(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 February 2008 (28.02.2008)

PCT

(10) International Publication Number
WO 2008/024129 A2

(51) International Patent Classification:
*C07H 21/04* (2006.01) *C12P 1/04* (2006.01)
*C12N 5/06* (2006.01)

(21) International Application Number:
PCT/US2006/046803

(22) International Filing Date:
6 December 2006 (06.12.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/742,542 6 December 2005 (06.12.2005) US

(71) Applicant *(for all designated States except US)*: **J. CRAIG VENTER INSTITUTE** [US/US]; 9704 Medical Center Drive, Rockville, MD 20850 (US).

(71) Applicant *(for US only)*: **HUTCHISON, Clyde, A., III** [US/US]; c/o J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **VENTER, Craig, J.** [US/US]; c/o J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 (US). **SMITH, Hamilton, O.** [US/US]; c/o J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850 (US).

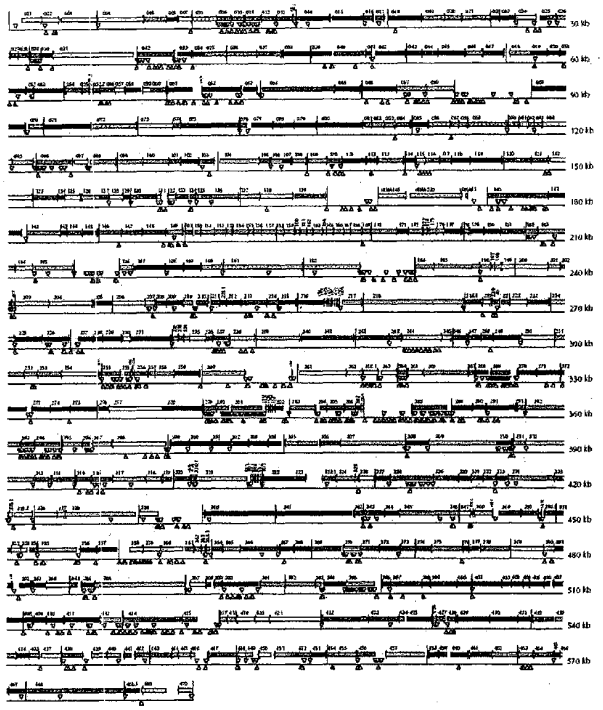
(74) Agents: **BATHURST, Brian** et al.; Carr & Ferrell LLP, 2200 Geng Road, Palo Alto, CA 94303 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH,

*[Continued on next page]*

(54) Title: SYNTHETIC GENOMES



Map of *Mycoplasma genitalium*. The triangles mark the position of transposon insertions (upper triangles are from Smith *et al.* (1999) *Proc Natl Acad Sci USA* 87, 826-830 1999. Vertical lines delineate the borders of the "5 kb" segments.

(57) Abstract: Methods are provided for constructing a synthetic genome, comprising generating and assembling nucleic acid cassettes comprising portions of the genome, wherein at least one of the nucleic acid cassettes is constructed from nucleic acid components that have been chemically synthesized, or from copies of the chemically synthesized nucleic acid components. In one embodiment, the entire synthetic genome is constructed from nucleic acid components that have been chemically synthesized, or from copies of the chemically synthesized nucleic acid components. Rational methods may be used to design the synthetic genome (e.g., to establish a minimal genome and/or to optimize the function of genes within a genome, such as by mutating or rearranging the order of the genes). Synthetic genomes of the invention may be introduced into vesicles (e.g., bacterial cells from which part or all of the resident genome has been removed, or synthetic vesicles) to generate synthetic cells. Synthetic genomes or synthetic cells may be used for a variety of purposes, including the generation of synthetic fuels, such as hydrogen or ethanol.

WO 2008/024129 A2

GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *without international search report and to be republished upon receipt of that report*

# SYNTHETIC GENOMES

By J. Craig Venter, Hamilton O. Smith and Clyde A. Hutchison III

## CROSS-REFERENCE TO RELATED APPLICATIONS

**[001]** The present application claims benefit and priority from U.S. Provisional Patent Application Serial No. 60/742,542 filed on Dec. 6, 2005, entitled, "Synthetic Genomes;" the present application is related to U.S. Provisional Patent Application Serial No. 60/752,965 filed on Dec. 23, 2005, entitled, "Introduction of Genomes into Microorganisms;" U.S. Provisional Patent Application Serial No. 60/741,469 filed on Dec. 2, 2005, entitled, "Error Correction Method;" and U.S. Non-Provisional Patent Application Serial No. 11/502,746 filed on Aug. 11, 2006, entitled "*In Vitro* Recombination Method," all of which are incorporated herein by reference.

## STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

**[002]** This invention was made with U.S. government support (DOE grant number DE-FG02-02ER63453). The government has certain rights in the invention.

## BACKGROUND OF THE INVENTION

### Field of the Invention

**[003]** The present invention relates generally to molecular biology, and more particularly to synthetic genomes.

Description of Related Art

[004] Conventional genetic engineering techniques are limited to allowing manipulation of existing sequences. It would thus be desirable to have the ability to implement dramatic alterations and arrangements of genetic content, beyond that made possible by conventional techniques. Consequently, there is a need for synthetic genomes.

## SUMMARY OF THE INVENTION

**[005]** Embodiments and methods are provided for the design, synthesis, assembly and expression of synthetic genomes. Included are methods for rationally designing components of a genome; generating small nucleic acid fragments and assembling them into cassettes comprising portions of the genome; correcting errors in the sequences of the cassettes; cloning the cassettes (*e.g.*, by *in vitro* methods such as rolling circle amplification); assembling the cassettes to form a synthetic genome (*e.g.*, by methods of *in vitro* recombination); and transferring the synthetic genome into a biochemical system (*e.g.*, by transplanting it into an intact cell, ghost cell devoid of functioning DNA, or other vesicle). In one embodiment, the synthetic genome comprises sufficient information to achieve replication of a vesicle (*e.g.*, a cell) in which it resides. The technology extends to useful end products that a synthetic genomic system can produce, such as energy sources (*e.g.*, hydrogen or ethanol), and biomolecules such as therapeutics and industrial polymers.

