

VARIANT LIBRARIES OF THE IMMUNOLOGICAL SYNAPSE AND SYNTHESIS THEREOF

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/435,650 filed on December 16, 2016, which is incorporated herein by reference in its entirety.

BACKGROUND

[0002] The cornerstone of synthetic biology is the design, build, and test process—an iterative process that requires DNA, to be made accessible for rapid and affordable generation and optimization of these custom pathways and organisms. In the design phase, the A, C, T and G nucleotides that constitute DNA are formulated into the various gene sequences that would comprise the locus or the pathway of interest, with each sequence variant representing a specific hypothesis that will be tested. These variant gene sequences represent subsets of sequence space, a concept that originated in evolutionary biology and pertains to the totality of sequences that make up genes, genomes, transcriptome and proteome.

[0003] Many different variants are typically designed for each design-build-test cycle to enable adequate sampling of sequence space and maximize the probability of an optimized design. Though straightforward in concept, process bottlenecks around speed, throughput and quality of conventional synthesis methods dampen the pace at which this cycle advances, extending development time. The inability to sufficiently explore sequence space due to the high cost of acutely accurate DNA and the limited throughput of current synthesis technologies remains the rate-limiting step.

[0004] Beginning with the build phase, two processes are noteworthy: nucleic acid synthesis and gene synthesis. Historically, synthesis of different gene variants was accomplished through molecular cloning. While robust, this approach is not scalable. Early chemical gene synthesis efforts focused on producing a large number of nucleic acids with overlapping sequence homology. These were then pooled and subjected to multiple rounds of polymerase chain reaction (PCR), enabling concatenation of the overlapping nucleic acids into a full length double stranded gene. A number of factors hinder this method, including time-consuming and labor-intensive construction, requirement of high volumes of phosphoramidites, an expensive raw material, and production of nanomole amounts of the final product, significantly less than required for downstream steps, and a large number of separate nucleic acids required one 96 well plate to set up the synthesis of one gene.

[0005] Synthesizing nucleic acids on microarrays provided a significant increase in the throughput of gene synthesis. A large number of nucleic acids could be synthesized on the microarray surface, then cleaved off and pooled together. Each nucleic acid destined for a specific gene contains a unique barcode sequence that enabled that specific subpopulation of polynucleotides to be depooled and assembled into the gene of interest. In this phase of the process, each subpool is transferred into one well in a 96 well plate, increasing throughput to 96 genes. While this is two orders of magnitude higher in throughput than the classical method, it still does not adequately support the design, build, test cycles that require thousands of sequences at one time due to a lack of cost efficiency and slow turnaround times. Thus, there is a need for more efficient generation of variant sequence libraries.

[0006] The immune system has the ability to seek and destroy harmful cells. T cells play an important role in such a process. In the context of surveillance for cancer cells, T cell immune responses induced from the endogenous T cell repertoire can be insufficient for the removal of such cancer cells. The immunological synapse includes a repertoire of T cell proteins, antigen presenting proteins, and antigens. Genetically modified components of this system provide a means for enhancing the T cell-mediated immune responses. However, cell-intrinsic factors as well as immunosuppressive factors in the tumor microenvironment can limit the function of such gene-modified T cells. Thus, there is a need for improved compositions and methods for cancer and antiviral therapies utilizing T cell mediated signaling.

BRIEF SUMMARY

[0007] Provided herein is a nucleic acid library, wherein the nucleic acid library comprises at least 3000 variant nucleic acids, wherein each variant nucleic acid encodes for a variant gene or gene fragment for a T cell receptor (TCR) protein, and wherein a region selected for variation spans up to 1000 bases in length. Further provided herein is a nucleic acid library, wherein the variant gene or gene fragment is from a variable domain. Further provided herein is a nucleic acid library, wherein the variable domain is a variable domain of TCR alpha, TCR beta, TCR gamma, or TCR delta. Further provided herein is a nucleic acid library, wherein the variable domain is specific to a cancer antigen. Further provided herein is a nucleic acid library, wherein the cancer antigen is MAGE A3, MAGE A12, MAGE A2, MAGE A6, NY-ESO-1, or CEA. Further provided herein is a nucleic acid library, wherein the variant gene or gene fragment is from a constant domain. Further provided herein is a nucleic acid library, wherein the variant gene or gene fragment allows for generation of a TCR protein having increased specificity, avidity, affinity, stability, or expression.

[0008] Provided herein is a polynucleotide library, wherein the polynucleotide library comprises at least 3000 variant nucleic acids, wherein each polynucleotide is at least 15 bases in length, wherein each polynucleotide encodes for a variant within a variable domain of a T cell receptor (TCR) protein or fragment thereof, wherein the variable domain comprises up to 1000 bases, and wherein each variant nucleic acid comprises at least one variation at a preselected codon for an amino acid residue in an antigen contacting interface. Further provided herein is a polynucleotide library, wherein the variable domain is a variable domain of TCR alpha, TCR beta, TCR gamma, or TCR delta. Further provided herein is a polynucleotide library, wherein the antigen is a cancer antigen. Further provided herein is a polynucleotide library, wherein the cancer antigen is MAGE A3, MAGE A12, MAGE A2, MAGE A6, NY-ESO-1, or CEA. Further provided herein is a polynucleotide library, wherein each variant nucleic acid comprises a plurality of variations at preselected codons for amino acid residues in the antigen contacting interface. Further provided herein is a polynucleotide library, wherein the plurality of variations at preselected codons for amino acid residues comprises up to 100 residues. Further provided herein is a polynucleotide library, wherein the plurality of variations at preselected codons for amino acid residues comprises up to 30 residues. Further provided herein is a polynucleotide library, wherein the plurality of variations at preselected codons for amino acid residues comprises up to 5 residues. Further provided herein is a polynucleotide library, wherein the variant allows for generation of a TCR protein having increased specificity, avidity, affinity, stability, or expression. Further provided herein is a polynucleotide library, wherein each variant nucleic acid further comprises at least one variation at a preselected codon for an amino acid residue within a constant domain of the T cell receptor (TCR) protein or fragment thereof.

[0009] Provided herein is a polynucleotide library, wherein the polynucleotide library comprises at least 3000 variant polynucleotides, wherein each polynucleotide is at least 15 bases in length, wherein each polynucleotide encodes for a variant within a constant domain of a T cell receptor (TCR) protein or fragment thereof, wherein the constant domain comprises up to 1000 bases, and wherein the variant comprises at least one variation at a preselected codon for an amino acid residue in an antigen contacting interface. Further provided herein is a polynucleotide library, wherein the constant domain is a constant domain of TCR alpha, TCR beta, TCR gamma, or TCR delta. Further provided herein is a polynucleotide library, wherein the antigen is a cancer antigen. Further provided herein is a polynucleotide library, wherein the cancer antigen is MAGE A3, MAGE A12, MAGE A2, MAGE A6, NY-ESO-1, or CEA. Further provided herein is a polynucleotide library, wherein each polynucleotide comprises a plurality of variations at preselected codons for amino acid residues. Further provided herein is a polynucleotide library,

wherein the plurality of variations at preselected codons for amino acid residues comprises up to 100 residues. Further provided herein is a polynucleotide library, wherein the plurality of variations at preselected codons for amino acid residues comprises up to 30 residues. Further provided herein is a polynucleotide library, wherein the plurality of variations at preselected codons for amino acid residues comprises up to 5 residues.

[0010] Provided herein is a nucleic acid library, wherein the nucleic acid library comprising at least 10,000 variant nucleic acids, wherein each variant nucleic acid is about 500 to about 1000 bases in length, wherein each variant nucleic acid encodes for a variant of a reference sequence that encodes an exon sequence for a T cell receptor protein or fragment thereof, and wherein the reference sequence is from a subject that has a cancer. Further provided herein is a nucleic acid library, wherein the library comprising at least about 1,000,000 variant nucleic acids, wherein each variant nucleic acid is about 500 to about 1000 bases in length, wherein each variant nucleic acid encodes for a variant of a reference sequence that encodes an exon sequence for a T cell receptor protein or fragment thereof, wherein the reference sequence of the T cell receptor protein or fragment thereof is selected based on comparing T cell receptor gene sequence from a plurality of subjects for a nucleic acid sequence that is common amongst the subjects, and wherein at least a portion of the subjects are diagnosed with a cancer. Further provided herein is a nucleic acid library, the cancer is a solid cancer or a hematopoietic cancer. Further provided herein is a nucleic acid library, wherein the library comprises about 10,000,000 variant nucleic acids. Further provided herein is a nucleic acid library, wherein each variant nucleic acid is about 600 to about 900 bases in length. Further provided herein is a nucleic acid library, wherein each variant nucleic acid is in a vector sequence. Further provided herein is a nucleic acid library, wherein the vector sequence is a viral vector sequence.

