Generating a Synthetic Genome by Whole Genome Assembly: phiX174 Bacteriophage from Synthetic Oligonucleotides

Hamilton O. Smith, Clyde A. Hutchison III*, Cynthia Pfannkoch, and J. Craig Venter

The Institute for Biological Energy Alternatives
1901 Research Blvd., #600
Rockville, MD 20850

*Current address: Department of Microbiology and Immunology, The University of North Carolina at Chapel Hill, Chapel Hill, NC27599-7290

Corresponding author: J. Craig Venter, the Institute for Biological Energy Alternatives, 1901 Research Blvd., #600, Rockville, MD 20850, Telephone: 301-309-3404, Fax: 301-309-3434, jcventer@tcag.org

Classification: Biological Sciences:Biochemistry
Manuscript information: The number of text pages (14); The number of words in the abstract (244) and the total number of characters in the paper (37,732 without figures)

Contributed by: J. Craig Venter

Abstract

We have improved upon the methodology and dramatically shortened the time required for accurate assembly of 5-6 kb segments of DNA from synthetic oligonucleotides. As a test of this methodology we have established conditions for the rapid (14 days) assembly of the complete infectious genome of bacteriophage ϕX174 (5,386 bp) from a single pool of chemically synthesized oligonucleotides. The procedure involves three key steps: 1) Gel purification of pooled oligonucleotides to reduce contamination with molecules of incorrect chain length, 2) Ligation of the oligonucleotides under stringent annealing conditions (55ºC) to select against annealing of molecules with incorrect sequences, and 3) Assembly of ligation products into full length genomes by polymerase cycling assembly (PCA); a non-exponential reaction in which each terminal oligonucleotide can be extended only once to produce a full-length molecule. We observed a discrete band of full-length assemblies upon gel analysis of the PCA product, without any PCR amplification. PCR amplification was then used to obtain larger amounts of pure full-length genomes for circularization and infectivity measurements. The synthetic DNA had a lower infectivity than natural DNA, indicating approximately one lethal error per 500 bp. However, fully infectious ϕX174 virions were recovered following electroporation into *E. coli*. Sequence analysis of several infectious isolates verified the accuracy of these synthetic genomes. One such isolate had exactly the intended sequence. We propose to assemble larger genomes by joining separately assembled 5-6 kb segments; approximately 60 such segments would be required for a minimal cellular genome.

Chemical synthesis of life in the laboratory has been a standing challenge to synthetic organic chemistry since Wöhler's synthesis of urea in 1828 (1) and the doctrine of spontaneous generation was put to rest by Pasteur in 1864, (an address delivered by Louis Pasteur at the "Sorbonne Scientific Soiree" of April 7, 1864). With an understanding of the genetic role of DNA, much work has focused on the synthesis of oligonucleotides and genes. The synthesis of the 207-bp gene for tyrosine suppressor tRNA in 1979 by Khorana and 17 co-workers was a monumental undertaking (2). Since then the automated DNA synthesizer has been developed based upon fundamental advances in synthetic methods from the laboratories of Letsinger (3, 4) and Caruthers (5, 6).

In 1999 we described a minimal prokaryotic genome based upon results from random whole genome transposon mutagenesis that inactivated one gene per cell (7). Using this approach, approximately 300 essential genes for self-replicating cellular life were described, and we proposed to make a synthetic chromosome to test the viability of the 300 genes hypothesis (7). Prior to attempting synthesis of a microbial chromosome, we commissioned an independent bioethical review of our proposed scientific plan (8). After more than one year of deliberation, the reviewers concluded that we were taking a reasonable scientific approach to an important biological question. The broader implications of the creation of life in the laboratory can now be considered as a realistic possibility. However, there are several technical barriers to the synthesis of microbial chromosome-sized stretches of DNA that are hundreds of thousands to millions of nucleotides long, the most notable being the contamination of the oligonucleotides by truncated species. While such oligonucleotides are highly useful as primers for PCR amplification and DNA sequencing, only small (a few hundred bp) synthetic genes can generally be accurately and directly synthesized without multiple repair/selection steps. For example, the recent report (9) of the assembly of a partially active poliovirus from cloned synthetic segments of DNA from which polio genomic RNA (7,440 bases) could be transcribed was quite complex and took many months to accomplish. First segments 400 to 600 bp in length were individually assembled. These were cloned and 5 to 15 isolates of each were sequenced to find one that was correct or readily correctable by oligonucleotide mutagenesis. These segments were then assembled into three larger segments of the polio genome, recloned, and finally assembled to produce a full-length product. This slow process would not be practical for synthesizing a 300,000 bp chromosome. We have now improved the methodology for synthesis of multi-gene segments of a genome as a step toward synthesis of a cellular genome. As a test of this methodology we have established conditions for global assembly of the infectious genome of bacteriophage ϕX174 (5,386 bp) from a single pool of chemically synthesized oligonucleotides. ϕX174 presents no known hazard since it infects only certain enteric bacteria and is not a human, plant, or animal pathogen. Therefore, its choice for synthesis serves to separate safety issues from ethical considerations and other potential risks associated with synthetic genomics.

ϕX174 (ϕX) is the prototypical minute icosahedral bacteriophage that was first characterized in the Sinsheimer laboratory beginning in the 1950s. The ϕX virion contains a circular single-stranded DNA that replicates through a double-stranded

replicative form of DNA (RF). Its chromosome was the first DNA molecule purified to homogeneity (10), and consequently ϕX DNA has been used in many landmark experiments. It was the first viral DNA shown to be infectious in 1961, and in 1967 Goulian, Kornberg and Sinsheimer (11) demonstrated that ϕX DNA synthesized with DNA polymerase using the intact genome as a template, was infectious. This feat was hailed as "life in the test tube" (see for example 12). The ϕX174 genome was also the first DNA completely sequenced by Sanger and co-workers in 1978 (13). The sequence describes a remarkably compact gene organization with several cases of genes expressed by translating overlapping regions of the DNA in two different reading frames. Thus, ϕX provides an historic and a demanding test bed for accurate synthetic chromosomes.

We set out to develop a procedure by which any 5 to 6 kb segment of DNA can be quickly and easily assembled from a single pool of synthetic oligonucleotides with high fidelity. In 1995, Stemmer et al (14) used a polymerase cycling reaction to assemble a 2.7kb plasmid from a single pool of oligonucleotides that was able to replicate in *E. coli*. Synthesis of the larger ϕX genome, however, provides a more stringent test for optimizing oligonucleotide assembly methods. Its compact genetic organization makes it relatively intolerant to mutation as compared to a plasmid, which may require only a replication origin and a drug resistance marker. We report here a procedure (Fig. 1), that utilizes sequential ligation and polymerase cycling reactions to accomplish the assembly of the ϕX genome an order of magnitude more rapidly than with previous methods described above (9) and with great savings in the sequence analysis of intermediate products, as well as human effort. This method should make it possible to assemble a minimal microbial genome from as few as 60 synthetic 5-6 kb segments.

