Rebuilding microbial genomes

Robert A. Holt,* Rene Warren, Stephane Flibotte, Perseus I. Missirlis, and Duane E. Smailus

Summary

Engineered microbes are of great potential utility in biotechnology and basic research. In principle, a cell can be built from scratch by assembling small molecule sets with auto-catalytic properties. Alternatively, DNA can be isolated or directly synthesized and molded into a synthetic genome using existing genomic blueprints and molecular biology tools. Activating such a synthetic genome will yield a synthetic cell. Here we examine obstacles associated with this latter approach using a model system whereby a donor genome from *H. influenzae* is fragmented, and the pieces are then modified and reassembled stepwise in an *E. coli* host cell. There are obstacles associated with this strategy related to DNA transfer, DNA replication, cross-talk in gene regulation and compatibility of gene products between donor and host. Encouragingly, analysis of gene expression indicates widespread transcription of *H. influenzae* genes in *E. coli*, and analysis of gap locations in *H. influenzae* and other microbial genome assemblies reveals few genes routinely incompatible with *E. coli*. In conclusion, rebuilding and booting a genome remains a feasible and pragmatic approach to creating a synthetic microbial cell. *BioEssays* 29:580–590, 2007.

Introduction

Genetic modifications to microbes have been undertaken for decades, but now the possibility of building completely engineered or synthetic cells is being contemplated. There are two basic approaches: the ''top-down'' approach of isolation, manipulation, re-assembly and self-replication of genetic material from an existing ''natural'' cell or, alternatively, the ''bottom-up'' approach of assembling a minimal self-replicating system of carefully defined small molecules.[1,2]

Canada's Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada.
*Correspondence to: Robert A. Holt, Canada's Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada.
E-mail: rholt@bcgsc.ca
DOI 10.1002/bies.20585
Published online in Wiley InterScience (www.interscience.wiley.com).

Abbreviations: BAC, Bacterial Artificial Chromosome; ORF, Open Reading Frame; WGS, whole genome shotgun.

Engineering a cell by total synthesis using well-understood components will ultimately allow greater flexibility in design and a broader range of capabilities, some of which cannot even be imagined today. However, the complexity of microbial life is such that the bottom-up approach is probably a long way from delivering a product recognizable as a free living organism or a product with utility. For this reason, we have focused our efforts on a top-down approach, using existing blueprints to build a microbial genome, and then activating that genome to produce the organism that it defines. With this approach, we are confined to our current and incomplete understanding of microbial genomics, but the trial and error involved will provide opportunities for scientific discovery that may be missed by observation alone.[3] Further, reducing to practice the methods for taking apart and rebuilding complete genomes and cells will offer a fast track to harnessing the biocatalytic potential of the microbial world.

To build a microbial genome one needs genes. The technology for synthesis of large DNAs capable of encoding multiple genes is based on serialized oligonucleotide assembly and amplification.[4,5] The power of this approach, in terms of length of DNA synthesized per unit cost, appears to be advancing on a trajectory reminiscent of Moore's law,[6] and the prospect of cost-effective direct synthesis of complete microbial genomes on the order of a million base pairs in size is not unrealistic in the near future. The main problem presented by a top-down approach to making a synthetic cell is not obtaining the genetic material, but rather how to activate this naked DNA to obtain the organism that it encodes. For viruses, activation of genomes assembled from cDNAs[7] or synthetic oligonucleotides[4,8,9] has already been accomplished, but the procedure is greatly facilitated by the small size of viral genomes (thousands of base pairs) and by the fact that part of the viral replication cycle involves the easily mimicked step of insertion of DNA into a host cell. When considering the activation of the genome of a free-living organism, the problems become more profound. There is no remotely practical way to assemble cellular components around a fabricated genome so, borrowing from virology, the most-reasonable approach is to deliver the genome to a host cell and booting it by exploiting the host cells transcriptional, translational, replicative and metabolic activities. In principle, once the new donor genome is functioning, removal of the redundant host genome leaves a new organism. While this approach is

simple in concept, in practice there are significant technical challenges including limitations to physically manipulating and transforming fragile high molecular weight DNA, recognition of donor regulatory sequences by host factors, donor gene products that are lethal to the host cell and, ultimately, independence of the nascent genome from host cell factors and its self-propagation.

A model system for rebuilding a microbial genome

We have established a model system whereby genomic DNA is isolated from a donor organism (*Haemophilus influenzae* Rd20), fragmented and re-assembled piece by piece in a host organism (*Escherichia coli* K12) (Fig. 1). Supplemental data pertaining to this model system and relevant to this article are available at ftp://icebox.bcgsc.ca/pub/uploads/BIES-06-0189/. In these early experiments, the approach of utilizing a donor genome from an existing distinct species allows donor and host genetic material to be readily distinguished in hybrid cells and avoids having to debug a novel design. While other donor–host pairs could be envisioned,[1] *H. influenzae* and *E. coli* are a good starting place because both are commensal gamma-proteobacteria and there are well-characterized non-pathogenic laboratory strains with quality reference genome sequences[10,11] readily available. Further, *H. influenzae* has a relatively small genome (1.83 Mbp) which minimizes the number of manipulations required. 1286 of the 1788 *H. influenzae* genes (70.1%) have a 1:1 ortholog (reciprocal best Blast match) in *E. coli* and these orthologs are typically present on short conserved strings containing two to four genes and show an average amino acid sequence identity of 59%.[12] There is little long-range synteny between these two species and the more compact genome of *H. influenzae* is explained by fewer paralogous gene expansions than are present in *E. coli*.[11]

Our proposed genome rebuild procedure involves the following steps: (1) isolation and shearing of *H. influenzae* genomic DNA, (2) construction of a Bacterial Artificial Chromosome (BAC) library and propagation of individual clones in *E. coli* (3) mapping clone end sequences against the reference *H. influenzae* genome and selecting a minimal tiling set, (4) manipulation of each clone to remove vector and add sequences to mediate recombination, (5) performing serial recombination of modified donor segments in the host cell to produce a intermediate hybrid organism and, finally, (6) removal of the host genome. This system is based on numerous assumptions and unknowns, some of which have already been addressed and others that cannot be tested until later stages in the protocol. Here we describe progress in this model system and address these considerations.