[0011] Provided herein is a protein library comprising proteins encoded by a nucleic acid library described herein. Provided herein is a cell library comprising the nucleic acid library described herein.

[0012] Provided herein is a method of synthesizing a nucleic acid library, comprising: (a) providing a first set of preselected polynucleotide sequences encoding for at least 3000 variant sequences of a TCR gene or gene fragment, wherein each variant sequence comprises at least one variation at a preselected codon for an amino acid residue in an antigen contacting interface; (b) synthesizing the first set of preselected polynucleotide sequences; and (c) screening binding activity for proteins encoded by the first set of polynucleotide sequences. Further provided herein is a method of synthesizing a nucleic acid library, wherein the at least one variation is in a variable domain coding region of the TCR gene or gene fragment. Further provided herein is a method of

synthesizing a nucleic acid library, wherein the variable domain is a variable domain of TCR alpha, TCR beta, TCR gamma, or TCR delta. Further provided herein is a method of synthesizing a nucleic acid library, wherein the at least one variation is in a constant domain coding region of the TCR gene or gene fragment. Further provided herein is a method of synthesizing a nucleic acid library, wherein the constant domain is a constant domain of TCR alpha, TCR beta, TCR gamma, or TCR delta. Further provided herein is a method of synthesizing a nucleic acid library, wherein the antigen is a cancer antigen. Further provided herein is a method of synthesizing a nucleic acid library, wherein the cancer antigen is MAGE A3, MAGE A12, MAGE A2, MAGE A6, NY-ESO-1, or CEA. Further provided herein is a method of synthesizing a nucleic acid library, wherein each variant sequence comprises up to 100 variations at preselected codons for amino acid residues in the antigen contacting interface. Further provided herein is a method of synthesizing a nucleic acid library, wherein each variant sequence comprises up to 30 variations at preselected codons for amino acid residues in the antigen contacting interface. Further provided herein is a method of synthesizing a nucleic acid library, further comprising (a) providing a second set of preselected polynucleotide sequences encoding for at least one variant sequence of a TCR gene or gene fragment, where each variant sequence comprises at least one variation at a preselected codon for an amino acid residue in the TCR gene or gene fragment in a region encoding a constant domain; (b) synthesizing the second set of preselected polynucleotide sequences; and (c) screening a second activity for proteins encoded by the second set of polynucleotide sequences. Further provided herein is a method of synthesizing a nucleic acid library, wherein the second activity is cancer cell killing, protein expression, or protein stability.

[0012a] Provided herein is a method of synthesizing a nucleic acid library, comprising: a. providing a first set of preselected polynucleotide sequences encoding for at least 3000 variant sequences of a TCR gene or gene fragment, wherein each variant sequence comprises at least one variation at a preselected codon for an amino acid residue in an antigen contacting interface; b. synthesizing the first set of preselected polynucleotide sequences; c. screening binding activity for proteins encoded by the first set of polynucleotide sequences; d. providing a second set of preselected polynucleotide sequences encoding for at least one variant sequence of a TCR gene or gene fragment, where each variant sequence comprises at least one variation at a preselected codon for an amino acid residue in the TCR gene or gene fragment in a region encoding a constant domain; e. synthesizing the second set of preselected polynucleotide sequences; and f. screening a second activity for proteins encoded by the second set of polynucleotide sequences.

INCORPORATION BY REFERENCE

[0013] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] **Figures 1A-1D** depict a process workflow for the synthesis of variant biological molecules incorporating a PCR mutagenesis step.

[0015] **Figures 2A-2D** depict a process workflow for the generation of a nucleic acid comprising a nucleic acid sequence which differs from a reference nucleic acid sequence at a single predetermined codon site.

[0016] **Figures 3A-3F** depict an alternative workflow for the generation of a set of nucleic acid variants from a template nucleic acid, with each variant comprising a different nucleic acid sequence at a single codon position. Each variant nucleic acid encodes for a different amino acid at their single codon position, the different codons represented by X, Y, and Z.

[0017] **Figures 4A-4E** depict a reference amino acid sequence (**FIG. 4A**) having a number of amino acids, each residue indicated by a single circle, and variant amino acid sequences (**FIGS. 4B, 4C, 4D, & 4E**) generated using methods described herein. The reference amino acid sequence and variant sequences are encoded by nucleic acids and variants thereof generated by processes described herein.

[0018] **Figures 5A-5B** depict a reference amino acid sequence (**FIG. 5A**, SEQ ID NO: 24) and a library of variant amino acid sequences (**FIG. 5B**, SEQ ID NOS 25-31, respectively, in order of appearance), each variant comprising a single residue variant (indicated by an "X"). The reference amino acid sequence and variant sequences are encoded by nucleic acids and variants thereof generated by processes described herein.

[0019] **Figures 6A-6B** depict a reference amino acid sequence (**FIG. 6A**) and a library of variant amino acid sequences (**FIG. 6B**), each variant comprising two sites of single position variants. Each variant is indicated by differently patterned circles. The reference amino acid sequence and variant sequences are encoded by nucleic acids and variants thereof generated by processes described herein.

[0020] **Figures 7A-7B** depict a reference amino acid sequence (**FIG. 7A**) and a library of variant amino acid sequences (**FIG. 7B**), each variant comprising a stretch of amino acids (indicated by a box around the circles), each stretch having three sites of position variants (encoding for histidine) differing in sequence from the reference amino acid sequence. The reference amino acid sequence and variant sequences are encoded by nucleic acids and variants thereof generated by processes described herein.

[0021] **Figures 8A-8B** depict a reference amino acid sequence (**FIG. 8A**) and a library of variant amino acid sequences (**FIG. 8B**), each variant comprising two stretches of amino acid sequences (indicated by a box around the circles), each stretch having one site of single position variants (illustrated by the patterned circles) differing in sequence from the reference amino acid sequence. The reference amino acid sequence and variant sequences are encoded by nucleic acids and variants thereof generated by processes described herein.

[0022] **Figures 9A-9B** depict a reference amino acid sequence (**FIG. 9A**) and a library of amino acid sequence variants (**FIG. 9B**), each variant comprising a stretch of amino acids (indicated by patterned circles), each stretch having a single site of multiple position variants

differing in sequence from the reference amino acid sequence. In this illustration, 5 positions are varied where the first position has a 50/50 K/R ratio; the second position has a 50/25/25 V/L/S ratio, the third position has a 50/25/25 Y/R/D ratio, the fourth position has an equal ratio for all amino acids, and the fifth position has a 75/25 ratio for G/P. The reference amino acid sequence and variant sequences are encoded by nucleic acids and variants thereof generated by processes described herein.

[0023] **Figures 10A-10C** illustrates the structure and function of the T cell receptor (TCR) (**FIG. 10A**) and a TCR expressed on a T cell interacting with an antigen peptide on a target cell when the target peptide sequence is presented by the appropriate major histocompatibility complex (MHC-1 for cytotoxic T cells) (**FIG. 10B**). The TCR has two chains: the alpha (α) and beta (β) chains. Both chains have a constant region (c) and a variable region (v), hinge region (h), transmembrane region (tm), and cytoplasmic tail region (ct). The TCR is associated with the CD3 complex, which comprises three transmembrane signaling molecules (CD3 $\zeta\zeta$, CD3 $\delta\epsilon$ and CD3 $\gamma\epsilon$). Efficient T-cell activation also requires the simultaneous binding of the T cell co-receptor (CD8 for cytotoxic T cells). “ss” refers to a disulfide bridge. **FIG. 10C**, illustrates an exemplary workflow for engineering T cells of a subject and administering them back to the subject.

[0024] **Figure 11** depicts an exemplary number of variants produced by interchanging sections of two expression cassettes (e.g., promoters, open reading frames, and terminators) to generate a variant library of expression cassettes.

[0025] **Figure 12** presents a diagram of steps demonstrating an exemplary process workflow for gene synthesis as disclosed herein.

[0026] **Figure 13** illustrates an example of a computer system.

[0027] **Figure 14** is a block diagram illustrating an architecture of a computer system.

[0028] **Figure 15** is a diagram demonstrating a network configured to incorporate a plurality of computer systems, a plurality of cell phones and personal data assistants, and Network Attached Storage (NAS).

[0029] **Figure 16** is a block diagram of a multiprocessor computer system using a shared virtual address memory space.

[0030] **Figure 17** depicts a BioAnalyzer plot of PCR reaction products resolved by gel electrophoresis.

[0031] **Figure 18** depicts an electropherogram showing 96 sets of PCR products, each set of PCR products differing in sequence from a wild-type template nucleic acid at a single codon position, where the single codon position of each set is located at a different site in the wild-type

template nucleic acid sequence. Each set of PCR products comprises 19 variant nucleic acids, each variant encoding for a different amino acid at their single codon position.