Materials and Methods

Synthetic oligonucleotides. Oligonucleotides were obtained from Integrated DNA Technologies (IDT) in 96-well format and normalized to 100μM each. Based on the accompanying mass spectrometer data, as well as information from the company, we ascertained that on average, each 42-mer contained about 50% truncated species. To purify the oligonucleotides, 10μl of each of the top (or bottom) strand oligonucleotides were pooled in two separate pools (TOP1-130 and BOT131-259), dried, and dissolved in 50μl of water. 20μl of concentrated pool plus 20μl of formamide was heated to 95°C for 2 min and 10μl aliquots were loaded into 1.5mm x 15mm slots of a preparative 12% sequencing gel and electrophoresed at 1300V for ~4h. The bands, which migrated close to the xylene cyanol marker, were visualized with a handheld 254nm UV lamp, and excised. The gel was extruded through a tuberculin syringe into 0.7ml of TE buffer, frozen at - 20°C overnight, eluted at 37°C on a rotating wheel for 1h, and filtered through a glass wool-plugged 1ml Eppendorf pipette tip. The recovered oligonucleotides were ethanol-precipitated and dissolved in 50μl of water.

Phosphorylation of oligonucleotides. The oligonucleotides were phosphorylated prior to ligation. A reaction mixture (100μl) containing 20μl of purified top (or bottom) oligonucleotides, 10μl of 10X T4 polynucleotide kinase buffer (New England BioLabs,

NEB), 1mM ATP, and 40 units of T4 polynucleotide kinase (NEB) was incubated at 37°C for 1h. The reaction was terminated by phenol-chloroform extraction and ethanol precipitation. The phosphorylation reaction was repeated a second time. After extraction and precipitation, the oligonucleotides were dissolved in 50µl of water.

Ligation reactions. The ligation reaction mixture (100µl) contained 10µl of top 5'P-oligonucleotides, 10 µl of bottom 5'P-oligonucleotides, 10µl of 10X Taq ligation buffer (NEB), and 60 µl of water. The mixture was heated to 95°C for 2 min and slowed cooled over 30 min to 55°C. Ten µl of Taq ligase (40units/µl, NEB) was added and incubation was continued at 55°C for 18h. The reaction was terminated by phenol-chloroform extraction and ethanol-precipitation. The products were dissolved in 90µl of TE buffer.

Polymerase cycling assembly of oligonucleotides (PCA). PCA was carried out in reaction mixtures (50µl) containing 5µl of 10X Advantage 2 buffer, 5µl of 10X dNTP mixture, 1µl of 50X HF polymerase mixture (Clonetech Advantage HF 2 PCR Kit Cat.# K1914-1) and either 0.2, 0.5, or 1µl of the Taq ligation product. The polymerase mixture contains an N-terminal deletion mutant of Taq DNA polymerase that lacks 5'-exonuclease activity, and Deep Vent polymerase (NEB) with 3'-exonuclease proofreading activity. Cycling parameters were 94°C for 15 sec, slow cool at -0.1°C/sec to 55°C, annealing at 55°C for 2', and extension at 72°C for 6 min. 35 cycles were carried out followed by finishing at 72°C for 5 min.

PCR amplification of the synϕX DNA molecules produced by PCA. The PCR reaction mixtures (25µl) contained 2.5µl of 10X advantage 2 buffer, 2.5µl of 10X dNTPs, 0.5µl of HF polymerase mixture, 0.5µl of each second stage PCA product, and 1µl each of 10µM top-1 and bot-259 oligonucleootides (unpurified). PCR parameters were 94°C for 15 sec, 55°C for 30 sec, and 72°C for 6 min for 25 cycles, finishing with 72°C for 5 min. The PCR reaction mixtures were pooled, phenol-chloroform extracted, ethanol precipitated and redissolved in 10µl of TE buffer.

Conversion of linear synϕX DNA molecules to infectious circular molecules. The pooled PCR-amplified synϕX DNA was cleaved with PstI and the linear DNA was gel-purified to yield approximately 1µg of linear synϕX. The linear synϕX molecules were circularized by ligation under dilute conditions (~1µg/ml) with T4 ligase (NEB) using recommended conditions. The ligation mixture was phenol-chloroform extracted, ethanol precipitated and re-dissolved in 10µl of TE/5 buffer in preparation for infectivity testing.

Assay of ϕX DNA infectivity. One µl of synϕX ligation product (an estimated 3 to 5ng of circular molecules based on ethidium bromide staining intensity) was electroporated into DH10B cells (InVitrogen), immediately diluted with 500µl of SOC broth, and then aliquoted into two screw-capped glass culture tubes (A and B) containing 2ml of KC broth each (10g Bacto Tryptone, 5 g KCl, and 0.5 ml 1 M $CaCl_2$ per liter). The tubes were rotated at 37°C for approximately 40min, and then 225µl of 2mg/ml lysozyme (Sigma), and 47.5µl of 0.5M EDTA, pH 8 were added, and incubation was continued on

ice for 30 minutes. The tubes were freeze-thawed twice in dry ice-ethanol to release the synϕX phage. Aliquots of the synϕX-A and -B lysates were plated undiluted in 3ml of top agar containing 0.3 ml of log phase HF4704 at $5x10^8$ cells/ml on LB plates. Phage plaques were visualized after 6 to 18 hours of incubation at 37°C.

DNA sequencing. Plaques were picked directly into 50µl PCR reactions using a 10µl micropipette tip, and subjected to 30 cycles of amplification consisting of 10sec at 94°C, 30sec at 55°C and 6min at 72°C. Fifty µl of a mixture of 40 µl of shrimp alkaline phosphatase (1 unit/µl), 8 µl of exonuclease I (20 units/µl), and 752 µl of water was added to each PCR reaction mixtures and digestion was at 37°C for 45min and 45°C for 15min, followed by 72°C for 15min to inactivate the enzymes. The synϕX DNA was gel-purified prior to carrying out standard sequencing reactions and sequencing on a Model 3730 XL ABI sequencer. Both strands were sequenced and depth was generally >2-fold.

Rationale and Theoretical Considerations.