A large insert H.influenzae genomic library

The *Haemophilus* genome project used lambda clones for long-range linking[9] so it was not initially clear that large (~100 kbp) contiguous DNA segments could be propagated in *E. coli*. To address this question and to provide clones for subsequent manipulation, we commissioned the construction of an *H. influenzae* BAC library (Amplicon Express, Pullman, WA). Partially digested *H. influenzae* DNA was cloned into the single copy pECBAC1 vector containing the F replicon.[2] We also constructed a 10 kbp plasmid library in a pBR322-based vector. Inserts from both libraries were sequenced to high redundancy and mapped to the reference genome using paired end sequences (Fig. 2A). From mapped clones, a minimal tiling containing nineteen BACs was selected using criteria to maximize genome coverage, minimize clone number and overlap, and break the fewest possible number of genes and predicted operons.

Genome assembly

To reassemble the *H. influenzae* genome, neighboring clones in the tiling path need to be fused in a stepwise manner in the host cell. This raises many interesting issues. First, why bother with these multiple steps? Why not just isolate an intact *H. influenzae* chromosome and use this in its entirety to transform the host cell? Notwithstanding the fact that high molecular weight DNA is fragile and difficult to handle, the *H. influenzae* chromosome is well beyond 200 to 300 kbp practical size limit for delivering DNA into a cell by chemical transformation or electroporation. Likewise, cells that are naturally competent take up small DNA molecules rather than complete genomes. Finally, even if a complete chromosome could be transferred, perhaps by conjugation or vesicle fusion, the recipient hybrid cells would not be viable due to the payload of incompatible, lethal genes and the opportunities to synthesize or otherwise supplement sections of the genome with novel coding sequences would be lost. Thus, it seems that

[1]Much attention has been paid to mollicutes, such as *Mycoplasma genitalium* (Hutchison et al. 1999. Science 286:2165–2169) that are attractive due to their small genome sizes (<600 kb). However, as an obligate intracellular pathogen, *M. genitalium* is difficult to culture in the laboratory. Further, the UGA codon in *M. genitalium* has been reassigned from a stop codon to tryptophan. Therefore, *M. genitalium* genes expressed in an *E. coli* host would frequently be truncated.

[2]The enzyme *Hin*dIII was used for the first, unsuccessful attempt at partial digestion. The reason for failure was of course because we had overlooked the fact that this commercial enzyme is purified from *H. influenzae* Rd, so that all sites in our DNA sample were protected. Better results were obtained with *Mbo*I.

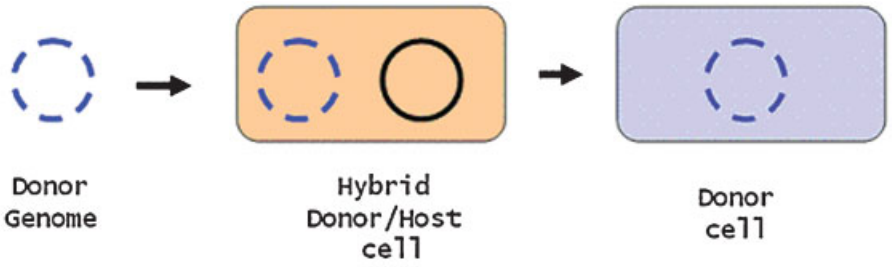

Figure 1. Model system for whole genome replacement strategy, whereby donor genomic DNA is isolated from a donor organism (*Haemophilus influenzae*), fragmented and assembled piece-by-piece in a host organism (*Escherichia coli*). Removal of the host chromsome leaves a new cell under control of the assembled donor genome.

the most practical approach to building a genome, be it natural, modified, or fully synthetic, appears to be stepwise assembly of moderately large segments in a host cell.

Remarkable progress has already been made in rebuilding a microbial genome in a host organism. Itaya and co-workers[13] used an iterative process of homologous recombination to integrate most (3.5 Mbp) of the genome of the photosynthetic bacterium *Synechocystis* directly into the genome of a *Bacillus subtilis* host. Their process, termed "inchworm elongation", involves insertion of a target sequence into the *Bacillus* genome followed by delivery of a *Synechocystis* genome segment tens of kilobases in length that recombines at that site. This is done iteratively, in an "inchworm" fashion in order to establish longer donor segments within the host chromosome. The purpose of this exercise by the Itaya group was to establish a method for assembling Mbp-sized segments of DNA, exploiting the *B. subtilus* genome as a cloning vector, rather than to assemble an independent genome for the purpose of top-down construction of a synthetic organism. However, it is conceivable that this approach could be modified for the latter purpose using an episomal target for recombination and a more closely related donor and host pair.