DETAILED DESCRIPTION

[0032] The present disclosure employs, unless otherwise indicated, conventional molecular biology techniques, which are within the skill of the art. Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of ordinary skill in the art.

[0033] Definitions

[0034] Throughout this disclosure, numerical features are presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of any embodiments. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range to the tenth of the unit of the lower limit unless the context clearly dictates otherwise. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual values within that range, for example, 1.1, 2, 2.3, 5, and 5.9. This applies regardless of the breadth of the range. The upper and lower limits of these intervening ranges may independently be included in the smaller ranges, and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention, unless the context clearly dictates otherwise.

[0035] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting of any embodiment. As used herein, the singular forms “a,” “an” and “the” are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprises” and/or “comprising,” when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or groups thereof. As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0036] Unless specifically stated or obvious from context, as used herein, the term “about” in reference to a number or range of numbers is understood to mean the stated number and numbers +/- 10% thereof, or 10% below the lower listed limit and 10% above the higher listed limit for the values listed for a range.

[0037] As used herein, the terms “preselected sequence”, “predefined sequence” or “predetermined sequence” are used interchangeably. The terms mean that the sequence of the polymer is known and chosen before synthesis or assembly of the polymer. In particular, various aspects of the invention are described herein primarily with regard to the preparation of nucleic acids molecules, the sequence of the oligonucleotide or polynucleotide being known and chosen before the synthesis or assembly of the nucleic acid molecules.

[0038] Provided herein are methods and compositions for production of synthetic (i.e. de novo synthesized or chemically synthesizes) polynucleotides. The term oligonucleotide, oligo, and polynucleotide are defined to be synonymous throughout. Libraries of synthesized polynucleotides described herein may comprise a plurality of polynucleotides collectively encoding for one or more genes or gene fragments. In some instances, the polynucleotide library comprises coding or non-coding sequences. In some instances, the polynucleotide library encodes for a plurality of cDNA sequences. Reference gene sequences from which the cDNA sequences are based may contain introns, whereas cDNA sequences exclude introns. Polynucleotides described herein may encode for genes or gene fragments from an organism. Exemplary organisms include, without limitation, prokaryotes (e.g., bacteria) and eukaryotes (e.g., mice, rabbits, humans, and non-human primates). In some instances, the polynucleotide library comprises one or more polynucleotides, each of the one or more polynucleotides encoding sequences for multiple exons. Each polynucleotide within a library described herein may encode a different sequence, i.e., non-identical sequence. In some instances, each polynucleotide within a library described herein comprises at least one portion that is complementary to a sequence of another polynucleotide within the library. Polynucleotide sequences described herein may, unless stated otherwise, comprise DNA or RNA.

[0039] Provided herein are methods and compositions for production of synthetic (i.e. de novo synthesized) genes. Libraries comprising synthetic genes may be constructed by a variety of methods described in further detail elsewhere herein, such as PCA, non-PCA gene assembly methods or hierarchical gene assembly, combining (“stitching”) two or more double-stranded polynucleotides to produce larger DNA units (i.e., a chassis). Libraries of large constructs may involve polynucleotides that are at least 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 250, 300, 400, 500 kb long or longer. The large constructs can be bounded by an independently selected upper limit of about 5000, 10000, 20000 or 50000 base pairs. The synthesis of any number of polypeptide-segment encoding nucleotide sequences, including sequences encoding non-ribosomal peptides (NRPs), sequences encoding non-ribosomal peptide-synthetase (NRPS) modules and synthetic variants, polypeptide segments of other modular proteins, such as antibodies, polypeptide segments from other protein families, including non-

coding DNA or RNA, such as regulatory sequences e.g. promoters, transcription factors, enhancers, siRNA, shRNA, RNAi, miRNA, small nucleolar RNA derived from microRNA, or any functional or structural DNA or RNA unit of interest. The following are non-limiting examples of polynucleotides: coding or non-coding regions of a gene or gene fragment, intergenic DNA, loci (locus) defined from linkage analysis, exons, introns, messenger RNA (mRNA), transfer RNA, ribosomal RNA, short interfering RNA (siRNA), short-hairpin RNA (shRNA), micro-RNA (miRNA), small nucleolar RNA, ribozymes, complementary DNA (cDNA), which is a DNA representation of mRNA, usually obtained by reverse transcription of messenger RNA (mRNA) or by amplification; DNA molecules produced synthetically or by amplification, genomic DNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers. cDNA encoding for a gene or gene fragment referred to herein, may comprise at least one region encoding for exon sequence(s) without an intervening intron sequence found in the corresponding genomic sequence. Alternatively, the corresponding genomic sequence to a cDNA may lack intron sequence in the first place.

[0040] Engineering Variance in the Immunological Synapse

[0041] Provided herein are methods for the synthesis of a variant nucleic acid library, wherein each variant nucleic acid encodes for a sequence that is varied in comparison to a reference T cell receptor (TCR) protein sequence. The varied sequence may be a nucleic acid sequence that encodes for a fragment or an entire TCR protein sequence. In some cases, the varied sequence is a peptide serving as an antigen in the immunological synapse. The resulting nucleic acid library may be an oligo library, gene fragment library or gene library. In some instances, the nucleic acid library is expressed in cells to generate a variant protein library.

[0042] Referring to **FIGS. 10A-10C**, the TCR protein is a heterodimeric complex comprising an α and a β chain, each of which has an extracellular, transmembrane and intracellular region. Both chains are linked by a disulfide bond, with each receptor providing an antigen binding site. The antigen is typically presented by a major histocompatibility complex (MHC) molecule. Accessory molecules, such as CD4 for MHC class II and CD8 for MHC class I are also expressed by T cells and involved in the complex. In addition, the CD3 complex is also involved and provides a signal transduction role in the T cell.

[0043] Provided herein are methods to increase tumor cell killing mediated by T cells by engineering T cells. For example, tissue samples can be taken from subjects to obtain sequencing information on TCRs present in the subject's T cells. Tissue samples may be obtained by

collecting blood, peripheral blood mononuclear cells, or collecting isolated tumor infiltrating lymphocytes from excised tumor tissue. The subject may have cancer or be cancer free.

[0044] When obtaining samples from a plurality of subjects, the subjects are screened for having a common HLA donor profile. For example, the HLA may be an HLA-A, HLA-B, or HLA-C allele. Exemplary alleles are provided in **Table 1** below.

Table 1. Exemplary alleles

HLA-A allele:	HLA-C allele:	HLA-B allele:
010101g	0102	0702
02010101g	0202/0210	0705g
0205	0302	0714
0206	0303	0801g
0217	030401/030403	1301
03010101g	030402	1302
0302	0332 (V)	1401
0325 (V)c	04010101g	1402
1101	0403	15010101g
2301g	0501g	1503
24020101g	0602	1506
240301g	070101g	1510
2501	0702	1517
2601	070401g	1518
29010101g	0713	1534
2902	0732N (V)	180101g
3001	0801	1803
3002	0802	2702
3004	0803	270502g
3101	1202	2707
3201	1203	350101g
3208	1402	3502
3301	1502	3503
3303	1505	3508
3402	1601	3512
3601	1602	3701
6601	1604	3801
680101	1701g	3901
680102g		3906
6802		400101g
7401g		400201g

[0045] Various processes are described herein to identify nucleic acid regions of interest for the selection for generation of a variant library having preselected diversity within that region. For example, T cells may be obtained from a plurality of subjects. At least a portion of the subjects

may have cancer, or none of the subjects may be diagnosed as having cancer. The T cells are then sequenced, and a common sequence is identified in gene sequence encoding for TCR protein amongst the subjects. In some instances, 1, 2, 3, 4, 5 or more separate stretches of common sequences are identified in the gene encoding for a TCR amongst the subjects. The common sequence(s) are typically exon sequence in TCR gene sequence. Intron sequence may also be analyzed amongst the subjects. Alternatively, T cells are obtained from a single subject before and after being diagnosed with cancer. The genetic profile of the TCR genes in the T cells is compared in a similar manner in order to identified common sequences for variation. Alternatively, T cells are obtained from a single subject, either diagnosed as having cancer or cancer free, and the genetic profile of the TCR genes in the T cells is compared against a sequence database. The sequence database may comprise TCR genetic profiles from similarly HLA mapped subjects.

[0046] A common sequence is then varied using methods described herein, including (i) de novo synthesis of a variant library of primer nucleic acids followed by PCR mutagenesis, (ii) de novo synthesis of the entire variant version of the common sequence, or (iii) de novo synthesis of multiple fragments of the variant version of the common sequence for annealing and assembly including polymerase chain assembly.