**[006]** Included are methods for constructing a synthetic genome, comprising generating and assembling the nucleic acid components of the genome, wherein at least part of the genome is constructed from nucleic acid components that have been chemically synthesized, or from copies of chemically synthesized nucleic acid components. In one embodiment, an entire synthetic genome is constructed from nucleic acid components that have been chemically synthesized, or from copies of chemically synthesized nucleic acid components. Further, a synthetic genome may be a synthetic cellular genome (a genome which comprises all of the sequences required for replication of a vesicle (e.g., a cell or synthetic vesicle) in which it resides).

**[007]** Methods are provided for constructing a synthetic cell, comprising use of certain exemplary methods to construct a synthetic genome

and introducing (transplanting) the synthetic genome into a vesicle (e.g., a cell or a synthetic membrane bound vesicle). Another method includes constructing a self-replicating synthetic cell, comprising use of exemplary methods to construct a synthetic cellular genome and introducing (transplanting) the synthetic cellular genome into a vesicle (e.g., a cell or a synthetic membrane bound vesicle), under conditions effective for the synthetic cell to replicate. Further methods include producing a product of interest, comprising culturing an exemplary synthetic cell under conditions effective to produce the product. When the product is produced from a synthetic cell comprising a synthetic cellular genome, the genome is contacted with the vesicle under conditions effective to replicate the synthetic cell and to produce the product.

**[008]** Other exemplary methods include making a synthetic cell, comprising removing part of or all of the resident (original) genome from a microorganism, such as a unicellular microorganism (*e.g.*, a bacterium, fungus, etc.) and replacing it with a synthetic genome that is foreign to the organism (*e.g.*, is from a different species of microorganism (*e.g.*, bacterium)), which exhibits at least one property that is different from the resident genome. Various exemplary embodiments include a synthetic cell produced by this method.

**[009]** One exemplary embodiment includes a synthetic genome that is capable of directing replication of a vesicle (*e.g.*, cell) in which it resides, under particular environmental (*e.g.*, nutritional or physical) conditions. In one embodiment, the cellular genome is supplemented in the vesicle (*e.g.*, cell) by small molecules, such as nutrients, ATP, lipids, sugars, phosphates etc, which serve as precursors for structural features or substrates for metabolic functions; and/or is supplemented with complex components, such as ribosomes, functional cell membranes, etc. These additional elements may complement or facilitate the ability of the genome to achieve (*e.g.*,

program) replication of the vesicle/cell. In another embodiment, the sequences in the genome are capable of providing all of the machinery and components required to produce a cell and to allow the cell to replicate under particular energy or environmental (*e.g.*, nutritional) conditions.

**[0010]** Further embodiments and methods include a "minimal genome" that may serve as a platform for introducing other sequences of interest, such as genes encoding biologic agents (*e.g.*, therapeutic agents, drugs, vaccines or the like), or genes encoding products that, in the presence of suitable precursors, can produce useful compounds (*e.g.*, biofuels, industrial organic chemicals, etc.). In one embodiment, the other sequences of interest result in the production of products in sufficient quantities to be commercially valuable.

**[0011]** According to one exemplary embodiment and method, a synthetic version of the *Mycoplasma genitalium* genome having 482 protein-coding genes and 43 RNA genes comprising a 580-kilobase circular chromosome is assembled from gene cassettes. Each cassette may be made from chemically synthesized oligonucleotides. Several versions of each cassette may be made such that combinatorial assembly into a complete chromosome results in millions of different genomes. These genomes may be tested for functionality by "genome transplantation," replacement of a cell's resident chromosome by the synthetic genome. According to further embodiments and methods, synthetic cells may be assembled from various subcellular components. Additionally, a genome in a cell-free environment may be established, comprising the necessary transcriptional and translation "machinery" to express genes.

## BRIEF DESCRIPTION OF THE DRAWINGS AND EXAMPLES

**[0012]** FIG. 1 shows an illustration of suitable oligonucleotides for preparing *M. genitalium*.

**[0013]** Example I illustrates the use of mycoplasma comparative genomics to identify genes that may be involved in a minimal gene set for mycoplasma.

**[0014]** Example II shows the design of 682 48-mer oligonucleotides, and the assembly of those oligonucleotides into three overlapping segments (cassettes).

**[0015]** Example III shows the synthesis of 5 kb cassettes of an essential region of the *M. genitalium* genome.

## DETAILED DESCRIPTION OF THE INVENTION

**[0016]** The following descriptions of various terms as used herein are not exhaustive and may include other descriptive matter.

**[0017]** "Cellular genome" or a "synthetic cellular genome" is a genome that comprises sequences which encode and may express nucleic acids and proteins required for some or all of the processes of transcription, translation, energy production, transport, production of cell membranes and components of the cell cytoplasm, DNA replication, cell division, and the like. A "cellular genome" differs from a viral genome or the genome of an organelle, at least in that a cellular genome contains the information for replication of a cell, whereas viral and organelle genomes contain the information to replicate themselves (sometimes with the contribution of cellular factors), but they lack the information to replicate the cell in which they reside.

**[0018]** "Foreign" gene or genome is a gene or genome derived from a source other than the resident (original) organism, e.g., from a different species of the organism.

**[0019]** "Genome" may include viral genomes, the genomes of organelles (e.g., mitochondria or chloroplasts), and genomes of self-replicating organisms, such as bacteria, yeast, archebacteria, or eukaryotes. A genome may also be an entirely new construct for an organism that does not fall into any known Linnean category. In one embodiment, the genes are from a microorganism, *e.g.*, a unicellular microorganism, such as a bacterium. The genes may be in the order found in the microorganism, or they may be shuffled; and mutant versions of some of the genes may also be included.

**[0020]** "Membrane-bound vesicle," refers to a vesicle in which a lipid-based protective material encapsulates an aqueous solution.