Our model system relies on assembling a non-redundant set of *H. influenzae* BAC clones in an *E. coli* host. A significant hurdle in this approach is presented by the BAC vector itself. BAC vectors use an origin of replication and partitioning system derived from the *E. coli* F plasmid.[14] The F plasmid, like other episomes, encodes only a few of the proteins required for replication and is dependent on the host for other factors. For example, the F replicon encodes RepE, which initiates replication by binding 19 bp repeats in the oriS region, but F also depends on host primosome genes *dna*B, *dna*C and *dna*G, plus the host initiation factors (*dna*A) and polIII (*dna*E).[15] The F-encoded *par*A and *par*B genes maintain strict copy number control at one molecule per cell. This means

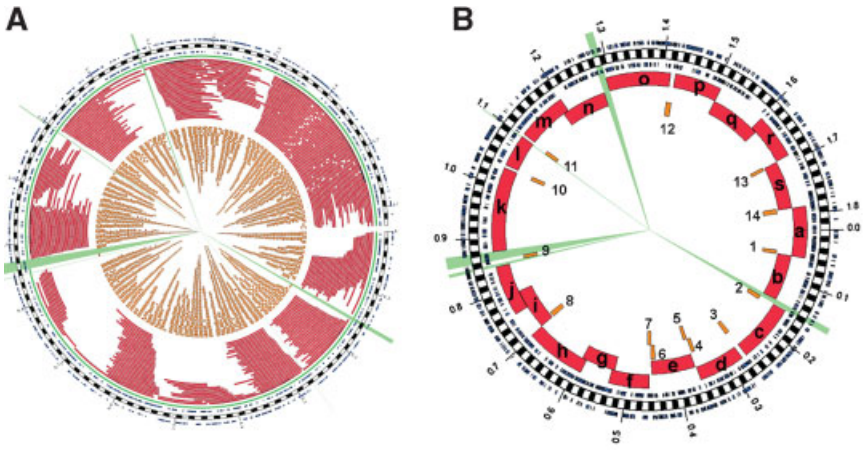

Figure 2. A: BAC clones (red) and 10kb plasmid clones (orange) end-sequenced and mapped to the *H. influenzae* reference genome sequence. Also shown are gene structures (blue), scale (black and white bars, 10kb each) and physical gaps (green rays). **B**: The minimal tiling set. Genome coordinates and end sequences for all clones are available at ftp://icebox.bcgsc.ca/pub/uploads/BIES-06-0189/.

that multiple independent BAC clones cannot be propagated in the same cell and, therefore, for genome reconstruction, vector segments must be removed from all but the first BAC being assembled. Since the pECBAC vector that our *H. influenzae* genome is cloned in has *Not*I restriction sites flanking the insert, and because there is only a single *Not*I site in the *H. influenzae* genome, we have been able to isolate from BAC clones vector-free genomic segments by *Not*I digestion and pulsed field gel electrophoresis purification (Fig. 3). Isolated linear segments can been joined to a given clone already propagating in *E. coli* by several approaches. For example, BACs can be retrofitted with appropriately placed homology segments prior to linearization such they can be transformed and fused with a resident BAC by lambda Red recombination.[16] This system utilizes a host strain carrying a segment of the phage lambda genome that contains the *exo, bet* and *gam* genes under the control of a temperature-sensitive repressor. These lambda genes mediate recombination between the ends of a linear incoming DNA segment with homologous sequences in a target DNA. The homology regions can be very short (~50 bp) and the target can be any chromosomal or episomal DNA molecule present in the host cell. Site-specific Cre/loxP[17] or FLP-FRT[18] recombination is an alternative means of assembly. If BACs are first retrofitted with non-promiscuous loxP sequences[19] then conceivably *Not*I cut linear segments can be re-circularized by ligation, transformed, and fused one after another with a resident BAC in a host cell expressing the Cre enzyme. In all cases, it would be sensible to avoid RecA-mediated homologous recombination for genome assembly because this approach could permit undesirable recombination events at off-target regions of homology. Finally, it is important to note that any in vitro approaches to genome assembly such as traditional restriction/ligation will be unsuitable when constructs exceed the size of DNA that can be effectively delivered to the host.

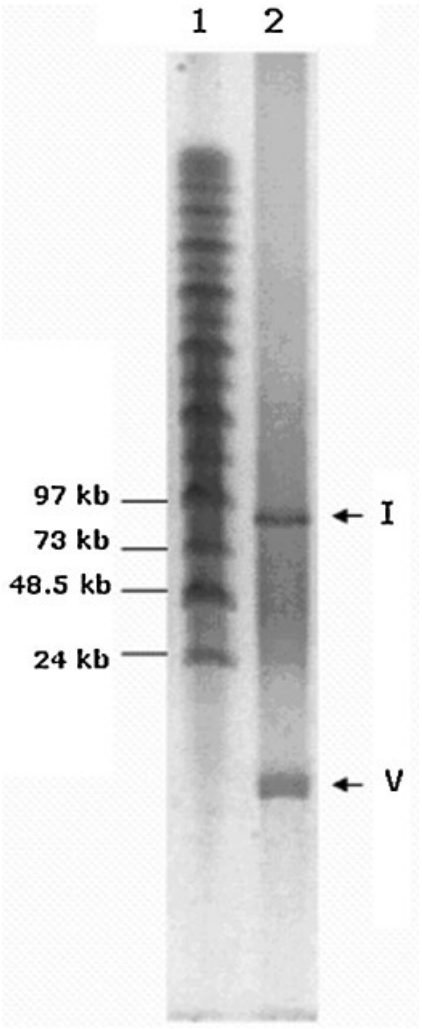

Figure 3. Isolation of BAC insert by restriction digestion and PFGE. A 1% agarose gel was run 20 hours, 6 V/cm, at 120° and 11°C then stained 30 minutes in SYBR® Green I nucleic acid stain. **Lane 1:** New England Biolabs Midrange II PFG Markers. **Lane 2:** *Not*I restriction digested *H. influenzae* BAC clone showing linearized 88 kbp insert (I) spanning genome coordinates 1,420,954 to 1,509,395, and 7.5 kbp linearized vecor segment vector (V).