[0047] Common sequence(s) described herein may be about 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 or more bases in length. In some instances, the common sequence(s) is at least about 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 or more bases in length. In some instances(s), the common sequence is about 100 to about 1000, about 100 to about 2000, or about 400 to about 900 bases in length. In some instances, multiple common sequences are identified. In some instances, about 100, 250, 500, 1000 or more common TCR coding gene sequences amongst the subjects are selected for variation. Variation includes the designing of at least about 500, 1000, 2000, 5000, 10000, 20000, 50000, 100000 or more variants for each of the common TCR sequences selected. The total variant library may result in 10^6 , 10^8 , 10^{10} , 10^{11} , 10^{12} , 10^{13} or more different nucleic acids.

[0048] Variant TCR sequence libraries described herein may comprise common sequence from any region of a TCR nucleotide sequence encoding for a gene or gene fragment. Exemplary regions of TCR nucleotide sequence for gene or gene fragment variation include, without limitation: variable region of the α chain, variable region of the β chain, constant region of the α chain, constant region of the β chain, hinge region, transmembrane region, or cytoplasmic tail region. In some instances, the variant TCR sequence library comprises common sequence for variation, wherein nucleic acids encoding for residues outside of the complementarity determining region of the variable domains are selected for variation. In some instances, the variant TCR

sequence library comprises common sequence for variation, wherein nucleic acids encoding for residues within the complementarity determining region of the variable domains are selected for variation.

[0049] Sequence encoding for the TCR α and beta chains each comprise separate variable (V), diversity (D), joining (J) gene segments, and constant (C) genes. The TCR α locus (chromosome 14) consists of 70–80 V α gene segments, each preceded by an exon encoding the leader sequence (L). A cluster of 61 J α gene segments is located a considerable distance from the V α gene segments. The J α gene segments are followed by a single C gene, which contains separate exons for the constant and hinge domains and a single exon encoding the transmembrane and cytoplasmic regions (not shown). The TCR β locus (chromosome 7) has a different organization, with a cluster of 52 functional V β gene segments located distantly from two separate clusters each containing a single D gene segment, together with six or seven J gene segments and a single C gene. Each TCR β C gene has separate exons encoding the constant domain, the hinge, the transmembrane region, and the cytoplasmic region (not shown). The TCR α locus is interrupted between the J and V gene segments by another T-cell receptor locus—the TCR δ locus. In some embodiments, the common sequence selected for variation includes one or more of the V α gene segment, J α gene segment, C α gene segment, V β gene segment, J β gene segment or C β gene segment.

[0050] Exemplary variable genes of a TCR α chain include, but are not limited to, TRAV1-1 (TRAV11, TCRAV1S1, TCRAV7S1), TRAV1-2, TRAV2, TRAV3, TRAV4 (TCRAV20S1, TCRAV4S1), TRAV5, TRAV6 (TCRAV5S1), TRAV7, TRAV8-1, TRAV8-2, TRAV8-3, TRAV8-4, TRAV8-5 (TRAV8S, TCRAV8S5), TRAV8-6, TRAV8-7, TRAV9-1, TRAV9-2, TRAV10 (TCRAV10S1, TCRAV24S1), TRAV11, TRAV12-1, TRAV12-2, TRAV12-3 (TRAV123, TCRAV2S2, TCRAV12S3), TRAV13-1, TRAV13-2, TRAV14DV4 (TRAV14/DV4, TCRAV6S1-hDV104S1, hADV14S1), TRAV15 (TCRAV15S1), TRAV16 (TCRAV16S1, TCRAV9S1), TRAV17 (TCRAV17S1, TCRAV3S1), TRAV18 (TCRAV18S1), TRAV19 (TCRAV12S1, TCRAV19S1), TRAV20, TRAV21, TRAV22 (TCRAV13S1, TCRAV22S1), TRAV23DV6 (TRAV23/DV6, TCRAV17S1), TRAV24, TRAV25 (TCRAV25S1, TCRAV32S1), TRAV26-1, TRAV26-2 (TRAV262, TCRAV4S1, TCRAV26S2), TRAV27, TRAV28, TRAV29DV5 (TRAV29/DV5), TRAV30, TRAV31 (TCRAV31S1), TRAV32, TRAV33 (C14orf12), TRAV34 (TCRAV26S1, TCRAV34S1), TRAV35 (TCRAV25S1, TCRAV35S1), TRAV36DV7 (TRAV36/DV7), TRAV37, TRAV38-1 (TRAV381, TCRAV14S2, TCRAV38S1), TRAV38-2DV8 (TRAV38-2/DV8, TRAV382DV8, TCRAV14S1, hADV38S2), TRAV39 (TCRAV27S1, TCRAV39S1), TRAV40 (TCRAV31S1, TCRAV40S1), and TRAV41

(TCRAV19S1, TCRAV41S1). Exemplary gene sequences for variable genes of a TCR α chain are listed in **Table 2**.

[0051] Exemplary variable genes of a TCR beta chain include, but are not limited to, TRBV1 (TCRBV27S1P, TCRBV1S1P) TRBV2 (TCRBV22S1A2N1T, TCRBV2S1), TRBV3-1 (TRBV31, TCRBV3S1, TCRBV9S1A1T), TRBV3-2 (TRBV32, TCRBV3S2, TCRBV9S2A2PT), TRBV4-1 (TRBV41, TCRBV4S1, TCRBV7S1A1N2T, BV07S1J2.7), TRBV4-2 (TRBV42, TCRBV4S2, TCRBV7S3A2T), TRBV4-3 (TRBV43, TCRBV4S3, TCRBV7S2A1N4T), TRBV5-1 (TRBV51, TCRBV5S1, TCRBV5S1A1T), TRBV5-2 (TRBV52, TCRBV31S1, TCRBV5S2P), TRBV5-3 (TRBV53, TCRBV5S3, TCRBV5S5P), TRBV5-4 (TRBV54, TCRBV5S4, TCRBV5S6A3N2T), TRBV5-5 (TRBV55, TCRBV5S3A2T, TCRBV5S5), TRBV5-6 (TRBV56, TCRBV5S2, TCRBV5S6), TRBV5-7 (TRBV57, TCRBV5S7, TCRBV5S7P), TRBV5-8 (TRBV58, TCRBV5S4A2T, TCRBV5S8), TRBV6-1 (TRBV61, TCRBV13S3, TCRBV6S1), TRBV6-2 (TRBV62, TCRBV13S2A1T, TCRBV6S2), TRBV6-3 (TRBV63, TCRBV13S9/13S2A1T, TCRBV6S3), TRBV6-4 (TRBV64, TCRBV13S5, TCRBV6S4), TRBV6-5 (TRBV65, TCRBV13S1, TCRBV6S5), TRBV6-6 (TRBV66, TCRBV13S6A2T, TCRBV6S6), TRBV6-7 (TRBV67, TCRBV13S8P, TCRBV6S7), TRBV6-8 (TRBV68, TCRBV13S7P, TCRBV6S8), TRBV6-9 (TRBV69, TCRBV13S4, TCRBV6S9), TRBV7-1 (TRBV71, TCRBV6S7P, TCRBV7S1), TRBV7-2 (TRBV72, TCRBV6S5A1N1, TCRBV7S2), TRBV7-3 (TRBV73, TCRBV6S1A1N1, TCRBV7S3), TRBV7-4 (TRBV74, TCRBV6S8A2T, TCRBV7S4), TRBV7-5 (TRBV75, TCRBV6S9P, TCRBV7S5), TRBV7-6 (TRBV76, TCRBV6S3A1N1T, TCRBV7S6), TRBV7-7 (TRBV77, TCRBV6S6A2T, TCRBV7S7), TRBV7-8 (TRBV78, TCRBV6S2A1N1T, TCRBV7S8), TRBV7-9 (TRBV79, TCRBV6S4A1, TCRBV7S9), TRBV8-1 (TRBV81, TCRBV30S1P, TCRBV8S1P), TRBV8-2 (TRBV82, TCRBV32S1P), TRBV9 (TCRBV1S1A1N1, TCRBV9S1), TRBV10-1 (TRBV101, TCRBV10S1, TCRBV12S2A1T, TCRBV12S2), TRBV10-2 (TRBV102, TCRBV10S2, TCRBV12S3), TRBV10-3 (TRBV103, TCRBV10S3, TCRBV12S1A1N2), TRBV11-1 (TRBV111, TCRBV11S1, TCRBV21S1), TRBV11-2 (TRBV112, TCRBV11S2, TCRBV21S3A2N2T), TRBV11-3 (TRBV113, TCRBV11S3, TCRBV21S2A2), TRBV12-1 (TRBV121, TCRBV12S1, TCRBV8S4P), TRBV12-2 (TRBV122, TCRBV12S2, TCRBV8S5P), TRBV12-3 (TRBV123, TCRBV12S3, TCRBV8S1), TRBV12-4 (TRBV124, TCRBV12S4, TCRBV8S2A1T), TRBV12-5 (TRBV125, TCRBV12S5, TCRBV8S3), TRBV13 (TCRBV13S1, TCRBV23S1A2T), TRBV14 (TCRBV14S1, TCRBV16S1A1N1), TRBV15 (TCRBV15S1, TCRBV24S1A3T), TRBV16 (TCRBV16S1, TCRBV25S1A2PT), TRBV17 (TCRBV17S1, TCRBV26S1P), TRBV18 (TCRBV18S1), TRBV19 (TCRBV17S1A1T, TCRBV19S1), TRBV20-1 (TRBV201, TCRBV20S1, TCRBV2S1), TRBV21-1 (TRBV211,

