37–39 (2008).
55. Ben Yehzekel, T. *et al.* De novo DNA synthesis using single molecule PCR. *Nucleic Acids Res.* **36**, e107 (2008).
56. Tener, G.M., Gilham, P.T., Razzell, W.E., Turner, A.F. & Khorana, H.G. Studies on the chemical synthesis and enzymatic degradation of desoxyribo-oligonucleotides. *Ann. NY Acad. Sci.* **81**, 757–775 (1959).
57. Hofgre, R. A short history of oligonucleotide synthesis. http://www.trilinkbiotech.com/tech/oligo_history.pdf
58. Stemmer, W.P., Cramer, 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).
59. Smith, H.O., Hutchison, C.A. III, Pfannkoch, C. & Venter, J.C. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. USA* **100**, 15440–15445 (2003).
60. 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).
61. Gibson, D.G. *et al.* One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome. *Proc. Natl. Acad. Sci. USA* **105**, 20404–20409 (2008).
62. Zahradka, K. *et al.* Reassembly of shattered chromosomes in *Deinococcus radiodurans*. *Nature* **443**, 569–573 (2006).
63. Gemski, P. Jr., Wohlhieter, J.A. & Baron, L.S. Chromosome transfer between *Escherichia coli* HFR strains and *Proteus mirabilis*. *Proc. Natl. Acad. Sci. USA* **58**, 1461–1467 (1967).
64. Larin, Z., Taylor, S.S. & Tyler-Smith, C. A method for linking yeast artificial chromosomes. *Nucleic Acids Res.* **24**, 4192–4196 (1996).
65. Kouprina, N. & Larionov, V. Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast *Saccharomyces cerevisiae*. *Nat. Protoc.* **3**, 371–377 (2008).
66. Itaya, M., Tsuge, K., Koizumi, M. & Fujita, K. Combining two genomes in one cell: stable cloning of the *Synechocystis* PCC6803 genome in the *Bacillus subtilis* 168 genome. *Proc. Natl. Acad. Sci. USA* **102**, 15971–15976 (2005).
67. Donaldson, E.F. *et al.* Systematic assembly of a full-length infectious clone of human coronavirus NL63. *J. Virol.* **82**, 11948–11957 (2008).
68. Kuroiwa, Y. *et al.* Manipulation of human mitochondrial genomes to carry greater than megabase-sized chromosome inserts. *Nat. Biotechnol.* **18**, 1086–1090 (2000).
69. Cox, J.C., Lape, J., Sayed, M.A. & Helling, H.W. Protein fabrication automation. *Protein Sci.* **16**, 379–390 (2007).
70. Newcomb, J., Carlson, R. & Aldrich, S.C. *Genome Synthesis and Design Futures: Implications for the US Economy* (Bio Economic Research Associates, Cambridge, MA, USA, 2007).
71. <http://en.wikipedia.org/wiki/Bootstrapping>.
72. Lartigue, C. *et al.* Genome transplantation in bacteria: changing one species to another. *Science* **317**, 632–638 (2007).
73. Lartigue, C. *et al.* Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science* (2009).
74. Tumpey, T.M. *et al.* Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* **310**, 77–80 (2005).
75. Mizoguchi, H., Sawano, Y., Kato, J. & Mori, H. Superpositioning of deletions promotes growth of *Escherichia coli* with a reduced genome. *DNA Res.* **15**, 277–284 (2008).
76. Psofai, G. *et al.* Emergent properties of reduced-genome *Escherichia coli*. *Science* **312**, 1044–1046 (2006).
77. Forster, A.C. & Church, G.M. Towards synthesis of a minimal cell. *Mol. Syst. Biol.* **2**, 45 (2006).
78. Forster, A.C. & Church, G.M. Synthetic biology projects in vitro. *Genome Res.* **17**, 1–6 (2007).
79. Rackham, O. & Chin, J.W. A network of orthogonal ribosome x mRNA pairs. *Nat. Chem. Biol.* **1**, 159–166 (2005).

80. Maerkl, S.J. & Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
81. Ivics, Z. *et al.* Transposon-mediated genome manipulation in vertebrates. *Nat. Methods* **6**, 415–422 (2009).
82. Church, G.M. Safeguarding biology. *Seed* **20**, 84–86 (2009).
83. Ciccarelli, R.B., Gunyuzlu, P., Huang, J., Scott, C. & Oakes, F.T. Construction of synthetic genes using PCR after automated DNA synthesis of their entire top and bottom strands. *Nucleic Acids Res.* **19**, 6007–6013 (1991).
84. Reese, C.B. The chemical synthesis of oligo- and poly-nucleotides: a personal commentary. *Tetrahedron* **58**, 8893–8920 (2002).