Heterologous gene expression

For an assembled donor genome to eventually assume control of a host cell, there must be expression of the genes carried by the donor chromosome, and this expression must be initiated by the host transcriptional machinery. Thus, there must be some degree of evolutionary conservation between the donor and host organisms for this cross expression to occur. One of the more-startling findings from the *Synechocystis/B. subtilis* megacloning experiments by Itaya and co-workers[13] was the small number of transcripts originating from the donor *Synechocystis* DNA in the hybrid genome. Due to the poor regulatory conservation between these two species the *Synechocystis* genome was propagated essentially as a payload of inert DNA. This is not surprising given the substantial evolutionary distance separating these organisms. Although cross-expression was not a goal of these experiments, this observation underscores the fact that, in order to function, a transplanted genome needs a compatible host.

To explore cross-expression between *H. influenzae* and *E. coli*, we designed a Nimblegen microarray with 60 bp oligonucleotide probes that covered the majority of open reading frames (ORFs) for each species. Several types of filters were applied in the oligonucleotide selection process in order to maximize the sensitivity, specificity and signal-to-noise ratio. For example, constraints were imposed on homology with non-specific targets, on the GC content, on the

presence of homopolymers, on self-folding energy and on overlap with neighboring oligonucleotides. The resulting expression array can simultaneously measure 4222 of the 4447 genes (95%) in *E. coli* and 1705 of the 1813 genes (94%) in *H. influenzae*. Each *E. coli* gene is covered with 40 ± 5 oligonucleotides, each *H. influenzae* gene with 39 ± 7 oligonucleotides. In pilot experiments, we fused two randomly selected segments of the *H. influenzae* genome (a 28 kbp segment starting at genome coordinate 1,409 bp and an 18 kbp segment starting at genome coordinate 1,555,863 bp) and propagated this construct in BAC vector pECBAC1 in *E. coli* DH10B cells. RNA was isolated from this clone and also from native *H. influenzae* grown under the same culture conditions. Surprisingly, we obtained very clear expression signal from all genes present in the fused *H. influenzae* genome segments propagated in *E. coli* and the signal intensities were consistent with expression of these same genes in native *H. influenzae* (Fig. 4). We are proceeding with experiments to map cross-expression across the entire genome of *H. influenzae*. From this initial result, however, it is very encouraging that a reassembled *H. influenzae* chromosome will likely achieve widespread gene expression in *E. coli*, as a step towards complete autonomy. The contrast between the extensive cross expression observed between *H. influenzae* and E. coli and the near non-existence of cross-expression between the more distantly related bacteria *Synechocystis* and *B. subtilis* suggests that profiling of cross expression will be a useful and perhaps necessary approach for defining suitable donor/host pairs for rebuilding microbial genomes.

Lethal genes

A central issue in genome transplantation is the non-compatibility of certain gene products between donor and host. The existence of "toxic genes" is well established from microbial sequencing projects where physical gaps in the sequence arise wherever clones in the genomic library cannot be propagated because they encode protein products lethal to the host strain. We end-sequenced our *H. influenzae* BAC library and a 10 kbp plasmid library and mapped paired reads to the reference *H. influenzae* genome sequence. This exercise revealed five gaps covering 24,034 bp and containing 14 annotated genes (Table 1).

Initially, it came as a surprise that, in our experiments, the native *Haemophilus* restriction enzymes appeared to be easily clonable in *E. coli*. Bacteria have evolved restriction systems that protect the cell by digesting invading DNA at short

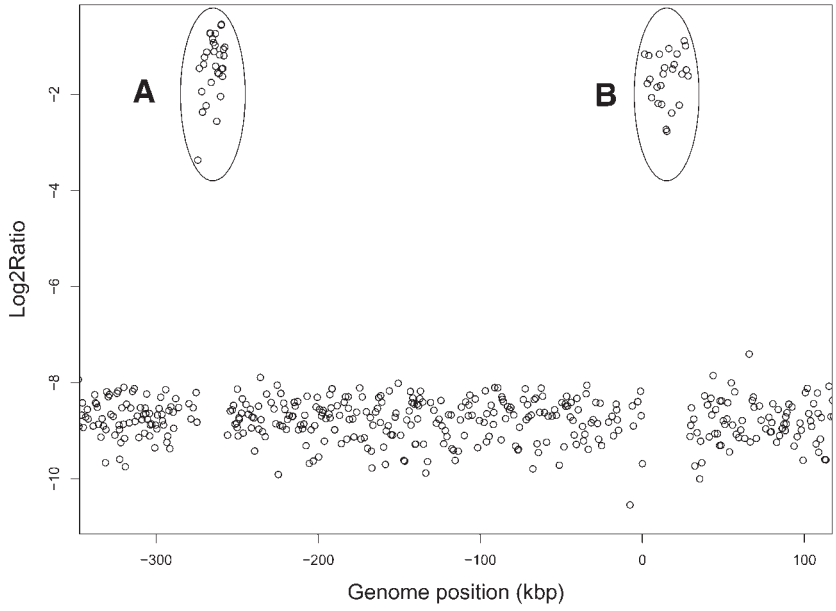

Figure 4. Expression of genes present on two segments of the H. influenze genome (**A,B**) fused in a single BAC clone and propagated in *E. coli*. Each point represents a single gene and is the mean multiple independent probes. The X-axis gives *H. influenzae* genome position and the Y-axis gives expression values of *H. influenzae* gene probes as the log2 ratio between intensity measured in the fused *H. influenzae* segments propagating as a BAC in *E. coli* (test sample) and the intensity measured in RNA from a culture of untransformed *H. influenzae* Rd (reference sample). A log2 value of 0 reflects equal expression, and a strongly negative log2 value indicates no expression. All genes present on the fused *H. influenzae* segments propogating in *E. coli* are expressed. Log2 values of 0 are likely not attained due to the fact that equal amounts of total RNA were used for the test and reference experiments, and ectopic *H. influenzae* transcripts constitute less of the total RNA in *E. coli*. Probe sequences and expression levels for each gene are available at ftp://icebox.bcgsc.ca/pub/uploads/BIES-06-0189/.