TCRBV10S1P, TCRBV21S1), TRBV22-1 (TCRBV22S1, TCRBV29S1P), TRBV23-1 (TCRBV231, TCRBV19S1P, TCRBV23S1), TRBV24-1 (TRBV241, TCRBV15S1, TCRBV24S1), TRBV25-1 (TRBV251, TCRBV11S1A1T, TCRBV25S1), TRBV26 (TCRBV26S1, TCRBV28S1P), TRBV27 (TCRBV14S1, TCRBV27S1), TRBV28 (TCRBV28S1, TCRBV3S1), TRBV29-1 (TRBV291, TCRBV29S1, TCRBV4S1A1T), and TRBV30 (TCRBV20S1A1N2, TCRBV30S1). Exemplary gene sequences for variable genes of a TCR beta chain are listed in **Table 2**.

Table 2. T cell receptor variable alpha chain and T cell receptor variable beta chain sequences

SEQ ID NO	Gene Name	Accession Number	Sequence
32	TRAV1-1	NC_000014.9	ATGTGGGGAGCTTTCCTTCTCTATGTTTCCATGAAGATGGG AGGTGAGTCTCAATCTAATAGTAAATGCTGCTAGGAATTT TCAAAACAATTCCTTTCAGCTAAATTATTGCAAATTTTGA CATTTGTAATGAGAGTATTCCTGAATATGCATTTTCTAA CGTGGTGCTAATTGTCCTCCTGTTACTATTGCTGCTGCTGT TACTGCAACCATTTATTTTCAGTCTAAGAAATTCTCCCATCA ATGGCAGTTCTTTTGTGACCACATGGAAGCATCATTTAAA AAATTATTCCAATAGTTTTTGGAGGAAACATCATTTTTAAT AATGATGGGGCTTCTGGGGGTGCTGCCCTAGTAACAATCA TGTATCTTGTTCATAGGCACTGCAGGACAAAGCCTTGAGCA GCCCTCTGAAGTGACAGCTGTGGAAGGAGCCATTGTCCAG ATAAACTGCACGTACCAGACATCTGGGTTTTATGGGCTGT CCTGGTACCAGCAACATGATGGCGGAGCACCCACATTTCT TTCTTACAATGCTCTGGATGGTTTGGAGGAGACAGGTCGT TTTTCTTCATTCCTTAGTCGCTCTGATAGTTATGGTTACCTC CTTCTACAGGAGCTCCAGATGAAAGACTCTGCCTCTTACT TCTGCGCTGTGAGAGA
33	TRAV4	NG_001332.3	ATGAGGCAAGTGGCGAGAGTGATCGTGTTCCCTGACCCTGA GTGAGTTATTTTGGGATGAAGAGGAATGGGATCTGGGCCT GATGATGCTGGAAAGGAATCTGGAACCTTGCCTGCTAGCA GTTGCTCTTTATCCAAGATGTAGAGGGATAGCTTCAGGGT TCCATTTTTCTCCAGGCAGCTCCTGAGCATTATGTGAGAT GTGTCTCAGGGCAGCAAAGATTTTCGGGGTTTCTAATCCC TAGGGTCTATCCATGGGAGGGAGATAGTAAACCTGACACT TCTGTGTATGCCTGGGATAAATATTTTTGTACTGAAACATA AATGAGAAGTAAATTGTATGTACATGTATATGGCAGGGAA GTAGAGCGTGGGGATGGATTTGACTGTGAGAGGGAACCG TGGGGCTACTGAGTGGGACCGTCTAACTTACCTTGCTTTC AGGATAGGGGCTAGGATTGTGTTTTACTCCCATAGGTACT TTGAGCCTTGCTAAGACCACCCAGCCCATCTCCATGGACT CATATGAAGGACAAGAAGTGAACATAACCTGTAGCCACA ACAACATTGCTACAAATGATTATATCACGTGGTACCAACA GTTTCCCAGCCAAGGACCACGATTTATTATTCAAGGATAC AAGACAAAAGTTACAAACGAAGTGGCCTCCCTGTTTATCC CTGCCGACAGAAAGTCCAGCACTCTGAGCCTGCCCCGGGT

			TTCCCTGAGCGACACTGCTGTGTACTACTGCCTCGTGGGT GACA
34	TRAV8- 1	NG_0013 32.3	ATGCTCCTGTTGCTCATACCAGTGCTGGGGATGATTTTTGC CCTGAGTGAGTAACATTCTATTATGGTCTCTAGTTCCACAG AAGTAACTGTTTTCTGATTCAAATCTTAGTAGAAATACTTT TCATAGACAAGTCTGCACTATTTTCACTGATACAACATTG ATTTTTTTCAGGAGATGCCAGAGCCCAGTCTGTGAGCCAGC ATAACCACCACGTAATTCTCTCTGAAGCAGCCTCACTGGA GTTGGGATGCAACTATTCCTATGGTGGAAGTGTAAATCTCT TCTGGTATGTCCAGTACCCTGGTCAACACCTTCAGCTTCTC CTCAAGTACTTTTCAGGGGATCCACTGGTTAAAGGCATCA AGGGCTTTGAGGCTGAATTTATAAAGAGTAAATTCTCCTT TAATCTGAGGAAACCCTCTGTGCACTGGAGTGACACAGCT GAGTACTTCTGTGCCGTGAATGC
35	TRBV2	NG_0013 33.2	ATGGATACCTGGCTCGTATGCTGGGCAATTTTTAGTCTCTT GAAAGCAGGTCGATGCTTAGACTCTAGGAAATTCTTGCTT TGAACCTACCTAAGACAATTCTAAACCATTCTCTTAATCTT CTTCTTTTCTCACAGGACTCACAGAACCTGAAGTCACCCA GACTCCCAGCCATCAGGTCACACAGATGGGACAGGAAGT GATCTTGCGCTGTGTCCCCATCTCTAATCACTTATACTTCT ATTGGTACAGACAAATCTTGGGGCAGAAAGTCGAGTTTCT GGTTTCCTTTTATAATAATGAAATCTCAGAGAAGTCTGAA ATATTCGATGATCAATTCTCAGTTGAAAGGCCTGATGGAT CAAATTTCACTCTGAAGATCCGGTCCACAAAGCTGGAGGA CTCAGCCATGTACTTCTGTGCCAGCAGTGAAGC
36	TRBV3- 1	NG_0013 33.2	ATGGGCTGCAGGCTCCTCTGCTGTGTGGTCTTCTGCCTCCT CCAAGCAGGTGAGTCCCGGGCCAGGTGACATGATCCTAT TGGAGTCCCTAAGCCTTTTACCATGACAACAACAGCAGG CCGTCTCCTAGGATTTGCCTGAATTCTGCTTCTTTCCTTTG CAGGTCCCTTGGACACAGCTGTTTCCCAGACTCCAAAATA CCTGGTCACACAGATGGGAAACGACAAGTCCATTAAATGT GAACAAAATCTGGGCCATGATACTATGTATTGGTATAAAC AGGACTCTAAGAAATTTCTGAAGATAATGTTTAGCTACAA TAATAAGGAGCTCATTATAAATGAAACAGTTCCAAATCGC TTCTCACCTAAATCTCCAGACAAAGCTCACTTAAATCTTCA CATCAATTCCCTGGAGCTTGGTGACTCTGCTGTGTATTCT GTGCCAGCAGCCAAGA
37	TRBV9	NG_0013 33.2	ATGGGCTTCAGGCTCCTCTGCTGTGTGGCCTTTTGTCTCCT GGGAGCAGGTGAGTCCTGGGCACAACCTTGAAAGTCTCCG ATCTTCATTTCTTGTCCCTGAAATGCATGTGGGCCAACGAT GGCTTCAGCAGGAGGCTTTCTTCTGTGCCTTATGGTTAACT TTTGTCTTCTGACACACAGGCCAGTGGATTCTGGAGTCA CACAAACCCCAAAGCACCTGATCACAGCAACTGGACAGC GAGTGACGCTGAGATGCTCCCCTAGGTCTGGAGACCTCTC TGTGTACTGGTACCAACAGAGCCTGGACCAGGGCCTCCAG TTCCTCATTCACTATTATAATGGAGAAGAGAGAGCAAAAG GAAACATTCTTGAACGATTCTCCGCACAACAGTTCCCTGA CTTGCACTCTGAACTAAACCTGAGCTCTCTGGAGCTGGGG GACTCAGCTTTGTATTTCTGTGCCAGCAGCGTAG
38	TRBV15	NG_0013 33.2	ATGGGTCTGGGCTTCTCCACTGGATGGCCCTTTGTCTCCT TGGAACAGGTGAGTACTGGGCAGAAAGGAAATCTTTGAG