The steps we used to accomplish the synthesis of infectious ϕX double-stranded RF DNA from a single pool of oligonucleotides are summarized in Fig. 1. In developing this strategy we considered how impurities present in oligonucleotide preparations can result in assembly errors and mutations in the final product. This led us to purify the oligonucleotides prior to assembly. We also analyzed and determined the theoretical endpoints and limitations of the two basic assembly steps: ligation and polymerase cycle assembly (PCA).

Oligonucleotide purity. If automated DNA synthesizers produced pure oligonucleotides of the programmed sequence, then assembly of long double-stranded DNA molecules would be straightforward. In reality, only approximately 50% of the molecules in preparations such as those used in our work have the correct chain length. The population of other molecules includes truncated species capped at the growing end, and also uncapped molecules containing errors, mostly deletions. These incorrect molecules will either block assembly of the oligonucleotides, or result in mutations in the assembled DNA. Since on average one of every two molecules is correct, the probability that a strand of our ϕX genome would completely assemble correctly by random selection from 130 unpurified oligonucleotides is $(1/2)^{130}$ or about 10^{-39}. We estimated that, even with selection for infectivity, it was essential to reduce the number of incorrect oligonucleotides to 10% or less to allow detection of correct molecules. Because purification of each individual oligonucleotide would be time consuming and laborious, and because all our synthetic oligonucleotides were of equal chain length, we chose to gel purify pooled oligonucleotides. We pooled the nucleotides from each strand in two separate pools to minimize the likelihood of annealed structures that could interfere with gel purification on the basis of chain length.

Ligation. We chose *Taq* ligation as the first step in our ϕX assembly because ligation under stringent annealing conditions (55°C) would diminish the possibility of incorrect pairing, and might also select against ligation of oligonucleotides containing mutations. In principle, it should have been possible to obtain full-length ϕX genomes by ligation. A major reason why this was not attained is that the concentrations of all the

oligonucleotides in the ligation mixture are not equal. Growing assemblies terminate prematurely when particular oligonucleotides become exhausted. Fig. 2A shows the results of a computer simulation of the ligation reaction. Both the fraction of full-length product and the average chain length of assemblies drop rapidly as the percent of random variation in the oligonucleotide concentrations increases. For example, if the oligonucleotide concentrations vary by as little as 20%, essentially no full-length assemblies are made. Other contributing factors are incomplete oligonucleotide phosphorylation and sequence errors that interfere with efficient ligation. For these reasons we should not expect complete ϕX genomes to assemble in our ligation reactions and subsequent assembly of the ligation products by polymerase cycling is needed.

Polymerase cycle assembly (PCA). The PCA reaction is a thermocycling polymerase reaction, similar to a PCR reaction, but without a pair of primers present in excess compared with template (14). At each cycle, DNA is melted and overlapping single strands reanneal. If the 3'-ends of reannealed strands are such that they can be extended using the opposite strand as template, then polymerase extends the strands to form duplex molecules. DNA strands continue to elongate at each cycle until they are either full-length or can no longer be extended. It should be noted that PCA is not an amplification reaction. Only limited total synthesis can occur and, because the polymerase mixture we used contains no 5' exonuclease, no DNA is degraded. It is difficult to analyze the kinetics of the process; however the final end products are simple to describe (Fig. 2B). Molecules present in the starting mixture can be extended in the 5'>3' direction to the end of the genome just once, over a number of cycles. Only the molecules containing the 5' ends of each strand can be extended to full-length. The total increase in mass of the DNA is limited to ~ n/2, and the fraction of final mass that can achieve full-length is ~ 2/n, where n is the number of oligonucleotides on one strand (Fig. 2B). We assumed it would be unlikely that either the ligation or the PCA reactions are a major source of errors since no synthesis occurs in the former and it is limited in the latter (although some errors could occur by deamination or depurination during these steps). Most errors can be attributed to incorrect synthesis of the oligonucleotides. If our oligonucleotides are 90% pure, then each synthetic chain would, on average, contain about 13 mutations. The fraction of correct synthetic ϕX174 (synϕX) molecules would be $(0.9)^{130} \sim 10^{-6}$.

Results

The oligonucleotides were designed to synthesize a ϕX genome with exactly the sequence reported by Sanger and colleagues in 1978 (ref. 13, several database entries have exactly this sequence, NC_001422 [genome database], J02482, and V01128). In designing the oligonucleotide set, we adopted the strategy of appending sequence to either end of the ϕX sequence to make a slightly larger molecule that could be cleaved to size with PstI and then circularized to produce infectious DNA. We appended 5'-TAACGCTGCA to the left end and 1 G plus a randomly generated 73 nucleotide sequence to the right end thereby restoring a PstI site at each end. Starting at the left end, 130 oligonucleotides (top-1 to top-130) each 42-bases in length, were consecutively generated. Similarly, starting on the bottom strand at position 22, oligonucleotides

numbered bot-131 to bot-259 were generated (see Fig. 1, upper left). These oligonucleotides were pooled and gel purified as described (see Materials and Methods).

We phosphorylated the oligonucleotides and mixed the top and bottom pools. To improve the ligation fidelity we used Taq ligase at 55°C for 18 hours. A sample of the ligation reaction was analyzed on an InVitrogen 2% E-gel (Fig. 3E). The average size of the double-stranded products was around 700bp with a small fraction extending up to 2 or 3kb (Fig. 3E, lane N). Fig 3E, lane D shows a sample denatured in formamide prior to loading on the gel. The single-stranded products ranged in size with the largest being about 1kb.

To obtain full-length ϕX DNA molecules, we diluted 0.2µl, 0.5µl and 1µl samples of the ligation product to to 50 µl and subjected them to 35 cycles of polymerase cycle assemble (PCA) as shown in Fig. 3*A*. With the 0.2 and 0.5µl samples, single-stranded assemblies approaching full-length or nearly full-length were produced, while assemblies from the 1µl sample were shorter. PCA of ligation samples exceeding 1µl yielded even shorter PCA products (data not shown). Apparently, dilute conditions favor the annealing of only two partially overlapping strands at a time followed by fill-in synthesis, whereas concentrated conditions favor multi-branched, annealed structures that interfere with fill-in synthesis.

To improve the yield of full-length ϕX, diluted samples of the first stage of PCA assembly were subjected to an additional 35 cycles of PCA. A small fraction of the products was now visible as full-length single-strands (Fig.3*B*) and as full-length duplexes (Fig. 3*C*). The presence of products of size apparently greater than full-length suggests either that some incorrect assemblies had accumulated during the later PCA cycles or that branched structures might have formed by reannealing either during electrophoresis (Fig. 3*B*) or during the final cycle of PCA (Fig. 3*C*).