[0021] "Minimal genome," with respect to a cell, as used herein, refers to a genome consisting of or consisting essentially of a minimal set of genetic sequences that are sufficient to allow for cell survival under specified environmental (*e.g.*, nutritional) conditions. A "minimal genome," with respect to an organelle, as used herein, refers to a genome consisting of or consisting essentially of a minimal set of genetic sequences that are sufficient to allow the organelle to function. A minimal genome must contain sufficient information to allow the cell or organelle to carry out essential biological processes, such as, for example, transcription, translation, use of an energy source, transport of salts, nutrients and the like into and out of the organelle or cell, etc. A "minimal replicating genome," with respect to either a cell or an organelle, contains, in addition, genetic sequences sufficient to allow for self replication of the cell or organelle, Thus, a "minimal replicating synthetic genome" is a single polynucleotide or group of polynucleotides that is at least partially synthetic and that contains the minimal set of genetic sequences for a cell or organelle to survive and replicate under specific environmental conditions.

[0022] "Nucleic acid" and "Polynucleotide" are used interchangeably herein. They include both DNA and RNA. Other types of nucleic acids, such as PNA, LNA, modified DNA or RNA, etc. are also included, provided that they can participate in the synthetic operations described herein and exhibit the desired properties and functions. A skilled worker will recognize which forms of nucleic acid are applicable for any particular embodiment or method described herein.

[0023] "Synthetic genome," includes a single polynucleotide or group of polynucleotides that contain the information for a functioning organelle or organism to survive and, optionally, replicate itself where particular environmental (e.g., nutritional or physical) conditions are met. All

or at least part of the genome (e.g., a cassette) is constructed from components that have been chemically synthesized, or from copies of chemically synthesized nucleic acid components. The copies may be produced by any of a variety of methods, including cloning and amplification by *in vivo* or *in vitro* methods. In one embodiment, an entire genome is constructed from nucleic acid that has been chemically synthesized, or from copies of chemically synthesized nucleic acid components. Such a genome is sometimes referred to herein as a "completely synthetic" genome. In other embodiments, one or more portions of the genome may be assembled from naturally occurring nucleic acid, nucleic acid that has been cloned, or the like. Such a genome is sometimes referred to herein as a "partially synthetic" genome.

**[0024]** Synthetic genomes offer numerous advantages over traditional recombinant DNA technology. For example, the selection and construction of synthetic genome sequences allow for easier manipulation of sequences than with classical recombination techniques, and permits the construction of novel organisms and biological systems. Furthermore, various embodiments and methods are amenable to automation and adaptation to high throughput methods, allowing for the production of synthetic genomes and synthetic cells by computer-mediated and robotic methods that do not require human intervention. The inventive technology opens the door to an integrated process of synthetic genome design, construction, introduction into a biological system, biological production of useful products, and recursive improvement to the design.

**[0025]** Various forms of rational or intelligent design of nucleic acids may be employed according to various exemplary embodiments and methods. According to one method, a gene set is identified that constitutes a minimal genome, e.g., of a bacterium, such as *Mycoplasma genitalia* (*M. genitalium*), *M. capricolum* (e.g., subspecies *capricolum*), *E. coli*, *B. subtilis*, or

others. One or more conventional or novel methods, or combinations thereof, may be used to accomplish this end. One method includes using random saturation global transposon mutagenesis to knock out the function of each gene in a microbial genome (e.g., a bacterial genome) individually, and to determine on this basis putative genes that may be eliminated without destroying cell viability. See, e.g., *Smith et al.* (1999) *Proc Natl Acad Sci USA* 87, 826-830. Another method is to use comparative genomics of a variety of related genomes (*e.g.*, analyzing the sequences of orthologous organisms, metagenomics, etc.) to predict common genes which are basic to the function of a microorganism of interest (*e.g.*, a bacterium), *e.g.*, to identify genes common to all members of a taxon. Existing databases may be used, or new databases may be generated by sequencing additional organisms, using conventional methods. According to one method, the identification of genes in a minimal genome is facilitated by isolating and expanding clones of individual cells, using a method for disrupting cell aggregates. Example I herein illustrates the use of mycoplasma comparative genomics to identify genes that may be involved in a minimal gene set for mycoplasma.

**[0026]** Following the identification of a putative minimal set of genes required for viability and, optionally, replication under a defined set of conditions, a candidate minimal genome may be constructed as described herein. According to one method, a set of overlapping nucleic acid cassettes are constructed, each generally having about 5 kb, which comprise subsets of the genes; and the cassettes are then assembled to form the genome. The function/activity of the genome may be further studied by introducing the assembled genome into a suitable biological system and monitoring one or more functions/activities encoded by the genome. The synthetic genome may be further manipulated, for example, by modifying (*e.g.*, deleting, altering individual nucleotides, etc.) portions of genes or deleting entire genes within

one or more of the cassettes; by replacing genes or cassettes by other genes or cassettes, such as functionally related genes or groups of genes; by rearranging the order of the genes or cassettes (*e.g.*, by combinatorial assembly); etc. The consequences of such manipulations may be examined by re-introducing the synthetic genes into a suitable biological system. Factors that may be considered include, *e.g.*, growth rate, nutritional requirements and other metabolic factors. In this manner, one may further refine which genes are required for a minimal genome.

**[0027]** Another aspect of rational design according to further methods involves the determination of which sites within a synthetic genome may withstand insertions, such as unique identifiers (*e.g.*, watermarks), expressible sequences of interest, etc., without disrupting gene function. In general, sites within a genome that can withstand such disruption lie at the junctions between genes, in non-coding regions, or the like.

**[0028]** Another aspect of rational design according to even further methods includes the selection of suitable regulatory control elements. For instance, in the case of prokaryotic-type cells, such regulatory control elements include promoters, terminators, signals for the modulation of gene expression (*e.g.*, repressors, stimulatory factors, etc.), signals involved in translation, signals involved in modification of nucleic acids (*e.g.*, by methylation), etc. In the case of eukaryotic-type cells, further regulatory control elements include signals involved in splicing, post-translational modification, etc.