			CAAAGCTATCTTGTCTCCTCAGTCTGCACCTTTCATTTCACAGC AGTAACACTGTTCTCCTTAAGTCTGACTCCAAATTTGTCTT CTTTCTCTACAGGTCATGGGGATGCCATGGTCATCCAGAA CCCAAGATACCAGGTTACCCAGTTTGGAAAGCCAGTGACC CTGAGTTGTTCTCAGACTTTGAACCATAACGTCATGTACTG GTACCAGCAGAAGTCAAGTCAGGCCCAAAGCTGCTGTTC CACTACTATGACAAAGATTTTAACAATGAAGCAGACACCC CTGATAACTTCCAATCCAGGAGGCCGAACACTTCTTTCTG CTTTCTTGACATCCGCTCACCAGGCCTGGGGGACACAGCC ATGTACCTGTGTGCCACCAGCAGAGA
--	--	--	---

[0052] A common sequence selected for variation in a library described herein may include one or more of the J α gene segment, C α gene segment, J β gene segment, or C β gene segment.

Exemplary J α genes include, but are not limited to, TRAJ1, TRAJ2, TRAJ3, TRAJ4, TRAJ5, TRAJ6, TRAJ7, TRAJ8, TRAJ9, TRAJ10, TRAJ11, TRAJ12, TRAJ13, TRAJ14, TRAJ15, TRAJ16, TRAJ17, TRAJ18, TRAJ19, TRAJ20, TRAJ21, TRAJ22, TRAJ23, TRAJ24, TRAJ25, TRAJ26, TRAJ27, TRAJ28, TRAJ29, TRAJ30, TRAJ31, TRAJ32, TRAJ33, TRAJ34, TRAJ35, TRAJ36, TRAJ37, TRAJ38, TRAJ39, TRAJ40, TRAJ41, TRAJ42, TRAJ43, TRAJ44, TRAJ45, TRAJ46, TRAJ47, TRAJ48, TRAJ49, TRAJ50, TRAJ51, TRAJ52, TRAJ53, TRAJ54, TRAJ55, TRAJ56, TRAJ57, TRAJ58, TRAJ59, TRAJ60, and TRAJ61. In some instances, the C α gene is TRAC. Exemplary J β genes include, but are not limited to, TRBJ1-1 (TRBJ11, TCRBJ1S1), TRBJ1-2 (TRBJ12, TCRBJ1S2), TRBJ1-3 (TRBJ13, TCRBJ1S3), TRBJ1-4 (TRBJ14, TCRBJ1S4), TRBJ1-5 (TRBJ15, TCRBJ1S5), TRBJ1-6 (TRBJ16, TCRBJ1S6), TRBJ2-1 (TRBJ21, TCRBJ2S1), TRBJ2-2 (TRBJ22, TCRBJ2S2), TRBJ2-2P (TRBJ22P), TRBJ2-3 (TRBJ23, TCRBJ2S3), TRBJ2-4 (TRBJ24, TCRBJ2S4), TRBJ2-5 (TRBJ25, TCRBJ2S5), TRBJ2-6 (TRBJ26, TCRBJ2S6), and TRBJ2-7 (TRBJ27, TCRBJ2S7). In some instances, the C β gene is TRBC1 (TCRBC1, BV05S1J2.2) or TRBC2.

[0053] A variant TCR library described herein may comprise variants for sequences of TCR delta (TCR δ) locus. In some instances, the sequence encoding for the TCR δ chain comprises a variable (V), diversity (D), joining (J) gene segments, or constant (C) gene. Exemplary genes include, but are not limited to, TRD (TCRDV1), TRDV1 (hDV101S1), TRDV2 (hDV102S1, MGC117421), TRDV3 (hDV103S1), TRDD1, TRDD2, TRDD3, TRDJ1, TRDJ2, TRDJ3, TRDJ4, and TRDC.

[0054] A variant TCR library described herein may comprise variants for sequences of TCR gamma (TCR γ) locus. In some instances, the sequence encoding for the TCR γ chain comprises a variable (V), diversity (D), joining (J) gene segments, or constant (C) gene. Exemplary genes include, but are not limited to, TRG, TRGV1 (V1S1P), TRGV2 (VIS2), TRGV3 (V1S3), TRGV4

(V1S4), TRGV5 (V1S5), TRGV5P (V1S5P), TRGV6 (V1S5P, TCRGV5P), TRGV7 (V1S7P), TRGV8 (V1S8), TRGV9 (V2), TRGV10 (V3P), TRGV11 (V4P), TRGVA (V5P), TRGVB (V6P), TRGJ1 (J1), TRGJ2 (J2), TRGJP (JP), TRGJP1 (JP1), TRGJP2 (JP2), TRGC1 (C1), TRGC2 (TRGC2(2X), TRGC2(3X))

[0055] Provided herein are TCR libraries comprising nucleic acids encoding for a variant TCR, wherein each nucleic acid encodes for a predetermined variant of at least one predetermined reference nucleic acid sequence. In some cases, the predetermined reference sequence is a nucleic acid sequence encoding for a protein, and the variant library comprises sequences encoding for variation of at least a single codon such that a plurality of different variants of a single residue in the subsequent protein encoded by the synthesized nucleic acid are generated by standard translation processes. In some instances, the variant TCR library comprises varied nucleic acids collectively encoding variations at multiple positions. In some instances, the variant library comprises sequences encoding for variation of at least a single codon of a TCR α , TCR β , TCR δ , or TCR γ chain. In some instances, the variant library comprises sequences encoding for variation of at least a single codon of a variable domain of TCR β chain. In some instances, the variant library comprises sequences encoding for variation of at least a single codon of a variable domain of TCR α and TCR β chain. In some instances, the variant library comprises sequences encoding for variation of multiple codons of a variable domain of a TCR α chain, TCR β chain, or combinations thereof. An exemplary number of codons for variation include, but are not limited to, at least or about 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 225, 250, 275, 300, or more than 300 codons.

[0056] In some instances, variant TCR libraries comprise variation in one or more exons of a variable gene of a TCR alpha chain. In some instances, variant TCR libraries comprise variation in one or more exons of a variable gene of a TCR beta chain. For example, at least or about 10, 15, 20, 25, 30, 40, 50, 75, 100, 150, 200, or more than 200 nucleic acids of one or more exons of a variable gene of a TCR α chain is varied. In some instances, at least or about 10, 15, 20, 25, 30, 40, 50, 75, 100, 150, 200, or more than 200 nucleic acids of one or more exons of a variable gene of a TCR beta chain is varied. Exemplary exon sequences of variable genes of a TCR alpha chain and a TCR beta chain are seen in **Table 3**.

Table 3. Exon sequences of variable genes of a TCR alpha chain and a TCR beta chain