After viewing the intensity of the bands, and realizing that only 1/25 of the PCA reaction mixture was analyzed on the gel, we estimated that several nanograms of full-length product were produced. Since 1ng of ϕXRF is approximately 1.7×10^8 molecules, we must have generated more than 10^9 independent synϕX molecules, and we can infer that some of these would have the correct sequence.

A small sample (0.5µl) of each second stage PCA reaction was amplified by PCR using top-1 and bot-259 oligonucleotides as primers (Fig. 3D). The PCR reactions were pooled, cleaved with PstI, and the linear DNA was gel-purified to yield approximately 1µg of linear synϕX. The linear synϕX molecules were circularized by ligation under dilute conditions (<1µg/ml) in preparation for infectivity testing. Gel analysis showed that about 50% of the linear synϕX were converted to circular molecules (data not shown).

One µl of synϕX ligation product (an estimated 3 to 5ng of circular molecules based on ethidium bromide staining intensity) was electroporated into DH10B cells (InVitrogen), divided into two aliquots, A and B, and plated for infectious particles as described in

Materials and Methods. The synϕX-A lysate yielded 194 plaques (Fig. 4) and the synϕX-B lysate yielded 181 plaques per 100µl plated. Some variation in plaque size was observed. The phage yield from 3 to 5 ng of the synϕX product was 2 x 10^{-4} of the yield from 1 ng of commercial NEB RF DNA. We therefore estimate that the synϕX product is about 5 x 10^{-5} of that for natural ϕX, leading us to conclude that there are 9 to 10 inactivating mutations per synthetic genome. This is in reasonable agreement with our estimate of 90% purity for the gel-purified oligonucleotides.

Several plaques were picked from each plate directly into PCR reaction mixtures for amplification and sequencing. The A plaques are independent of the B plaques, but individual plaques from A, or from B, are not necessarily independent since they may have arisen from the same infected cell. Representative sequences from four plaques were compared in Fig. 5. Phage DNA from plaque B3 gave a sequence that was identical to GenBank J02482. The B1 DNA sequence contained one silent T>C transition, two silent G>A transitions, one G>A transition at position 4170 that resulted in a gly to ser amino acid change in the gene A protein and one T>G transversion at position 3606 resulting in a ser to ala change in gene H protein. DNA from plaque A4 contained a T>C silent mutation at position 1045 and A8 contained a G>A silent mutation at position 446, a C>T mutation at position 4399 that resulted in an ala>val amino acid change in gene A protein, and a C>A change at position 5144 resulting in a gln to lys change in gene B and silent changes in gene A and A*. Infectious titers of synthetic phage from plaque B3 were indistinguishable from commercially available phage.

Discussion.

Synthesis of a molecule with the same properties as a naturally occurring compound has traditionally been used as evidence for correctness of a proposed molecular structure. It is, therefore, interesting to consider whether synthesis of an infectious ϕX genome proves the correctness of the Sanger sequence (13) upon which our synthesis was based. Clearly, the sequence is accurate enough to specify an infectious phage. One synϕX isolate (B3) with wild-type plaque morphology is identical in sequence to the Sanger sequence. In a study of convergent evolution using ϕX, Bull et al., (15) determined that the wild-type ϕX used in their experiments differed at five positions from the Sanger sequence (13). In the present work we have also sequenced four different preparations of natural ϕX. It is interesting that none of these sequences is identical to the Sanger sequence (Fig. 6), differing from it by three to five single base substitutions in addition to the single base change associated with the *am*3 mutation, which is present in our natural DNA preparations, but absent from the Sanger sequence as presented in GenBank (13). SynϕX-A4, A8, and B3 are consequently closer to the Sanger sequence (13) than any of the natural ϕX DNA preparations (10) we have sequenced. It should be noted that the commercial ϕX DNA present in the lab while synthesis was ongoing differs at five positions (including the am3 site) from the Sanger sequence. None of those differences are found in our synϕX sequences, demonstrating that this isolate is synthetic rather than a contaminant. The natural sequences closest to the Sanger sequence are from DNA preparations contemporaneous with the original sequencing (Fig 6). Our sequences were determined using template DNA prepared in the late 1970s from ϕX brought back from a sabbatical in the Sanger laboratory by one of the authors (C.A.H.). It is difficult to

conclude whether any of the three differences between the Sanger sequence and contemporaneous ϕX preparations represent sequencing errors. Only two of these differences are shared by all four of the natural ϕX sequences and the sequence of Bull et al. (15), and these are both silent changes. SynϕX-B3 is identical to the Sanger sequence (13) at both these positions and is fully infectious. The quality of a synthetic genome sequence depends in part on the accuracy of the original DNA sequence from which it is derived. It is truly remarkable how well the first sequence of a DNA genome holds up to close scrutiny a quarter of a century later.

We have considered the source of mutations limiting the accuracy of assembled synthetic DNA. The predominant impurities in oligonucleotide preparations are molecules shorter than the desired product, which are expected to result in deletions following assembly. In our experiments, selection for viable ϕX strongly selects against such frameshift mutations, and none were seen in the viable synϕX isolates. The observed mutations are all single base substitutions, predominantly transitions (changes from one pyrimidine to the other [C<->T], or from one purine to the other [A<->G]). The two obvious sources of these substitutions are enzymatic replication errors and pre-existing base substitutions in the synthetic oligonucleotide preparations. The mutation frequency expected from enzymatic errors is simply due to the 25 cycles of PCR carried out prior to circularization of the genome. PCA reactions only extend molecules present in the original ligation mixture to full length a single time, and so are not expected to contribute significantly to the mutation frequency. The error rate for PCR using *Taq* polymerase has been measured to be approximately 10^{-5} (mutation frequency /bp/ duplication) and the HF polymerase mixture used in these experiments is reported to have significantly higher fidelity. This leads us to estimate an average of approximately one PCR induced mutation per genome. The observed rate appears to be somewhat higher, leading us to speculate concerning the possible origin of base substitutions in the oligonucleotide preparations. The predominant transition type observed (G.C to A.T which results in G>A and C>T changes) could arise by deamination of C to produce U. PCR amplification would then copy U in DNA as T to produce normal DNA that is insensitive to the uracil N-glycosylase of *E. coli*. The phosporamidite of U might exist as a contaminant of the C phosphoramidite used in oligonucleotide synthesis, or deamination might occur at any stage prior to PCR amplification of the full length ϕX genome.