**[0029]** A further design procedure that may be applied is the design of suitable cassettes to be combined to form a synthetic genome. Upon generating synthetically a substantially exact copy of a genome of known sequence, cassettes are selected which lie adjacent to one another in that

sequence and, preferably, which overlap one another in order to facilitate the joining of the cassettes. Factors to be considered in designing the cassettes include, *e.g.*, that the segments be about 4 to 6.5 kb in length, not including overlaps; that the segments contain only whole genes, except for the overlaps; and that the overlaps with adjacent sequences are about 200-250 (*e.g.*, about 216) bp. Thus, each synthetic about 5 kb piece is a cassette comprising one or more complete genes. An illustration of cassettes that are designed, following these constraints, for the synthesis of *M. genitalium*, is shown in Figure 1.

**[0030]** In another embodiment, cassettes are designed to be interchangeable, *e.g.*, the cassettes are bounded by unique sequences such as restriction enzyme or adaptor sites, which allow the cassettes to be excised from the genome. The cassettes may be: removed, manipulated (*e.g.*, mutated) and returned to the original location in the genome; substituted by other cassettes, such as cassettes having functionally related genes; re-assorted (rearranged) with other cassettes, for example in a combinatorial fashion; etc. Mutations or other changes may be introduced, for example, by inserting mutated nucleic acid from a natural source; by site-directed mutagenesis, either *in vivo* or *in vitro*; by synthesizing nucleic acids to contain a desired variation, etc,

**[0031]** As noted herein, genes of interest which directly or indirectly lead to the production of desired products (*e.g.*, therapeutic agents, biofuels, etc.) may be present in a synthetic genome. To optimize the production of such products, the genes may be manipulated and the effects of the manipulations evaluated by introducing the modified synthetic genome into a biological system. Features may be altered including, *e.g.*, coding or regulatory sequences, codon usage, adaptations for the use of a particular growth medium, etc. Among the factors that may be evaluated are, *e.g.*, the amount of desired end product produced, tolerance to end product,

robustness, etc. Additional rounds of such manipulations and assessments may be performed to further the optimization. Using such iterative design and testing procedures (sometimes referred to herein as "reiterative" or "recursive" improvement, "recursive design," or "use of feed-back loops") one may optimize the production of a product of interest or may optimize growth of a synthetic cell. One may make predictions about cellular behavior, which may be confirmed or, if desired, modified. Furthermore, by designing and manipulating genes in a synthetic genome according to methods described herein, experimental studies may be performed, *e.g.*, to identify features that are important for the maintenance, division, etc. of cells, features that are important to impart "life" to an organism, etc.

**[0032]** A variety of methods may be used to generate and assemble nucleic acid cassettes. As a first step, a cassette of interest is generally subdivided into smaller portions from which it may be assembled. Generally, the smaller portions are oligonucleotides of about between about 30 nt and about 1 kb, e.g., about 50 nt (*e.g.*, between about 45 and about 55). In one embodiment, the oligonucleotides are designed so that they overlap adjacent oligonucleotides, to facilitate their assembly into cassettes. For example, for *M. genitalium*, the entire sequence may be divided into a list of overlapping 48-mers with 24 nucleotide overlaps between adjacent top and bottom oligonucleotides. An illustration of suitable oligonucleotides for preparing *M. genitalium* is shown in Figure 1. The oligonucleotides may be synthesized using conventional methods and apparatus, or they may be obtained from well-known commercial suppliers.

**[0033]** Among the many methods which can be used to assemble oligonucleotides to form longer molecules, such as the cassettes described herein, are those described, e.g., in *Stemmer et al.* (1995) (*Gene* 164, 49-53) and *Young et al.* (2004) (*Nucleic Acids Research* 32, e59). One suitable method, called

polymerase cycle assembly (PCA), was used by *Smith et al.* (2003) (*Proc Natl Acad Sci USA* 100, 15440-5) for the synthesis of the 5386 nt genome of ϕX174. It is generally preferable to clone and/or amplify these cassettes in order to generate enough material to manipulate readily. In some embodiments, the cassettes are cloned and amplified by conventional cell-based methods. In one embodiment, *e.g.*, when it is difficult to clone a cassette by conventional cell-based methods, the cassettes are cloned *in vitro*. One such *in vitro* method, which is discussed in co-pending U.S. Provisional Patent Application Serial Nos. 60/675,850; 60/722,070; and 60/725,300, uses rolling circle amplification, under conditions in which background synthesis is significantly reduced.

**[0034]** Cassettes which may be generated according to various exemplary methods may be of any suitable size. For example, cassettes may range from about 1 kb to about 20 kb in length. A convenient size is about 4 to about 7 kb, *e.g.*, about 4.5 to about 6.5 kb, preferably about 5 kb. The term "about" with regard to a particular polynucleotide length, as used herein, refers to a polynucleotide that ranges from about 10% smaller than to about 10% greater than the size of the polynucleotide. In order to facilitate the assembly of cassettes, it is preferable that each cassette overlaps the cassettes on either side, *e.g.*, by at least about 50, 80, 100, 150, 200, 250 or 1300 nt. Larger constructs (up to the size of, *e.g.*, a minimal genome) comprising groups of such cassettes are also included, and may be used in a modular fashion according to various exemplary embodiments and methods.

**[0035]** A variety of methods may be used to assemble the cassettes. For example, cassettes may be assembled *in vitro*, using methods of recombination involving "chew-back" and repair steps, which employ either 3' or 5' exonuclease activities, in a single step or in multiple steps. Alternatively, the cassettes may be assembled with an *in vitro* recombination system that

includes enzymes from the *Dienocuccus radiodurans* homologous recombination system. Methods of *in vivo* assembly may also be used.

**[0036]** Example II describes the generation of a synthetic mouse mitochondrial genome of 16.3 kb by the assembly of three cassettes. Example II shows the design of 682 48-mer oligonucleotides, and the assembly of those oligonucleotides into three overlapping segments (cassettes). The oligonucleotides are then assembled into cassettes, by such methods as the method described in *Smith et al.* (2003), *supra*, modified in order to reduce heat damage to the synthetic DNA.