SEQ ID NO	Gene Name	Accession Number	Sequence

39	TRAV1-1	X04939.1	ATGTGGGGAGCTTTCCTTCTCTATGTTTCCATGAAGATGGG AGGCACTGCAGGACAAAGCCTTGAGCAGCCCTCTGAAGT GACAGCTGTGGAAGGAGCCATTGTCCAGATAAACTGCAC GTACCAGACATCTGGGTTTTATGGGCTGTCCTGGTACCAG CAACATGATGGCGGAGCACCCACATTTCTTTCTTACAATG GTCTGGATGGTTTGGAGGAGACAGGTCGTTTTTCTTCATTC CTTAGTCGCTCTGATAGTTATGGTTACCTCCTTCTACAGGA GCTCCAGATGAAAGACTCTGCCTCTTACTTCTGCGCTGTTT TTAACCAGGCAGGAAGTCTCTGATCTTTGGGAAGGGAAC CACCTTATCAGTGAGTTCCAATATCCAGAACCCT
40	TRAV8-1	X04949.1	ATGCTCCTGTTGCTCATACCAGTGCTGGGGATGATTTTTGC CCTGAGAGATGCCAGAGCCCAGTCTGTGAGCCAGCATAAC CACCACGTAATTCTCTCTGAAGCAGCCTCACTGGAGTTGG GATGCAACTATTCCTATGGTGGAAGTGTAAATCTCTTCTGG TATGTCCAGTACCCTGGTCAACACCTTCAGCTTCTCCTCAA GTACTTTTCAGGGGATCCACTGGTTAAAGGCATCAAGGGC TTTGAGGCTGAATTTATAAAGAGTAAATTCTCCTTTAATCT GAGGAAACCCTCTGTGCAGTGGAGTGACACAGCTGAGTA CTTCTGTGCCGTGAATGAATACGACTACAAGCTCAGCTTT GGAGCCGGAACCACAGTAACTGTAAGAGCAAATATCCAG AACCC
41	TRAV13-1	X04944.1	ATGACATCCATTCGAGCTGTATTTATATTCCTGTGGCTGCA GCTGGACTTGGTGAATGGAGAGAATGTGGAGCAGCATCCT TCAACCCTGAGTGTCCAGGAGGGAGACAGCGCTGTTATCA AGTGTACTTATTCAGACAGTGCCTCAAACCTACTTCCCTTGG TATAAGCAAGAACTTGGA AAAAGACCTCAGCTTATTATAG ACATTCGTTCAAATGTGGGCGAAAAGAAAGACCAACGAA TTGCTGTACATTGAACAAGACAGCCAAACATTTCTCCCT GCACATCACAGAGACCCAACCTGAAGACTCGGCTGTCTAC TTCTGTGCAGCAAAGCGCAAGGCCTCTAGCAACACAGGCA AACTAATCTTTGGGCAAGGGACAACCTTTACAAGTAAAACC AGATATCCAGAACCCT
42	TRBV2	M62379.1	GAACCTGAAGTCACCCAGACTCCCAGCCATCAGGTCACAC AGATGGGACAGGAAGTGATCTTGCAGTGTGTCCCCATCTC TAATCACTTATACTTCTATTGGTACAGACAAATCTTGGGG CAGAAAGTCGAGTTTCTGGTTTCCTTTTATAATAATGAAAT CTCAGAGAAGTCTGAAATATTTCGATGATCAATTCTCAGTT GAAAGGCCTGATGGATCAAATTTCACTCTGAAGATCCGGT CCACAAAGCTGGAGGACTCAGCCATGTACTTCTGTGCCAG CAGT
43	TRBV3-1	L06889	ATGGGCTTCAGGCTCCTCTGCTGCGTCGCCTTCTGCCTCCT CCAAGCAGGTCCCTTGGACACAGCTGTTTCCCAGACTCCA AAATACCTGGTCACACAGATGGGAAACGACAAGTCCATT AAATGTGAACAAAATCTGGGCCATGATACTATGTATTGGT ATAAACAGGACTCTAAGAAATTTCTGAAGATAATGTTTAG CTACAATAACAAGGAGATCATTATAAATGAAACAGTTCCA AATCGATTCTCACCTAAATCTCCAGACAAAGCTAAATTAA ATCTTCACATCAATTCCCTGGAGCTTGGTGACTCTGCTGTG TATTTCTGTGCCAGC

44	TRBV15	M62376	ACCCCGACCTCTCAACCCAGGAATCAGAGCCTGAGACAG ACAGATGCTTCATTCTGTATGGGGTGGTATTCCTGCCATG GGTCCTGGGCTTCTCCACTGGATGGCCCTTTGTCTCCTTGG AACAGGTCATGGGGATGCCATGGTCATCCAGAACCCAAG ATACCGGGTTACCCAGTTTGGAAAGCCAGTGACCCTGAGT TGTTCTCAGACTTTGAACCATAACGTCATGTACTGGTACC AGCAGAAGTCAAGTCAGGCCCCAAAGCTGCTGTTCCACTA CTATAACAAAGATTTTAACAATGAAGCAGACACCCCTGAT AACTTCCAATCCAGGAGGCCGAACACTTCTTTCTGCTTTCT AGACATCCGCTCACCAGGCCTGGGGGACGCAGCCATGTAC CAGTGTGCCACCAGC
----	--------	--------	---

[0057] Libraries comprising nucleic acids encoding for variant TCRs as described herein comprise various lengths of amino acids when translated. In some instances, the length of each of the amino acid fragments or average length of the amino acid synthesized may be at least or about 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, or more than 150 amino acids. In some instances, the length of the amino acid is in a range of about 15 to 150, 20 to 145, 25 to 140, 30 to 135, 35 to 130, 40 to 125, 45 to 120, 50 to 115, 55 to 110, 60 to 110, 65 to 105, 70 to 100, or 75 to 95 amino acids. In some instances, the length of the amino acid is in a range of about 22 to about 75 amino acids.

[0058] A number of variant sequences for the at least one region of the TCR chain for variation are de novo synthesized using methods as described herein. In some instances, a number of variant sequences is de novo synthesized for a variable region of an alpha or beta TCR chain. The number of variant sequences may be at least or about 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, or more than 500 sequences. In some instances, the number of variant sequences is in a range of about 10 to 500, 25 to 475, 50 to 450, 75 to 425, 100 to 400, 125 to 375, 150 to 350, 175 to 325, 200 to 300, 225 to 375, 250 to 350, or 275 to 325 sequences.

[0059] Provided herein are variant nucleic acids encoding for variant TCRs, wherein the variant TCRs are antigen specific. In some instances, the antigen is involved in or associated with a disease, disorder, or condition. For example, the antigen is associated with a proliferative disease, a tumorous disease, an inflammatory disease, an immunological disorder, an autoimmune disease, an infectious disease, a viral disease, an allergic reaction, a parasitic reaction, a graft-versus-host disease or a host-versus-graft disease. In some instances, the antigen is an antigen expressed on a tumor cell. In some instances, an antigen is associated with a pathogen such as a virus or bacterium.

[0060] In some instances, the variant TCRs recognize antigens that are tissue-restricted. For example, the variant TCRs are restricted non-vital cell lineages or tissues. In some instances, the variant TCRs recognize antigens from mutated gene products.

[0061] Provided herein are variant TCR libraries, wherein the variant TCRs encode for variants in an antigen binding interface. In some instances, residues for variation are preselected or predicted residues that contact an antigen. In some instances, residues for variation are preselected or predicted residues located in the binding pocket.

[0062] Variant TCR libraries as described herein comprise one or more mutation in a library. In some instances, the TCR variant libraries are single variant libraries comprising variants at a single site across the library. In some instances, the TCR variant libraries are multiple variant libraries comprising variants at a number of sites. In some instances, the number of sites is 2, 3, 4, 5, 6, 7, 8, 9, 10, or more than 10 sites.

[0063] Provided here are libraries where one or more preselected codons in a TCR gene or gene fragment encode for variant amino acid residues to generate variation in resulting variant TCR protein libraries. In some instances, up to 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 amino acid residues are varied. In some instances, up to 30 amino acid residues are varied. In some instances, up to 5 amino acid residues are varied. In some instances, up to 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, or more than 1000 amino acid residues are varied. In some instances, at least or about 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 amino acid residues are varied. In some instances, at least or about 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, or more than 1000 amino acid residues are varied. In some instances, all amino acid residues in a preselected region are varied. In some instances, variant TCR libraries are highly diverse. In some instances, the libraries comprise at least or about 10^6 , 10^8 , 10^9 , 10^{10} , 10^{11} , 10^{12} , or 10^{13} variants. In some instances, the libraries comprise at least or about 10^9 variants.

[0064] Provided herein are variant nucleic acids encoding for a variant TCR for generation of soluble variant TCRs. In some instances, soluble variant TCRs are antigen specific. In some instances, the soluble variant TCRs comprise one or more chains of a TCR α chain, TCR β chain, TCR δ chain, TCR γ chain, or combinations thereof. In some instances, the soluble variant TCRs comprise one or more chains of a TCR α chain. In some instances, the soluble variant TCRs comprise one or more chains of a TCR β chain. In some instances, the soluble variant TCRs comprise one or more chains of a TCR α chain and one or more chains of a TCR β chain. In some instances, a number of TCR chains is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more than 10 TCR chains, wherein the TCR chain is a TCR α chain, TCR β chain, TCR δ chain, TCR γ chain, or combinations thereof.

[0065] A soluble variant TCR described herein may comprise variants in TCR α , TCR β , TCR δ , TCR γ chain, or combinations thereof. In some instances, the soluble variant TCRs comprise variants in the TCR α chain. In some instances, the soluble variant TCRs comprise variants in the

TCR β chain. In some instances, the soluble variant TCRs comprise variants in the TCR α chain and the TCR β chain.

[0066] A soluble variant TCRs described herein may comprise a linker between the TCR chains. For example, the linker is a peptide sequence between the TCR chains. In some instances, the soluble variant TCRs have one or more linker sequences. In some instances, a number of linker sequences is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more than 10 linker sequences. In some instances, a number of linker sequences is in a range of 1-10, 2-8, 3-7, or 4-6 linker sequences.