We have demonstrated the rapid, accurate synthesis of a large DNA molecule based only upon its published genetic code. The accuracy of our final product was demonstrated by DNA sequencing and by phage infectivity. Our methods will permit serial synthesis of gene-cassettes containing 4-7 genes in a highly robust manner. However, without selectivity, these cassettes will contain mutations (~2/kb) derived from errors contained in the oligonucleotide pool. Thus, without error correction, they would currently be unsuitable for assembly into a chromosome for an entire organism. Selection for infectivity, such as we have used, or for open reading frames in single genes provide advantages for synthesizing viruses or short sequences. However, when our method of synthesis is coupled with DNA sequencing and repair by site directed mutagenesis, it will enable rapid production of accurate cassettes that can be assembled into larger genomes. The capabilities of DNA synthesis have lagged far behind our ability to determine

sequences during the past 30 years. If this gap can be closed, then limitless possibilities for the application of synthetic methods to the study and practical application of genomics will emerge. There are many reasons to synthesize DNA chemically, rather than clone natural sequences, one of which is to prove correctness (or incorrectness) of a sequence, as many published sequences contain errors. The natural DNA may be unavailable to the experimenter for various reasons including: an uncooperative lab; an environmental sample that has been used up; an archaeological sample in short supply, or a sequence from badly degraded DNA; or from an extinct organism. As well, the sequence could be deduced rather than experimental from an ancestral sequence, a designer protein, or a fusion of domains from different proteins. There may be a hazard associated with the source of the natural sequence or the target sequence may be RNA or a protein sequence rather than a DNA sequence. The sequence may need to be re-engineered in order to alter the codon usage (or the code) for a particular host; to alter closely spaced regulatory signals, or protein initiation (ribosome binding sites), promoters, transcription terminators; to introduce restriction sites; or to allow convenient construction of a family of related, but different constructs. Synthesis may become the easiest way to get a sequence, as methods are refined. A desired construct may require the assembly of many pieces from different sources. Synthesis obviates the need to develop a special strategy for each construct providing complete flexibility of design. The combination of improved oligonucleotide synthesis combined with the methods described here will enable rapid, accurate synthesis of genomes of self-replicating organisms that will serve as a basis of understanding minimal cellular life. Synthetic genomics will become commonplace and provide the potential for a vast array of new and complex chemistries altering our approaches to production of energy, pharmaceuticals, and textiles.

1. Wöhler, F. (1828) *Ann. Phys. Chem.* **88**, 253.
2. Sekiya, T., Takeya, T., Brown, E. L., Belagaje, R., Contreras, R., Fritz, H. J., Gait, M. J., Lees, R. G., Ryan, M. J., Khorana, H. G., *et al*. (1979) *J. Biol. Chem.* **254,** 5787-5801.
3. Letsinger, R.L., Mahadevan, V. (1965) J. Am. Chem. Soc. **87**, 3526-7.
4. Letsinger, R.L., Ogillvie, K.K., Miller, P.S. (1969). J. Am. Chem. Soc. 91, 3360-5.
5. Matteucci, M.D., Caruthers, M.H. (1981) J. Am. Chem. Soc **103**, 3185-91.
6. McBride, L.J., Caruthers, M.H. (1983) Tetrahedron Lett. 24, 245-8.
7. Hutchison, C. A. 3rd, Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., Venter, J. C. (1999) *Science* **286,** 2165.
8. Cho, M. K., Magnus, D., Caplan, A. L., McGee, D., The Ethics of Genomics Group. (1999) *Science* **286,** 2087.
9. Cello, J., Paul, A. V., Wimmer, E. (2002) *Science* **297,** 1016-1018.
10. Sinsheimer, R. L. (1959) *J. Mol. Biol.* **1,** 43.
11. Goulian, M., Kornberg, A., Sinsheimer, R. L. (1967) *Proc. Natl. Acad. Sci. USA* **58,** 2321-2328.
12. Kornberg, A. (2000) *J. of Bacteriology* **182,** 3613-3618.
13. Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A. 3rd, Slocombe, P. M., Smith, M. (1978) *J. Mol. Biol.* **125,** 225-246.
14. Stemmer, W. P. C., Crameri, A., Ha, K. D., Brennan, T. M., Heyneker, H. L. (1995) *Gene* **164,** 49-53.
15. Bull, J. J., Badgett, M. R., Wichman, H.A., Huelsenbeck, Hillis, D. M., Gulati, A., Ho, C., Molineux, J. (1997) *Genetics* **147,** 1497-1507.

This work was supported by a U.S Department of Energy Grant and by a grant from the J. Craig Venter Science Foundation.

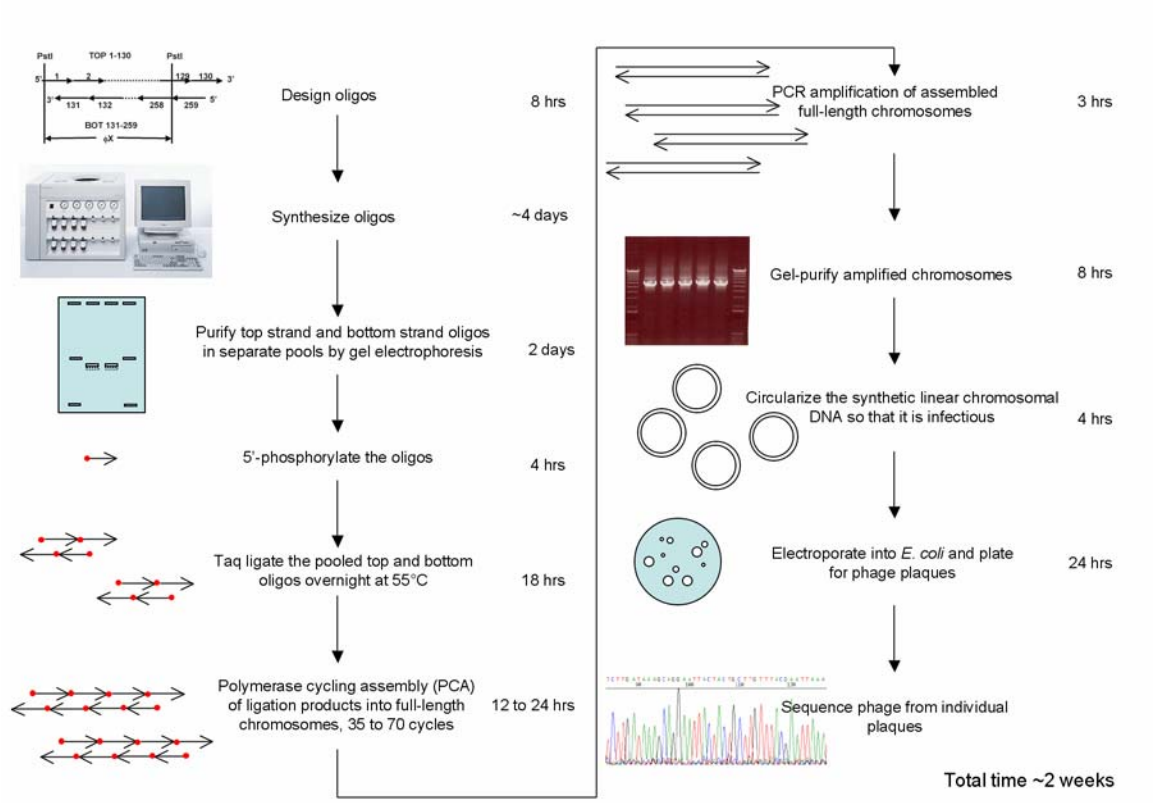


Figure 1. Schematic diagram of the steps in the global synthesis of infectious ϕX174 bacteriophage from synthetic oligonucleotides.