**[0037]** According to one method, once a cassette is assembled, its sequence may be verified. It is usually desirable to remove errors which have arisen during the preparation of the cassettes, *e.g.*, during the synthesis or assembly of the nucleic acid components. Among the error correction methods which may be used are; (1) methods to modify, tag and/or separate mismatched nucleotides so that amplification errors may be prevented; (2) methods of global error correction, using enzymes to recognize and cleave mismatches in DNA, having known or unknown sequences, to produce fragments from which the errors may be removed and the remaining error-free pieces reassembled; (3) methods of site-directed mutagenesis; and (4) methods to identify errors, select portions from independent synthetic copies which are error-free, and assemble the error-free portions, *e.g.*, by overlap extension PCR (OE-PCR). Other methods to recognize errors include, *e.g.*, the use of isolated mismatch or mutation recognition proteins, hybridization of oligonucleotide-fluorescent probe conjugates, electrophoretic/DNA chip methods, and differential chemical cleavage with reagents assaying for base access ability either in solution or the solid phase; such methods may be combined with conventional procedures to remove errors.

[0038] In one embodiment, one or more identifying features, such as a unique sequence (*e.g.*, encoding a particular symbol or name, or, *e.g.*, spelling with the alphabet letter designations for the amino acids) or an identifiable mutation which does not disrupt function are introduced into the synthetic genome. Such sequences, sometimes referred to herein as "watermarks," may serve not only to show that the genome has, in fact, been artificially synthesized and to enable branding and tracing, but also to distinguish the synthetic genome from naturally occurring genomes. Often, genes or cassettes contain selectable markers, such as drug resistance markers, which aid in selecting nucleic acids that comprises the genes or cassettes. The presence of such selectable markers may also distinguish the synthetic genomes from naturally occurring nucleic acids. A synthetic genome which is identical to a naturally occurring genome, but which contains one or more identifying markers as above, is sometimes referred to herein as being "substantially identical" to the naturally occurring genome.

[0039] A synthetic genome according to one embodiment may be present in any environment that allows for it to function. For example, a synthetic genome may be present in (*e.g.*, introduced into) any of the biological systems described herein, or others. The functions and activities of a synthetic genome, and the consequences of modifying elements of the genome, can be studied in a suitable biological system. Furthermore, a suitable biological system allows proteins of interest (*e.g.*, therapeutic agents) to be produced. In some embodiments, if suitable substrates are provided, downstream, non-proteinaceous products, such as energy sources (*e.g.*, hydrogen or ethanol) may also be produced, *e.g.*, in commercially useful amounts.

[0040] A variety of suitable biological systems may be used according to various embodiments and methods. For example, in one

embodiment, a synthetic genome is contacted with a solution comprising a conventional coupled transcription/translation system. In such a system, the nucleic acid may be able to replicate itself, or it may be necessary to replenish the nucleic acid, *e.g.*, periodically.

**[0041]** In another embodiment, a synthetic genome is introduced into a vesicle such that the genome is encapsulated by a protective lipid-based material. In one embodiment, the synthetic genome is introduced into a vesicle by contacting the synthetic genome, optionally in the presence of desirable cytoplasmic elements such complex organelles (*e.g.*, ribosomes) and/or small molecules, with a lipid composition or with a combination of lipids and other components of functional cell membranes, under conditions in which the lipid components encapsulate the synthetic genome and other optional components to form a synthetic cell. In other embodiments, a synthetic genome is contacted with a coupled transcription/translation system and is then packaged into a lipid-based vesicle. In a further embodiment, the internal components are encapsulated spontaneously by the lipid materials.

**[0042]** Exemplary embodiments also include a synthetic genome introduced into a recipient cell, such as a bacterial cell, from which some or all of the resident (original) genome has been removed. For example, the entire resident genome may be removed to form a ghost cell (a cell devoid of its functional natural genome) and the resident genome may be replaced by the synthetic genome. Alternatively, a synthetic genome may be introduced into a recipient cell which contains some or all of its resident genome. Following replication of the cell, the resident (original) and the synthetic genome will segregate, and a progeny cell will form that contains cytoplasmic and other epigenetic elements from the cell, but that contains, as the sole genomic material, the synthetic genome (*e.g.*, a copy of a synthetic genome). Such a cell is a synthetic cell according to various embodiments and methods, and differs

from the recipient cell in certain characteristics, *e.g.*, nucleotide sequence, nucleotide source, or non-nucleotide biochemical components.

**[0043]** A variety of *in vitro* methods may be used to introduce a genome (synthetic, natural, or a combination thereof) into a cell. These methods include, *e.g.*, electroporation, lipofection, the use of gene guns, etc. In one embodiment, a genome, such as a synthetic genome, is immobilized in agar; and the agar plug is laid on a liposome, which is then inserted into a host cell. In some embodiments, a genome is treated to fold and compress before it is introduced into a cell. Methods for inserting or introducing large nucleic acid molecules, such as bacterial genomes, into a cell are sometimes referred to herein as chromosome transfer, transport, or transplantation.

**[0044]** According to one embodiment, a synthetic cell may comprise elements from a host cell into which it has been introduced, *e.g.*, a portion of the host genome, cytoplasm, ribosomes, membrane, etc. In another embodiment, the components of a synthetic cell are derived entirely from products encoded by the genes of the synthetic genome and by products generated by those genes. Of course, nutritive, metabolic and other substances as well as physical conditions such as light and heat may be provided externally to facilitate the growth, replication and expression of a synthetic cell.

**[0045]** Various exemplary methods may be readily adapted to computer-mediated and/or automated (*e.g.*, robotic) formats. Many synthetic genomes (including a variety of combinatorial variants of a synthetic genome of interest) may be prepared and/or analyzed simultaneously, using high throughput methods. Automated systems for performing various methods as described herein are included. An automated system permits design of a desired genome from genetic components by selection using a bioinformatics

computer system, assembly and construction of numerous genomes and synthetic cells, and automatic analysis of their characteristics, feeding back to suggested design modifications.