Table 1. *H. influenzae* genes encoding products that are toxic to *E. coli*

Gap coordinates	Gap size (bp)	Predicted gene product(s)
149,741–154,917	5,176	GTP-binding protein, DNA polymerase III ε, ribonuclease H
850,667–851,075	408	Ribosomal proteins S5, S8, S11, S13, L6, L15, L18, L30, L36
858,693–873,416	14,723	Argininosuccinate lyase, gal. 1-phosphate uridyltransferase, alanyl-tRNA synthetase, aminopeptidase P, galactokinase, galactose operon repressor
1,097,159–1,098,085	926	Putative transcriptional regulator
1,279,425–1,282,226	2,801	ssDNA-specific exonuclease, disulfide isomerase

recognition sequences that are protected in the host cell by methylation. Purified restriction enzymes are an indispensable tool in molecular biology and most laboratory strains of *E. coli* have their restriction systems knocked out. This is not the case for our donor *H. influenzae* Rd genome, and will probably not be true when engineering other genomes. Thus, one might expect that donor restriction enzymes would be toxic to host cells and would be unclonable. We hypothesized that we do not observe this in our system due to the fact that restriction endonuclease genes are co-transcribed with the methylase gene that protects the cell from self-digestion. This was verified by the fact that genomic DNA extracted from *E. coli* carrying the *H. influenzae* clone bearing the *Hin*dIII operon was completely protected from *Hin*dIII digestion in vitro.

When discussing the list of *H. influenzae* genes unclonable in *E. coli*, we are routinely asked the question "is this set of lethal genes unique, or are these genes that commonly show toxicity when a bacterial genomic library is cloned in *E. coli*"? Even though hundreds of prokaryotic genomes have been sequenced, this is question is not easy to answer. If a genome project includes gap closure, a finished sequence is published and the location of the gaps and their gene content is typically not reported. Alternatively, if a genome project reports a draft sequence containing gaps, then the identity of genes present in the gaps remains, by definition, unknown.

To address this issue, we undertook a meta-analysis where raw data in the form of paired end-sequences (and excluding finishing reads) were gathered from published whole genome shotgun (WGS) data sets from a number of bacteria and the paired reads were mapped against the finished reference genome sequence for each respective organism to locate gaps and identify the genes present within these regions (Table 2). Genes with similar predicted function missing in both our *H. influenzae* BAC libraries and the libraries of at least one WGS project include those coding for RNase H, ribosomal proteins, disulfide isomerases, transcriptional regulators and kinases.

When expressed in *E. coli*, viral RNase H genes have been found to be either toxic[20] or have reduced expression.[21] Related to the *H. influenzae* ribosomal protein-coding genes (S8, L6, L18, S5, L30, L15, L36, S11, S13) missing in our libraries, it is striking that clones containing genes predicted to code for a variety of 30S/50S ribosomal proteins are frequently under-represented in WGS libraries. Five out of the eight WGS projects analyzed here have clones missing for either 50S ribosomal protein-coding genes L1, L2 or L3 and/or 30S ribosomal protein-coding genes S1 or S20. Interestingly, it has been demonstrated that overexpression of endogenous ribosomal proteins in *E. coli* often shows feedback inhibition at the translational level to prevent accumulation of unassociated ribosomal proteins.[22] Further, some but not all *Bacillus stearothermophilus* ribosomal proteins have been shown to be toxic to *E. coli*.[23] In this latter case, it was postulated that the toxicity might arise because (1) ribosomal proteins show sufficient similarity to their *E. coli* counterparts to participate in ribosome assembly, but alter the function of the ribosome complex or (2) the cell simply does not tolerate large excess of ribosomal proteins.

Consistent with *H. influenzae*, four genes predicted to code for sulfurtransferases, disulphide bond formation proteins and sulfite reductase also appear to be missing from the WGS libraries of three different genome-sequencing projects. The protein-folding catalyst disulfide isomerase likely resolves incorrect disulphide bonds.[24] Interference of endogenous disulfide isomerases by expression of recombinants could yield incorrect protein folding and/or increased sensitivity to redox-active metals, preventing *E. coli* from surviving oxidative stress.[24]

Genes missing multiple copies at distinct sites of *Streptococci* genomes that we sampled include those coding for ribonucleases, ABC transporters and other permeases. Nine genes predicted to code for cell-wall-associated hydrolases were missing from both *P. gingivalis* and the WGS genomic libraries of three proteobacteria genomes. Three distinct transposase-coding genes were also missing from the *P. gingivalis* libraries. In stark contrast, *H. influenzae* genes encoding 25 ABC transporters, 12 hydrolases, 3 transposases as well as other genes under-represented in the WGS libraries that we analyzed appear to have been easily clonable in *E. coli*. Although the reason for this is unclear, it is important to point out that *H. influenzae* has 20 ABC transporters, 12 hydrolases, 1 transposase and 1,253 other genes that have orthologs in *E. coli*. When expressed in *E. coli*, *H. influenzae* orthologs may be able to substitute for their endogenous counterparts.