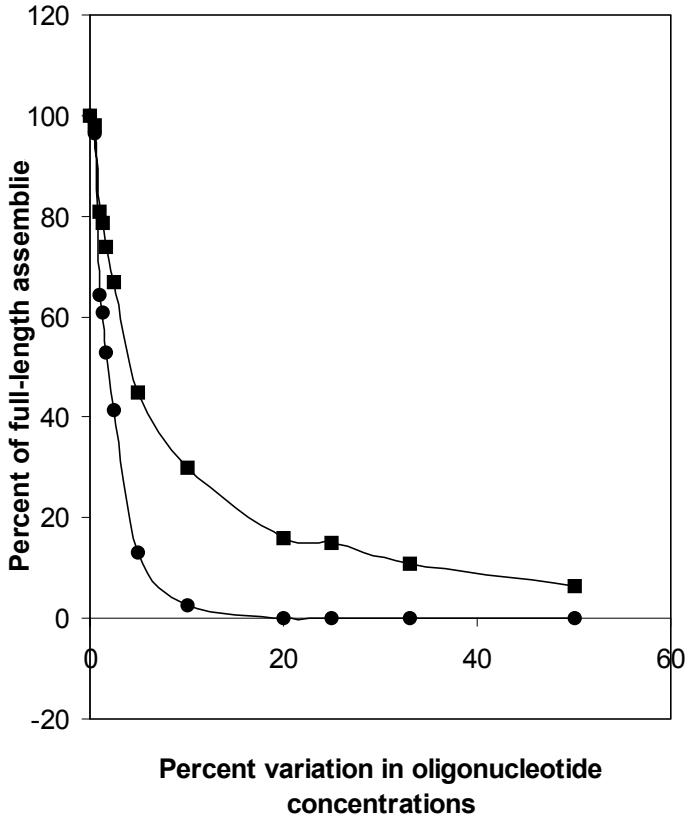
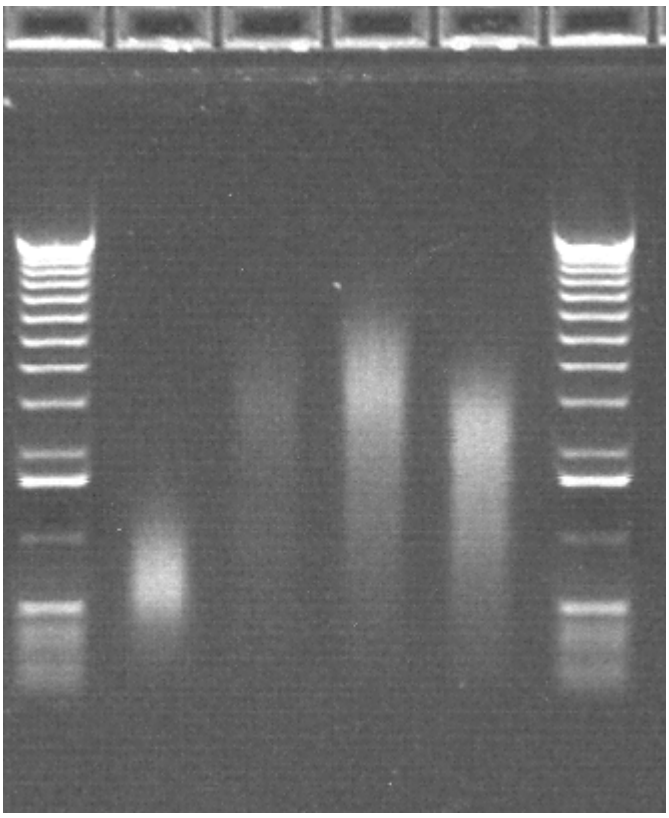


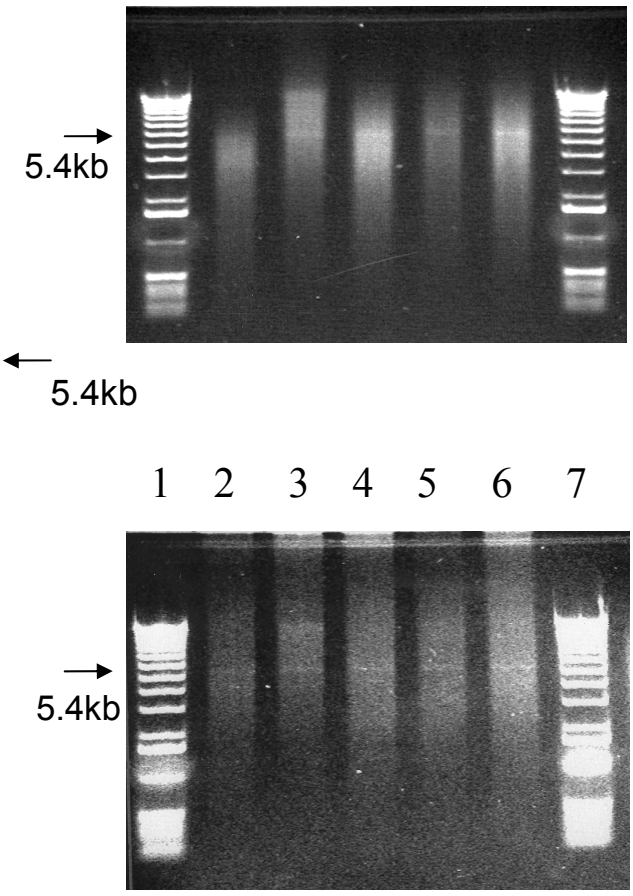
Figure 2. Panel A. Computer simulation of the ligation reaction as a function of variation in the concentrations of the input oligonucleotides. Computer simulation parameters are130 top oligonucleotide (with bottom oligonucleotides saturating) and 2000 molecules of each top oligonucleotide. Percent variation of oligonucleotide concentrations is determined by the randomly chosen number of molecules of each oligonucleotide. For example, at 10% variation in oligonucleotide concentrations, the number of each oligonucleotide is randomly chosen between 1800 and 2000. At each iteration of the program a random pair of assemblies is selected and the pairs are joined together to form a larger assembly if the end coordinate of one assembly is one less than the beginning of the other assembly. The process is iterated until the number of assemblies no longer changes during one million pairings. -■-■-, average percent length of assemblies; -♦-♦- percent of full-length assemblies. Panel B: The final products of a theoretical PCA reaction that starts with oligonucleotides of uniform size. Only the two terminal oligonucleotides can be extended to full-length since polymerization is only in the 5' to 3' direction. An assembly is considered to be "active" if it can be extended by overlapping with another assembly followed by fill-in synthesis. An assembly is "inactive" if it cannot be extended further. Mass increase factor = final mass / beginning mass = sn (n+2) / 2sn = (n + 2) / 2 ~ n/2. Fraction of final mass full length = [2 (sn + s/2)] / [sn (n + 2)] = [2 (n +1/2)] / [n (n + 2)]~ 2/n; s = nucleotide length of oligonucleotide; n = number of oligonucleotides on one strand.