**[0046]** While various embodiments and methods have been described herein, it should be understood that they have been presented by way of example only, and not limitation. Further, the breadth and scope of a preferred embodiment should not be limited by any of the above-described exemplary embodiments.

# CLAIMS

What is claimed is:

1. A method for constructing a synthetic genome comprising:

assembling nucleic acid cassettes that comprise portions of the synthetic genome, wherein at least one of the nucleic acid cassettes is constructed from nucleic acid components that have been chemically synthesized, or from copies of chemically synthesized nucleic acid components.

2. The method of claim 1, wherein one or more of the nucleic acid cassettes are prepared by assembling chemically synthesized, overlapping oligonucleotides of about 50 nucleotides.

3. The method of claim 1, wherein the cassettes are about 4 kilobases to about 7 kilobases in length.

4. The method of claim 1, wherein the cassettes are about 4.5 kilobases to about 6.5 kilobases in length.

5. The method of claim I, wherein the cassettes are about 5 kilobases in length.

6. The method of claim 1, wherein the cassettes overlap adjacent cassettes by at least about 200 nucleotides.

7. The method of claim 1, wherein the synthetic genome is a eukaryotic cellular organelle.

8. The method of claim 1, wherein the synthetic genome is a bacterial genome.

9. The method of claim 1, wherein the synthetic genome is a minimal genome.

10. The method of claim 1, wherein the synthetic genome is a minimal replicating genome.

11. The method of claim 1, wherein the synthetic genome is substantially identical to a naturally occurring genome.

12. The method of claim 1, wherein the synthetic genome is a non-naturally occurring genome.

13. The method of claim 1, wherein one or more of the cassettes can be readily removed and replaced in the synthetic genome.

14. The method of claim 1, wherein an entire synthetic genome is constructed from nucleic acid components that have been chemically synthesized, or from copies of the chemically synthesized nucleic acid components.

15. The method of claim 1, wherein the synthetic genome is a synthetic cellular genome.

16. The method of claim 15, wherein an entire synthetic cellular genome is constructed from nucleic acid components that have been chemically

synthesized or from copies of the chemically synthesized nucleic acid components.

17. The method of claim 1, wherein the synthetic genome further comprises sequences that allow production of a product of interest.

18. The method of claim 17, wherein the product of interest is an energy source.

19. The method of claim 14, wherein the entire synthetic genome further comprises sequences that allow production of a product of interest.

20. The method of paragraph 19, wherein the product of interest is an energy source.

21. The method of claim 15, wherein the synthetic cellular genome further comprises sequences that allow production of a product of interest.

22. The method of claim 21, wherein the product of interest is an energy source.

23. The method of claim 16, wherein the entire synthetic cellular genome further comprises sequences that allow production of a product of interest.

24. The method of claim 23, wherein the product of interest is an energy source.

25. The method of claim 1, further comprising rationally designing components of the synthetic genome.

26. The method of claim 1, further comprising constructing a synthetic cell, by introducing the synthetic genome into a vesicle.

27. The method of claim 15, further comprising constructing a self-replicating synthetic cell, by introducing the synthetic cellular genome into a vesicle, optionally under conditions effective for replication of the synthetic cell.

28. The method of claim 26, further comprising producing a product of interest by culturing the synthetic cell under conditions effective to produce the product.

29. The method of claim 27, further comprising producing a product of interest by culturing the self-replicating synthetic cell under conditions effective to replicate the synthetic cell and to produce the product.

30. The method of claim 1, further comprising automating the method.

31. The method of claim 1, further comprising the synthetic genome.

32. A synthetic genome.

33. The method of claim 26, further comprising the synthetic cell.

34. A synthetic cell comprising a synthetic genome.

35. A method for making a synthetic cell, comprising:
removing part or all of a resident genome from a microorganism; and

replacing the resident genome with a synthetic genome that exhibits at least one property that is different from the resident genome.

36. The method of claim 35, further comprising the synthetic cell.

37. A synthetic cell comprising:

a microorganism of one species from which part or all of the resident genome has been removed; and

a synthetic genome which exhibits at least one property that is different from the resident genome.