Table 2. Gene content in gap regions from eight representative bacterial genomes

Bacterial species	Genome size (Mbp)	% G+C	Group	Clone coverage	Gaps coordinates	Gap size[§] (bp)	Predicted gene product(s) (% sequence identity)[¶]
Campylobacter jejuni RM1221	1.78	30.3	Epsilon proteobacteria	28X	432201–437459	5258	Cell wall-associated hydrolase (99)
					775067–781117	6050	Cell wall-associated hydrolase (99)
Clostrodium perfingens	2.8	28.2	Firmicutes	26X	1154718–1156372	1654	Acetyltransferase, GNAT family (93)
					1332980–1335344	2364	Sodium:dicarboxylate symporter (84), glutamate/aspartate transporter (76)
					2632682–2635499		
					2699916–2701486	2817	50S ribosomal protein L1 (90), antitermination protein NusG (100)
						1570	S4 DNA binding protein (99)
Coxiella burnetii	2.03	42.6	Gamma proteobacteria	46X	165319–167620	2301	ORF16-lacZ fusion protein (70)
					390887–391756	869	tRNA (guanine-N1)-methyltransferase (91)
					1210872–1211545	673	ompA-like transmembrane domain protein (79)
Haemophilus somnus	2.01	37.2	Gamma proteobacteria	19X	64011–67622	3611	50S ribosomal protein L2 (100)
					670482–671950	1468	Disulfide bond formation protein (87)
					1065664–1067388	1724	dGTP triphosphohydrolase (99)
					1070654–1075864	5210	30S ribosomal protein S1 (85), N-acetylglucosaminyl transferase (94))
					1451575–1457089	5514	Aconitate hydratase (97)
					1586270–1587875	1605	Transposase (94)
					1796398–1798513	2115	Dehydrogenase (90)
					1946022–1947422	1400	Cell wall-associated hydrolase (71)
Neisseria meningitidis	2.27	51.5	Beta proteobacteria	10X	60714–66905	6191	ComE operon protein (70), cell wall-associated hydrolase (81)
					72007–73104	1097	dTDP-D-glucose 4,6-dehydratase (86)
					482682–483600	918	30S ribosomal protein S20 (100)
					936541–937109	568	Oxydoreductase (52)
					1167633–1199108	31475	Dehydratase (98), sulfite reductase (100), UDP-N-acetylmuramate ligase (100), chaperone protein HscA (92)
					1715587–1717589	2002	Disulphide bond formation protein (98)
					1991503–1993303	1800	DNA-binding protein (60)
					2131809–2137741	5932	Cell wall-associated hydrolase (80)
Porphyromonas gingivalisW83	2.34	48.3	Bacteroidetes	6X	119966–124689	4723	Cell wall-associated hydrolase (66)
					879134–885762	6628	MATE efflux protein (87), ISPg1 transposase (95)
					1093278–1093924	646	ISPg1, transposase (65)
					1341739–1346469	4730	Cell wall-associated hydrolase (66)
					1536011–1537997	1986	Integrase (84)
					1752409–1758616	6207	Cell wall-associated hydrolase (65)
					2155773–2156870	1097	ISPg5 transposase Orf1 (99)
					2170919–2176102	5183	Cell wall-associated hydrolase (65)

Streptococcus agalactiae A909	2.1	35	Firmicutes	26X	22626–27883	5257	Metal-dependent RNase (98)
					70685–71857	1172	Alcohol dehydrogenase (90)
					138684–140594	1910	L-2-hydroxyisocaproate dehydrogenase (100)
					174744–179649	4905	Metal-dependent RNase (98)
					259829–264645	4816	Metal-dependent RNase (98)
					1290268–1291699	1331	Glyoxalase (100)
					1400547–1402041	1494	Cytidylate kinase (100)
					1446751–1448703	1952	Amino acid ABC transporter (92)
					2042583–2044168	1585	Transposase (100)
Streptococcus thermophilus LMD-9	1.91	39.1	Firmicutes	5X	67777–71441	3664	Metal-dependent RNase (98)
					174804–176572	1768	Permease (86)
					200169–201378	1209	Auxin efflux carrier (AEC) family permease (96)
					204806–207712	2906	Glucose-6-phosphate isomerase (84)
					358481–364331	5850	Ribosomal subunit interface protein (92)
					477586–478589	1003	Ammonia permease (78)
					535433–538175	2742	Sulfurtransferase (52), Uridine phosphorylase (100)
					701003–702384	1381	Glucose kinase (91)
					913226–915330	2104	Na+/alanine symporter (95)
					1060338–1061924	1586	Proteinase (99)
					1066129–1067735	1606	ABC transporter substrate binding protein (92)
					1359052–1360292	1240	ABC-type amino acid transport (100)
					1408078–1410316	2238	Folylpolyglutamate synthase (100)
					1514863–1515597	734	ABC transporter (78)
					1768097–1769633	1536	50S ribosomal protein L3 (100)
					1795254–1796311	1057	RNase H-like ribonuclease (100)
					1822111–1823819	1708	Transcriptional regulator, TetR (100)

Whole genome shotgun sequence reads were downloaded from the NCBI trace archive (www.ncbi.nlm.nih.gov/Traces/), and paired end sequences were aligned to their respective, completed genome using wuBLASTn (wublast.wustl.edu, Gish 1996–2006). Regions devoid of clone coverage were identified and used to search the protein sequence database genbank-nr using wuBLASTx.