PANEL 3

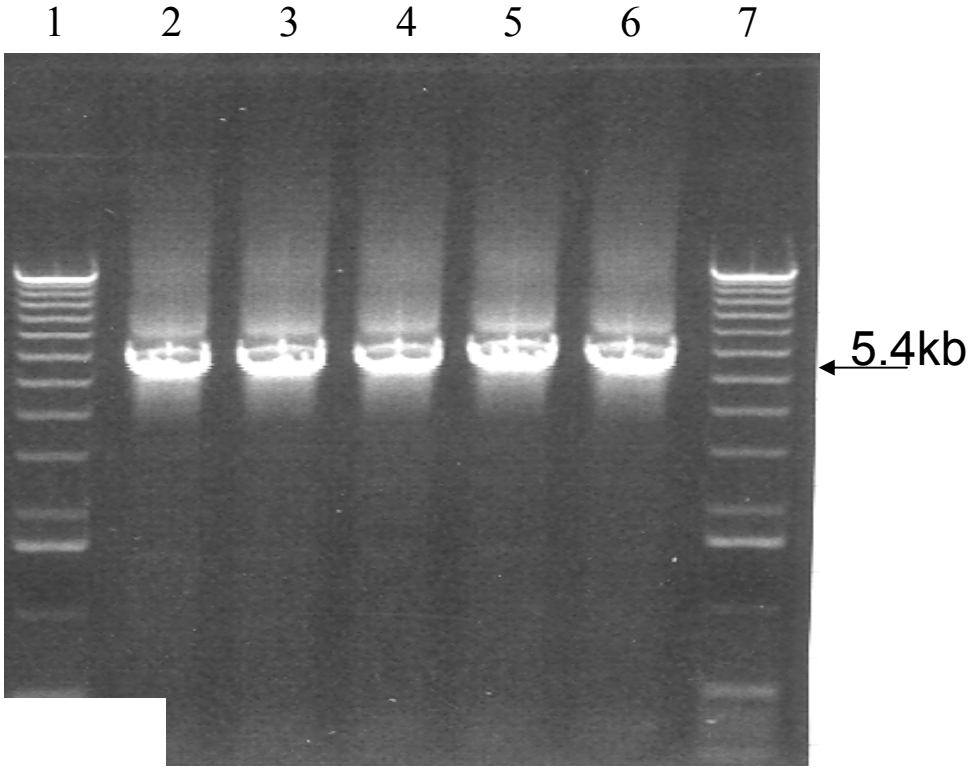
A



B



PANEL 3D



PANEL 3E

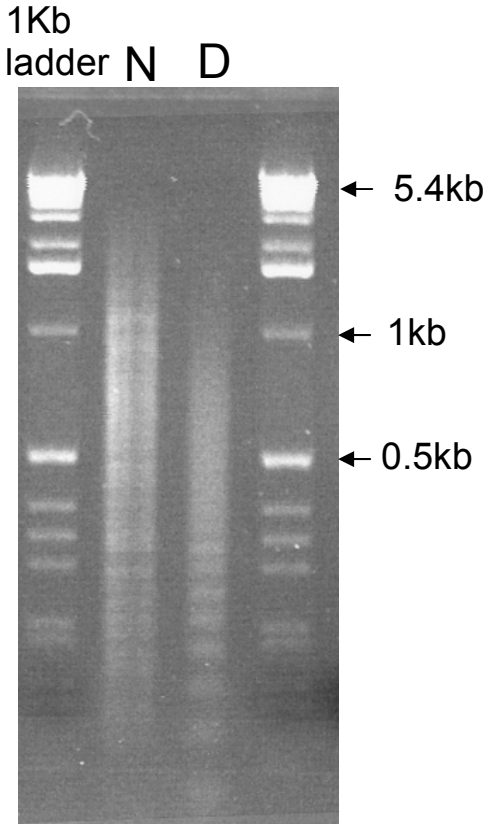


Figure 3: Polymerase cycle assembly of full-length synϕX molecules. The first stage of PCA (50µl reaction volume) was carried out for 35 cycles with either 0.2, 0.5, or 1µl of the Taq ligation product. PCA products were analyzed on a 0.8% E-gel. Panel A: Lane 1 and 6, 1 kb ladder; lane 2, 0.5µl of Taq ligation product; lane 3, 2µl of the 0.2µl PCA: lane 4, 2µl of the 0.5µl PCA; lane 5, 2µl of the 1µl PCA. The second stage of PCA was for an additional 35 cycles in 5 new 50µl reactions. For reaction 1, the 0.2µl first stage reaction was continued without change for another 35 cycles with the addition of 0.5µl of fresh HF polymerase mixture. For reactions 2 and 3, 10µl and 20µl of the 0.5µl first stage PCA product was used. For reactions 4 and 5, 5µl and 10µl of the 1µl first stage PCA product was used. Analysis was on 0.8% E-gels. Panel B, native DNA. Panel C

formamide denatured DNA. Lane 1 and 7, 1kb ladder, lane 2, 2µl of reaction 1, lanes 3 and 4, 2µl of reactions 2 and 3, lanes 5 and 6, 2µl of reactions 4 and 5. Panel D, PCR amplification of the products of the second set of PCA products as shown in Panel E. Panel E: Taq ligase assembly of 259 oligonucleotides. A 0.5µl sample of the ligation products was analyzed on a 2% E-gel (InVitrogen) in duplex form in lane N. 1µl of the ligation products was mixed with 20µl of formamide, heated to 95°C for 2 min and then analyzed on lane D. Denatured standards run approximately the same as native standards, based on other experiments (data not shown).