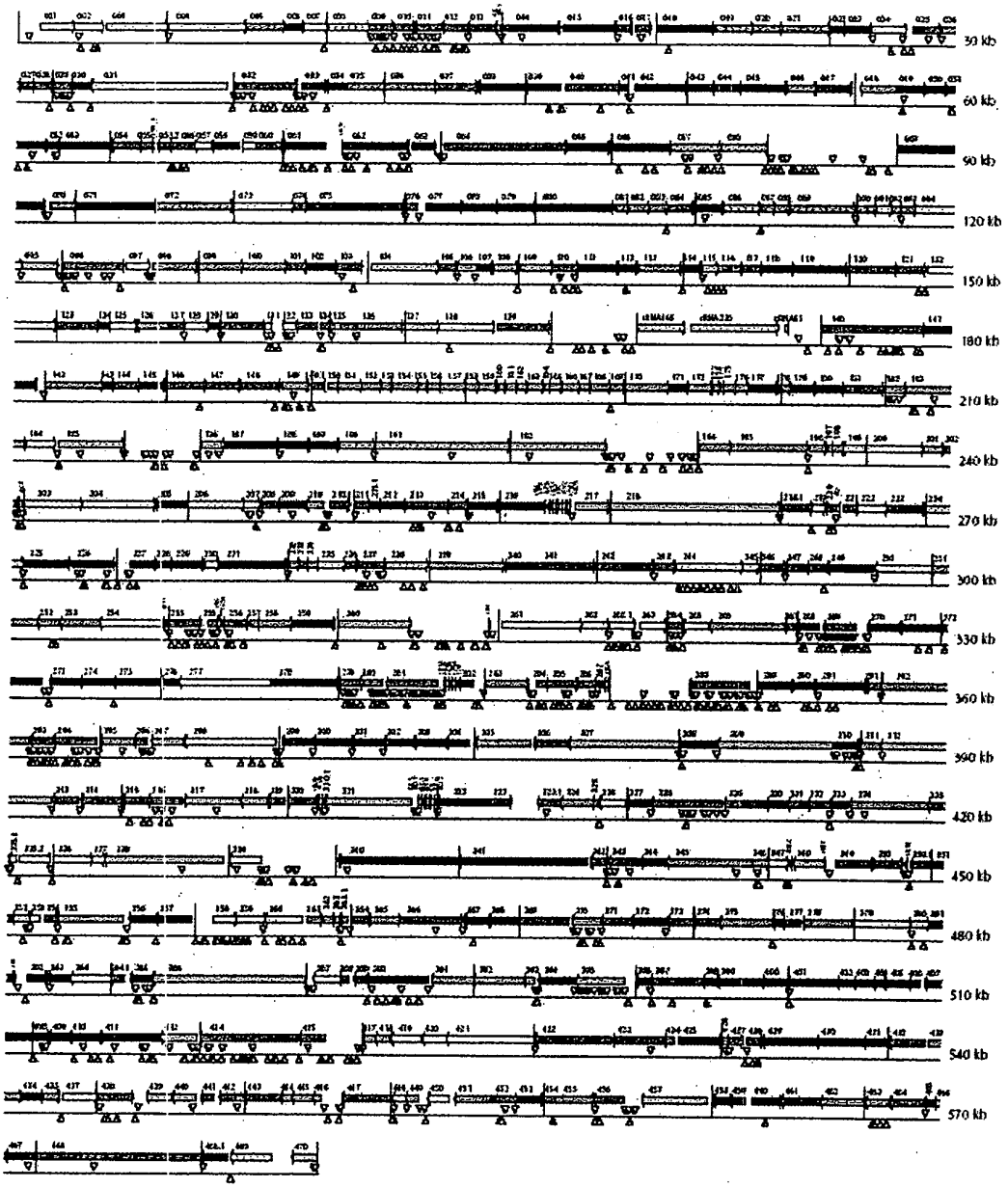
38. A method comprising:

designing a synthetic genome;

constructing the synthetic genome;

introducing the synthetic genome into a biological system; and

expressing the synthetic genome.




Map of *Mycoplasma genitalium*. The triangles mark the position of transposon insertions (upper triangles are from Smith *et al.* (1999) *Proc Natl Acad Sci USA* 87, 826-830 1999. Vertical lines delineate the borders of the "5 kb" segments.

FIGURE 1

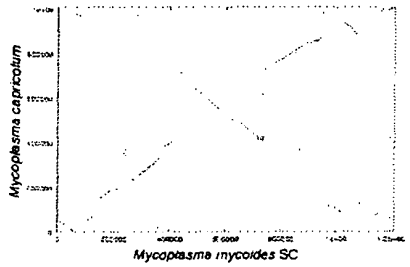
In the foregoing and in the following example, all temperatures are set forth in uncorrected degrees Celsius; and, unless otherwise indicated, all parts and percentages are by weight.

## EXAMPLES

**Example I - Mycoplasma comparative genomics to identify genes in a minimal gene set**

**Mycoplasma comparative genomics.** The 13 complete and 2 partial genome sequences currently in the inventors' dataset comprise an *in silico* laboratory

Comparisons of pairs of mycoplasma genomes using the whole genome alignment tool MUMMER shows between species that are closely related such as *M. capricolum. M. mycoides SC*, and *Mesoplasma florum*, genome rearrangements are symmetrical about an axis passing through the origins of replication and points that bisect the genome equally. The direction of transcription of rearranged genes almost always stays the same relative to the origin of replication. This phenomenon has been observed for other species bacteria but perhaps never so strikingly as with *M. capricolum* and *M. mycoides SC* (see figure below). These reciprocal crossovers suggest that any major removal of DNA from one side of the genome might need to be matched with a similar deletion from the other side so that the terminus of replication remains constant.




MUMMER comparison of the *M. capricolum* and *M. mycoides* genomes at the protein level showing the locations of orthologous genes. The cross pattern shows the conservation of gene position relative to the origin and terminus of replication for these two species.

**Core mycoplasma gene set.** The 13 complete genome sequences include 5 species from the pneumoniae branch of *Mollicutes* phylogeny, 4 from the hominis branch, three from the *Entomoplasmatales* branch and one from the *Acholeplasmatales* branch. For each of the complete genomes we used BLASTp to generate orthologs tables based on whether gene X in

one genome has a significant best BLASTp hit to gene Y in another genome, and Y is X's best hit, then those genes are called orthologs. Using these tables, we identified a core mycoplasma gene set, *i.e.* those orthologous genes common to all 13 completely sequenced mycoplasma species. Additionally, those tables identify orthologous gene sets for three of the main mycoplasma tree branches.

The core mycoplasma gene set is ~165 genes. That set can be expanded to ~200 by including those 45 genes missing only in the intracellular parasite *Phytoplasma asteris*, which obtains many of its essential metabolic products from the plant cytoplasm this species lives in. The set can be further expanded to ~310 genes by taking into account non-orthologous gene displacements that are obvious in some cases and suggested in others. Obvious examples include the 14 genes absent in either or both of the two non-glycolytic species *Ureaplasma parvum* or *Mycoplasma arthritidis*. An additional 96 genes are included in the expanded core gene set because orthologs are absent in only one of the 12 complete genome (*P. asteris* is so different from the other species it is usually ignored in this core set expansion process). Based on this 13 genome comparative genomics analysis only, we would predict that our model synthetic organism, *M. laboratorium*, would need only about 310 genes and would have a genome containing only about 372 kbp.

Given the significant evolutionary divergences of the 4 branches of mycoplasmas from each other because of their high rate of evolution and different responses to gene loss, we also determined the common gene sets for the pneumoniae, hominis and *Entomoplasmatales* groups of mycoplasmas. (see the Table below). It is instructive to consider an expanded core gene set for the 5 members of the pneumoniae group (which includes *M. genitalium*, our planned platform for *M. laboratorium* construction). If one includes only those genes present in at least 4 of the 5 group member genomes, that expanded core set is 391 protein coding genes.