There are a number of genes missing in our *H. influenzae* BAC library, but not from the WGS libraries of genome-sequencing projects listed here. These include genes coding for aminopeptidase P, GTP-binding protein, alanyl-tRNA synthetase, DNA polymerase III—epsilon subunit—and at least four genes involved in galactose metabolism. For the purpose of rebuilding and activating the *H. influenzae* genome, toxic but non-essential genes can likely be ignored, but toxic genes that are essential cannot. Genes that encode DNA polymerase III subunit epsilon along with ribosomal proteins and alanyl-tRNA synthetase appear toxic because they are missing from our libraries, yet they are part of the essential core of a minimal gene set required for cell viability.[25] Further, genes essential to the survival of *H. influenzae* have been identified by high-density transposon mutagenesis.[26] Genes missing from our libraries that also could not sustain transposon insertion include, again, those coding for DNA polymerase III subunit epsilon, numerous ribosomal proteins, alanyl-tRNA synthetase, plus the additional genes for argininosuccinate lyase, aminopeptidase P, galactose-1-phosphate uridylyltransferase and the galactose operon repressor. Unfortunately there is little information in the literature on the mechanism of toxicity of these proteins. It is believed that the peptide degrader aminopeptidase P is responsible for the inactivation of physiologically important peptides and cell detoxification by degradation of toxic peptides.[27] It is clear that interfering with endogenous aminopeptidase P could have unforeseen consequences for survival of *E. coli*. There has been a report on the toxicity of an aminoacyl-tRNA synthetase from *Bacillus stearothermophilus* overexpressed in *E. coli*.[28] Although the exact mechanisms have yet to be elucidated, the authors suggested that overproduction of tRNA synthetase (glutaminyl- or tyrosyl-) may lead to incorrect acetylation of substrates and misincorporation of amino acids, resulting in mistranslated proteins.[28,29]

To our knowledge, the meta-analysis analysis that we present here is the first to report on the commonality of protein-coding genes missing or under-represented in bacterial genomic libraries cloned in *E. coli*. The high clone coverage in our reassemblies makes it unlikely that gaps that we observe are due to under-sampling, but rather due to toxic gene content. This is supported by the occurrence of many of the same types of genes in independent species that, once expressed in a host, appear to interfere with native cellular processes.

Genome resolution

In prokaryotes, chromosomes are distinguished from episomes by different mechanisms of replication and copy number control and also by the presence of essential genes. Plasmid replication is not linked to the cell cycle or tied to replication of the host chromosome and plasmids are ultimately dispensable because they do not contain genes essential to the survival of the organism in all circumstances.[30,31] The DNA payload that can be supported by the F replicon is unknown and assembling a new chromosome in a host cell as an episome driven by a single F replicon is a reasonable proposition. In our system, this approach will give a hybrid *H. influenzae/E. coli* cell with a bipartite genome. A significant challenge will be resolving the donor and host elements of this bipartite genome to obtain a cell encoded and propagated by the donor *H. influenzae* genome alone. Several issues arise. First, the F replicon is derived from an *E. coli* plasmid but it does not have the properties of a shuttle vector and cannot propagate in *H. influenzae*. Therefore, upon completion of assembly of the donor *H. influenzae* genome, it must be propagating under the control of its native oriC, not the F replicon, and one of the key questions to be addressed in our system is the point at which the native *Haemophilus* oriC can assume this control. This could occur as soon as the *H. influenzae* oriC is present provided it can utilize *E. coli trans* factors, or it may be the case that expression of its own *H. influenzae* replication factors present elsewhere in its genome may be required. Given the extensive cross-expression between these species and the absence of all but one replication factor from the set of *H. influenzae* lethal genes, activation of the *H. influenzae* oriC is likely achievable.

The key factor in genome resolution is likely to be the balance of essential and lethal genes. If the donor *H. influenzae* genome contains all essential genes for cell survival, as the host genome already does, then the two components of the bipartite genome of the hybrid cell are redundant. In this case, the host genome is dispensable and, in principle, the host cell can be cured of its original genome. Curing might be accomplished by disrupting host chromosome partitioning, or by direct ablation of the host genome by a restriction enzyme with recognition sites in the host but not donor genome. Curing could be facilitated by antibiotic selection where the donor but not host genome carries resistance genes, or by adding an inducible "suicide" gene like the gyrase poison *ccd*B specifically to the host genome. Thus, means for resolving the donor and host genome can be envisioned, but these cannot ultimately be tested until the assembly of the donor genome is complete.

Regardless of mechanism, genome resolution will be dependent upon successful incorporation of all essential genes in the donor *H. influenzae* chromosome, and transcription and translation of their functional gene products. Thus, genes that are essential to the donor but toxic to the host present a conundrum. How can they be included if they kill their host? Again, a resolution to this problem will probably come only with testing a number of possible solutions. For example, lethal genes that are essential could be grouped on a single synthetic construct that is the final insertion into the donor chromosome. Addition of these genes could have the dual benefit of rendering the donor chromosome independent and

making cells that retain the host chromosome unviable, thereby facilitating genome resolution. Alternatively, the lethal yet essential coding sequences could be placed under the control of an inducible promoter or orthogonal expression system, and added at any time during genome assembly. Finally, empirical studies will be necessary to verify toxicity of suspected lethal genes, to determine their mechanism of toxicity, to establish whether toxicity can be altered by titrating expression, to test whether toxic genes can be replaced by orthologs from the host genome, and to explore the possibility of mitigating toxicity via alternative growth conditions.

Conclusions

Here we describe a central problem in synthetic biology today; the activation of isolated genetic material to generate the organism that it encodes. This will be a key hurdle in any top-down approach to producing a synthetic free-living organism. The most-feasible approach to genome activation today is stepwise assembly in a host cell. The critical points in this proposed process that require further research are (1) methods for physical manipulation of large segments of naked DNA, (2) determining the degree of cross talk between donor genetic material and host transcriptional machinery that is necessary for genome exchange, and finally (3) compatibility of coding sequences between donor and host cells, specifically with regard to delivery and activation of donor genes that are lethal or otherwise negatively impact host fitness. We have begun to explore each of these areas in our model system that involves stepwise transfer of the *H. influenzae* genome to *E. coli*, and we argue that these three stated challenges should not be insurmountable. We have propagated large segments of the *H. influenzae* genome in *E. coli*, demonstrated extensive regulatory cross-talk between these species, and mapped *H. influenzae* genes (plus genes from other bacteria) that are disruptive to an *E. coli* host. There are challenges remaining that can only be addressed at later stages of genome integration, such as planning the best approach for delivery of toxic genes and establishing the necessary steps for genome resolution. Although the DNA that we are working with is cloned from native *Haemophilus* cells, one or eventually all of these chromosomal elements could later be replaced by synthetic constructs, either native sequence or sequence that encodes new functionality. This system will provide a platform to address long-standing academic questions, such as what is the minimal gene set required to support cellular life in a given environment. Minimal gene sets that are now speculative can actually be tested once there is a method for their assembly and activation. Further, such a system will provide a means for constructing synthetic chromosomes for organisms that may be of utility in biotechnolgy. There will, however, undoubtedly be problems that we cannot foresee. For example, while we have propagated large *H. influenzae* genomic segments individually in *E. coli*, perhaps new incompatibilities will arise as more of the genome is assembled and biochemical pathways with components encoded on different genome segments are fully restored. An advantage of our model system, and of a synthetic approach in general, is that attention to problems that arise during synthesis may be missed by an observational approach alone, and attention to these problems has the potential to offer new insight into biology.