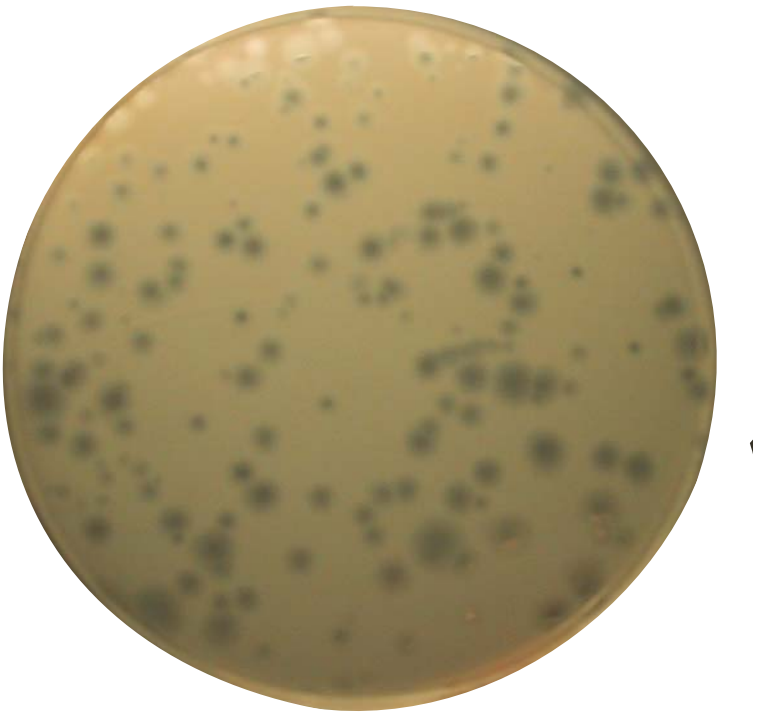


Figure 4. Plaques of synϕX-A. There appear to be several plaque morphologies: small plaques with sharp borders, medium sized plaques, and large plaques with fuzzy borders.

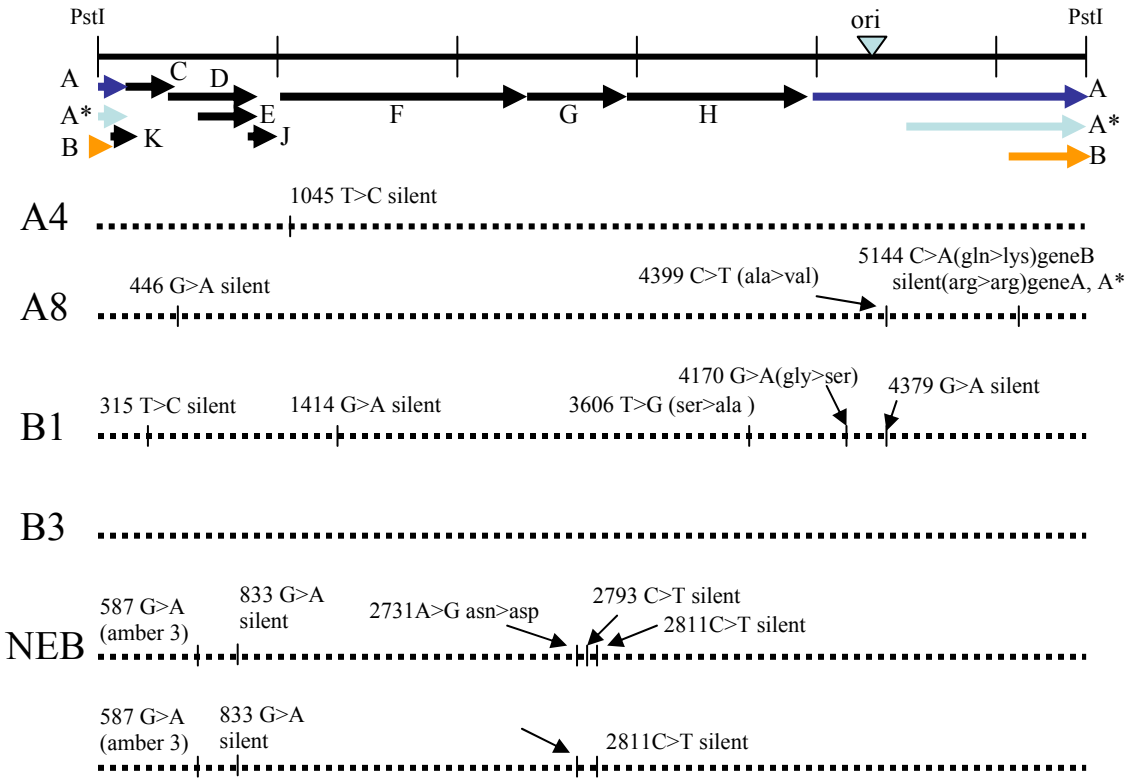


Figure 5. Sequence comparisons of natural and synthetic ϕX genomes. Differences from the Sanger sequence (13) are indicated. A4, A8, B1 and B3 are the synthetic ϕX described in the text. NEB is the ϕX 174 RF I DNA supplied by New England BioLabs, catalog no. #N3021S; RF70s is DNA prepared in the late 1970's and stored since then by C.A.H.

Position	*Genbank*	*RF70s*	*SS78*	*Bull*	*G'97*	*NEB'03*	*Gene*	*Genbank->Variant*
0587	G	A	A	G	A	A	D E	Val->Val (silent) Try->am (am3)
0833	G	A	A	A	A	A	D E	Ala->Ala (silent) Arg->Arg (silent)
1650	A	A	A	G	G	A	F	His->Arg
2731	A	G	G	A	A	G	G	Asn->Asp
2793	C	C	C	C	C	T	G	Pro->Pro (silent)
2811	C	T	T	T	T	T	G	Asn->Asn (silent)
3340	A	A	A	A	G	A	H	Asp->Gly
4518	G	G	(G)	A	A		A	Ala->Thr
4784	C	C	(C)	T	C	C	A	His->His (silent)

Figure 6. Sequence differences between the Sanger sequence and more recent sequencing of natural ϕX DNAs. RF70s: A preparation of ϕX double-stranded RF from the late 1970's. SS78: A preparations of ϕX virion single-stranded DNA from 1978. Bull: The sequence of wild-type ϕX used by Bull et al. (15); G'97: ϕX RF DNA from 1997. NEB'03: ϕX RF DNA from New England Biolabs in use at IBEA during the ϕX genome synthesis.