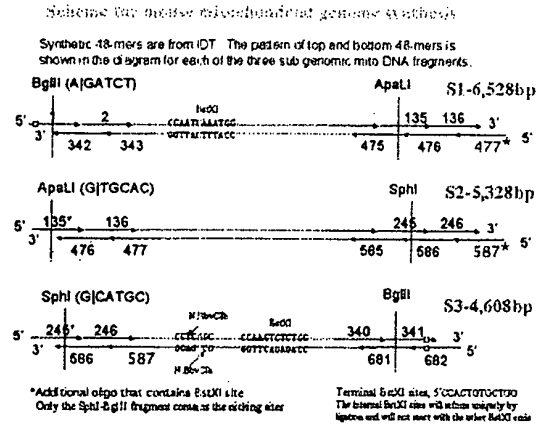
**Sizes of the orthologous gene sets shared by all and various subgroups of the mycoplasmas.**

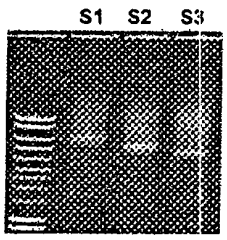
| *Orthologous gene sets* | *M. genitalium* | *M. arthritidis* | *M. capricolum* |
|---|---|---|---|
| Gene in all mycoplasmas | 165 | 168 | 152 |
| Genes in all mycoplasmas and *B. subtilis* and *C. perfringens* | 153 | 159 | 151 |
| Genes in all mycoplasmas except *Onion Yellows Phytoplasma* | 200 | 206 | 197 |
| Core mycoplasma genes lost in *Onion Yellows Phytoplasma* | 35 | 38 | 39 |
| Genes present in all 5 pneumoniae clade members | 294 | - | - |
| Genes present in all 5 pneumoniae clade members and in hominis group | 220 | 220 | - |
| Genes present in all 5 pneumoniae clade members and in mycoides | 244 | - | 236 |

| | | | |
|---|---|---|---|
| group | | | |
| Genes present in all 5 pneumoniae clade members and in *P. asteris* | 207 | - | - |
| Genes present in all 4 hominis clade members | - | 293 | - |
| Genes present in all hominis clade members and in mycoides group | - | 243 | 230 |
| Genes present in all hominis clade members and in *Phytoplasma asteris* | - | 206 | - |
| Genes present in all 3 mesoplasma/mycoides clade members | - | - | 438 |
| Genes present in all mycoides clade members and in *Phytoplasma asteris* | - | - | 230 |
| *M. capricolum* orthologs in *M. mycoides* | - | - | 715 |

## Example II - Synthesis of a mouse mitochondrial genome

The mouse mitochondrial genome is a 16,299 bp circular DNA and its sequence has been critically checked. We designed 682 48-mers to assemble it in three overlapping segments; S1 (6,528 bp), S2 (5,328 bp), and S3 (4,508 bp) as diagrammed below.




Taq ligations at 50° of the oligos for each mitochondrial fragment were subjected to 6 cycles of PCA and 20 cycles of PCR using end primers.

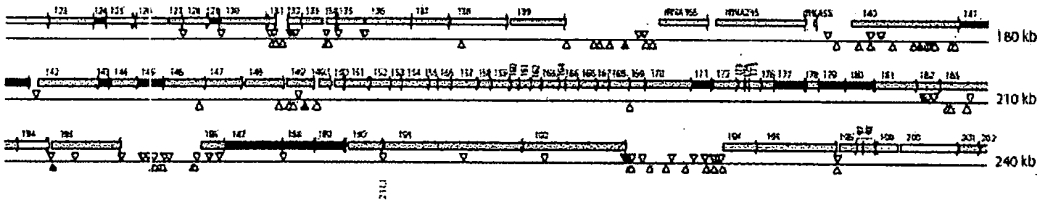
We assembled each of these three pieces by the method described in Smith *et al.* (2003), *supra*, modified to reduce the heating damage to the synthetic product. The gel above shown illustrates products from one such modified procedure that dramatically reduces the time spent at high temperature.

The synthesis of three overlapping segments comprising the entire mouse

mitochondrial genome illustrates that combinations of PCA and PCR can routinely assemble 5-6 kb segments of DNA and validates our plan to build 5 kb cassettes for the assembly of a cellular genome.

**Example III - Synthesis of 5 kb cassettes of an essential region of the *M. genitalium* genome**

5 kb cassettes are constructed to generate a synthetic copy of an essential region of the *M. genitalium* genome - the ribosomal protein genes MG149.1 through MG181. This 18.5 kb region is flanked by genes that tolerate transposon insertions (MG149 and MG182). Sets of 386 top strand and 386 bottom strand oligonucleotides, of 48 nt, were synthesized to cover this region. These nucleic acids are illustrated below:




The assembly of four overlapping segments (cassettes) comprising these oligonucleotides is performed.

Using these techniques, cassettes of, for example, 4-6 kb can be constructed that include gene sets of interest (*e.g.* a minimal genome from a unicellular microorganism), and can be "mixed and matched" with, or altered by substitutions from, *e.g.*, other species, to obtain a custom made genome, which can be introduced into a vesicle or ghost cell for testing, as described above. Synthetic cells thus constructed can be cultured under suitable conditions to determine function. After determination of functionality, the genome can be modified by substitution of cassettes, and the process repeated until a desired result is obtained.

From the foregoing description, one skilled in the art can easily ascertain the essential characteristics of this invention, and without departing from the spirit and scope thereof, can make changes and modifications of the invention to adapt it to various usage and conditions and to utilize the present invention to its fullest extent. The preceding specific embodiments are to be construed as merely illustrative, and not as limiting the scope of the invention in any way whatsoever. The entire disclosure of all applications, patents and publications cited above and in the figures are hereby incorporated by reference.