Acknowledgments

We thank Dr. Rosie Redfield (University of British Columbia) for helpful advice and for providing bacterial strains. RAH is a Michael Smith Foundation for Health Research Scholar.

References

1. Luisi PL. 2002. Toward the engineering of minimal living cells. Anat Rec 268:208–214.
2. Forster AC, Church GM. 2006. Towards synthesis of a minimal cell. Mol Syst Biol 2:45.
3. Benner SA, Sismour AM. 2005. Synthetic biology. Nat Rev Genet 6:533–543.
4. Smith HO, Hutchison CA III, Pfannkoch C, Venter JC. 2003 Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci USA 100:15440–15445.
5. Tian J, Gong H, Sheng N, Zhou X, Gulari E, et al. 2004. Accurate multiplex gene synthesis from programmable DNA microchips. Nature 432:1050–1054.
6. Carlson R. 2003. The pace and proliferation of biological technologies. Biosecur Bioterror 1:203–214.
7. Yount B, Curtis KM, Fritz EA, Hensley LE, Jahrling PB, et al. 2003. Reverse genetics with a full-length infectious cDNA of severe acute respiratory syndrome coronavirus. Proc Natl Acad Sci USA 100:12995–13000.
8. Cello J, Paul AV, Wimmer E. 2002. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science 297:1016–1018.
9. Tumpey TM, Basler CF, Aguilar PV, Zeng H, Solorzano A, et al. 2005. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science 310:77–80.
10. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512.
11. Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, et al. 1997. The complete genome sequence of Escherichia coli K-12. Science 277:1453–1462.
12. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, et al. 1996. Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. Curr Biol 6:279–291.
13. Itaya M, Tsuge K, Koizumi M, Fujita K. 2005. 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.
14. Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, et al. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 89:8794–8797.
15. Kornberg A, Baker TA. 1992. DNA Replication. University Science Books, USA p. 658–662.
16. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, et al. 2000 An efficient recombination system for chromosome engineering in Escherichiacoli. Proc Natl Acad Sci USA 97:5978–5983.

17. Hoess RH, Ziese M, Sternberg N. 1982. P1 site-specific recombination: nucleotide sequence of the recombining sites. Proc Natl Acad Sci USA 79:3398–3402.
18. Schlake T, Bode J. 1994. Use of mutated FLP recognition target (FRT) sites for the exchange of expression cassettes at defined chromosomal loci. Biochemistry 33:12746–12751.
19. Missirlis PI, Smailus DE, Holt RA. 2006. A high-throughput screen identifying novel functional Cre-LoxP inverted repeat and spacer mutants. BMC Genomics 7:73–79.
20. Cheng H, Zhang H-Z, Shen W-A, Liu Y-F, Ma FC. 2003. Expression of RNase H of human hepatitis B virus polymerase in Escherichia coli. World J. Gastroenterol 9:513–515.
21. Wei X, Peterson DL. 1996. Expression, Purification, and Characterization of an Active RNase H Domain of the Hepatitis B Viral Polymerase. J Biol Chem 271:32617–32622.
22. Nomura M, Yates JL, Dean D, Post LE. 1980. Feedback regulation of ribosomal protein gene expression in Escherichia coli: structural homology of ribosomal RNA and ribosomal protein mRNA. Proc Natl Acad Sci USA 77:7084–7088.
23. Ramakrishnan V, Gerchman SE. 1991. Cloning, sequencing, and overexpression of genes for ribosomal proteins from Bacillus stearothermophilus. J Biol Chem 266:880–885.
24. Hiniker A, Bardwell JC. 2004. In vivo substrate specificity of periplasmic disulfide oxidoreductases. J Biol Chem 279:12967–12973.
25. Gil R, Silva FJ, Pereto J, Moya A. 2004. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68:518–537.
26. Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N, et al. 2002. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci USA 99:966–971.
27. Gonzales T, Robert-Baudouy J. 1996. Bacterial aminopeptidases: Properties and functions. FEMS Microbiology Reviews 18:319–344.
28. Bedouelle H, Guez V, Vidal-Cros A, Hermann M. 1990. Overproduction of Tyrosyl-tRNA Synthetase Is Toxic toEscherichia coli: a Genetic Analysis. J bacteriology 172:3940–3945.
29. Swanson R, Hoben P, Sumner-Smith M, Uemura H, Watson L, et al. 1988. Accuracy of in vivo aminoacylationrequires proper balance of tRNA and aminoacyl-tRNA synthetase. Science 242:1548–1551.
30. Dasgupta S, Lobner-Olesen A. 2004. Host controlled plasmid replication: Escherichia coli minichromosomes. Plasmid 52:151–168.
31. Egan ES, Fogel MA, Waldor MK. 2005. Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol 56:1129–1138.