85. Sierzchala, A.B. *et al.* Solid-phase oligodeoxynucleotide synthesis: a two-step cycle using peroxy anion deprotection. *J. Am. Chem. Soc.* **125**, 13427–13441 (2003).
86. Khorana, H.G. *et al.* Total synthesis of the structural gene for the precursor of a tyrosine suppressor transfer RNA from *Escherichia coli*. 1. General introduction. *J. Biol. Chem.* **251**, 565–570 (1976).
87. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).
88. Prodromou, C. & Pearl, L.H. Recursive PCR: a novel technique for total gene synthesis. *Protein Eng.* **5**, 827–829 (1992).
89. Dillon, P.J. & Rosen, C.A. A rapid method for the construction of synthetic genes using the polymerase chain reaction. *Biotechniques* **9**, 298–300 (1990).
90. Stewart, L. & Burgin, A.B. Whole gene synthesis: a Gene-O-Matic future. *Front. Drug Des. Discov.* **1**, 297–341 (2005).
91. Czar, M.J., Anderson, J.C., Bader, J.S. & Peccoud, J. Gene synthesis demystified. *Trends Biotechnol.* **27**, 63–72 (2009).
92. Xiong, A.S. *et al.* Chemical gene synthesis: strategies, softwares, error corrections, and applications. *FEMS Microbiol. Rev.* **32**, 522–540 (2008).
93. Kolisnychenko, V. *et al.* Engineering a reduced *Escherichia coli* genome. *Genome Res.* **12**, 640–647 (2002).
94. Link, A.J., Phillips, D. & Church, G.M. Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization. *J. Bacteriol.* **179**, 6228–6237 (1997).
95. Li, X.T. *et al.* Identification of factors influencing strand bias in oligonucleotide-mediated recombination in *Escherichia coli*. *Nucleic Acids Res.* **31**, 6674–6687 (2003).
96. Van den Brulle, J. *et al.* A novel solid phase technology for high-throughput gene synthesis. *Biotechniques* **45**, 340–343 (2008).
97. Gao, X., Yo, P., Keith, A., Ragan, T.J. & Harris, T.K. Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Res.* **31**, e143 (2003).
98. Withers, S.T., Gottlieb, S.S., Lieu, B., Newman, J.D. & Keasling, J.D. Identification of isopentenol biosynthetic genes from *Bacillus subtilis* by a screening method based on isoprenoid precursor toxicity. *Appl. Environ. Microbiol.* **73**, 6277–6283 (2007).
99. Mandecki, W., Hayden, M.A., Shallcross, M.A. & Stotland, E. A totally synthetic plasmid for general cloning, gene expression and mutagenesis in *Escherichia coli*. *Gene* **94**, 103–107 (1990).
100. Davis, C.A. & Benzer, S. Generation of cDNA expression libraries enriched for in-frame sequences. *Proc. Natl. Acad. Sci. USA* **94**, 2128–2132 (1997).
101. Hoover, D.M. & Lubkowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43 (2002).
102. Drew, D. *et al.* A scalable, GFP-based pipeline for membrane protein overexpression screening and purification. *Protein Sci.* **14**, 2011–2017 (2005).
103. Cabantous, S. & Waldo, G.S. *In vivo* and *in vitro* protein solubility assays using split GFP. *Nat. Methods* **3**, 845–854 (2006).
104. Kodumal, S.J. *et al.* Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Natl. Acad. Sci. USA* **101**, 15573–15578 (2004).
105. Kim, C. *et al.* Progress in gene assembly from a MAS-driven DNA microarray. *Microelectron. Eng.* **83**, 1613–1616 (2006).
106. Fuhrmann, M., Oertel, W., Berthold, P. & Hegemann, P. Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Res.* **33**, e58 (2005).
107. Marsic, D., Hughes, R.C., Byrne-Steele, M.L. & Ng, J.D. PCR-based gene synthesis to produce recombinant proteins for crystallization. *BMC Biotechnol.* **8**, 44 (2008).
108. Binkowski, B.F., Richmond, K.E., Kaysen, J., Sussman, M.R. & Belshaw, P.J. Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Res.* **33**, e55 (2005).
109. Young, L. & Dong, Q. Two-step total gene synthesis method. *Nucleic Acids Res.* **32**, e59 (2004).
110. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
111. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).
112. Hutchison, C.A. III, Smith, H.O., Pfannkoch, C. & Venter, J.C. Cell-free cloning using phi29 DNA polymerase. *Proc. Natl. Acad. Sci. USA* **102**, 17332–17336 (2005).
113. Zhou, X. *et al.* Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.* **32**, 5409–5417 (2004).
114. Shimizu, Y., Kanamori, T. & Ueda, T. Protein synthesis by pure translation systems. *Methods* **36**, 299–304 (2005).