[0067] Various methods for obtaining soluble TCRs are contemplated herein. In some instances, variant TCR libraries are first de novo synthesized and subsequently transferred into expression vectors (i.e., in bacterial or mammalian cells). In some instances, purified TCRs are obtained to generate soluble TCRs. Soluble TCRs may lack a domain or residues required for insertion or stabilization in a cell membrane. Soluble TCRs may be further modified to be in the form of a fusion protein. Exemplary additional components for inclusion in a fusion protein include, without limitation, an immunoglobulin (Ig) domain of IgG, IgA and IgD, IgM or IgE. In particular, a fusion protein is designed to include a soluble TCR and the constant domain (Fc) or fragment thereof an Ig domain. Exemplary IgG subtypes include, without limitation, IgG1, IgG2, IgG3, and IgG4. For example, such a fusion protein may be composed of a soluble single-chain T cell receptor genetically linked to the Fc region of the human IgG1 heavy chain. Fc regions selected for fusion proteins may be selected based on binding to certain Fc receptors, e.g., Fc γ RI, Fc γ RIIa, Fc γ RIIb/c, Fc γ RIIIa, Fc γ RIIIb, and FcRn.

[0068] Libraries described herein having regions of a common sequence selected for variation may have a preselected variation profile. For example, at least about 1, 5, 10, 15, 20, 30, 40, 50 or more nucleic acids in the common sequence may be selected for invariance, or may be selected to have a limited variation profile. In some instances, variant nucleic acids each encoding for a variant TCR having a variant codon profile are inserted into expression vectors and transferred (e.g., by electroporation, transfection, or transduction) into mammalian cells for screening.

[0069] Variant nucleic acid libraries described herein may be used to generate variant protein libraries. Such protein libraries may be screened for changes in TCR activity, e.g., binding affinity or avidity, stability, protein expression, or target cell killing (e.g., cancer cell killing), or ADCC activity. Screening may include binding assays with an antigen(s) associated with a particular cancer of interest. The cancer may be a solid cancer or a hematologic cancer. In some instances, the cancer is bladder cancer, lung cancer, brain cancer, melanoma, breast cancer, Non-Hodgkin lymphoma, cervical cancer, ovarian cancer, colorectal cancer, pancreatic cancer, esophageal cancer, prostate cancer, kidney cancer, skin cancer, leukemia, thyroid cancer, liver cancer, or uterine

cancer. In some instances, the antigen is MART1, CEA, gp100, NY-ESO-1, WT1, MUC1, HER-2/neu, MAGE-A3, p53, PSMA, GD2, Proteinase 3 (PR1), Tyrosinase, Survivin, PSA, hTERT, and EphA2. Exemplary antigens for use in binding assays include, without limitation, those provided in the **Table 4**.

Table 4. Exemplary antigens for use in binding assays

Antigen peptide name	Protein Sequence	Disease Target
MAGE A3	KVAELVHFL	Melanoma
MAGE A12	KMAELVHFL	Melanoma
MAGE A2	KMVELVHFL	Melanoma
MAGE A6	KVAKLVHFL	Melanoma
NY-ESO-1	SLLMWITQV	Many cancers including melanoma and sarcoma
CEA peptide	IMIGVLVGV	Metastatic colorectal cancer

[0070] Screening may comprise comprises *in vitro*, *in vivo*, or *ex vivo* assays. Cells for screening include primary cells taken from living subjects or cell lines. Cells may be from prokaryotes (e.g., bacteria and fungi) or eukaryotes (e.g., animals and plants). Exemplary animal cells include, without limitation, those from a mouse, rabbit, primate, and insect. In some instances, cells for screening include a cell line including, but not limited to, Chinese Hamster Ovary (CHO) cell line, human embryonic kidney (HEK) cell line, or baby hamster kidney (BHK) cell line. In some instances, nucleic acid libraries described herein may also be delivered to a multicellular organism. Exemplary multicellular organisms include, without limitation, a plant, a mouse, rabbit, primate, and insect.

[0071] Variant TCR libraries generated using methods described herein may be screened to select a modified TCR gene sequence providing for protein complex with improved affinity (measure of the strength of interaction between an epitope and an antibody's antigen binding site) for a tumor antigen. Additional functional considerations, such as variant gene expression, avidity (measure of the overall strength of an antibody-antigen complex), stability, and target cell (i.e. cancer cell) specificity are also assessed. In some instances, the increased specificity of a TCR complex provides for reduced cross-reactivity to non-cancer associated antigens compared to a reference non-variant TCR complex. In some instances, the variant TCR libraries are screened for localization within a cell. In some instances, the variant TCR libraries are screened to identify properly localized variant TCRs.

[0072] Provided herein are variant TCR libraries with improved affinity or specificity for an antigen that are expressed in cells. In some instances, the variant TCR libraries are expressed in cells to generate cells with altered specificity for an antigen. In some instances, the cells are T cells. In some instances, the variant library comprises sequences encoding for variation of at least a single codon of a variable domain of a TCR alpha chain or TCR beta chain. In some instances, the variant library comprises sequences encoding for variation of multiple codons of a variable domain of a TCR alpha chain or TCR beta chain. For example, the variant library comprises variation of a sequence in **Table 2** or **Table 3**. In some instances, variant TCR libraries comprise variation in one or more exons of a variable gene of a TCR alpha chain. Exemplary exon sequences for variation are seen in **Table 2** or **Table 3**. In some instances, variant TCR libraries comprise variation in one or more exons of a variable gene of a TCR beta chain.

[0073] In some instances, variant TCR libraries that are expressed in cells are used to identify variant TCRs with improved variant gene expression, avidity, stability, affinity, or specificity. For example, a first variant TCR library is generated that comprises improved specificity to a tumor antigen. In some instances, following identification of variant TCRs with improved gene expression, avidity, stability, affinity, or specificity, those variant TCRs are further varied to produce a second library of variant TCRs with a second improvement. For example, variant TCRs with improved specificity are further varied to identify variants further comprising improved stability. In some instances, the second library comprises variants in a same region as the first library. In some instances, the second library comprises variants in a different region of the first library. For example, the first library comprises variants in a variable region of the TCR and the second library comprises variants in a constant region of the TCR. In some instances, a number of variant TCR libraries are generated. In some instances, the number of variant TCR libraries is at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, or more than 12 variant libraries. In some instances, each of the variant libraries has improvements in gene expression, avidity, stability, affinity, or specificity.

[0074] In some instances, cells engineered to express variant TCRs are introduced into a subject. In some instances, the cells are autologous, meaning derived from a subject's own cells. Alternately, cells expressing variant TCRs are allogeneic, meaning derived from another subject with a similar tissue type. In some instances, cells are tailored to the subject. In some instances, cells are compatible with local tissue.

[0075] Variant TCR libraries generated using methods described herein may be screened to select at least one modified TCR gene coding sequence providing for enhanced expression profile of an engineered TCR protein complex in cells. Strategies to enhance relative gene expression include codon optimization, introduction of one or more disulfide bonds between the TCR chains,

the introduction of murine residue sequence into the constant region domain, variation of non-complementary coding region of variable domain(s), use of TALEN or CRISPR technology to selectively knockdown or halt expression of endogenous TCR protein in the modified T cells, delivery of immune stimulatory cytokines to the modified T cells, or a combination of any or all of these strategies. The insertion of disulfide bond (e.g., in the hinge region) of a varied gene sequence encoding for TCR α or β subunits provides for advantageous assembly of the engineered variants when expressed in cells. The number of disulfide bonds added to TCR gene sequence described herein may be at least 1, 2, 3, 4, 5 or more.

[0076] In some instances, de novo synthesis methods described herein provide for generation of a variant library of antigens to identify an antigen which are both highly efficient at binding to a variant TCR complex generated by methods described herein as well as by the mapped MHC I molecule of a subject. In this way, both the TCR complex and the antigen binding portion can be optimized for an enhanced interaction at the immunological synapse. For screening purposes, nucleic acids encoding for a variant antigen library are first de novo synthesized, and subsequently transferred into expression vectors for expression (i.e., in bacterial or mammalian cells) in order to obtain purified peptide for binding assays with TCR complexes and MHC I complexes.

[0077] Variant Library Synthesis

[0078] Methods described herein provide for synthesis of a library of nucleic acids each encoding for a predetermined variant of at least one predetermined reference nucleic acid sequence. In some cases, the predetermined reference sequence is nucleic acid sequence encoding for a protein, and the variant library comprises sequences encoding for variation of at least a single codon such that a plurality of different variants of a single residue in the subsequent protein encoded by the synthesized nucleic acid are generated by standard translation processes. The synthesized specific alterations in the nucleic acid sequence can be introduced by incorporating nucleotide changes into overlapping or blunt ended nucleic acid primers. Alternatively, a population of nucleic acids may collectively encode for a long nucleic acid (e.g., a gene) and variants thereof. In this arrangement, the population of nucleic acids can be hybridized and subject to standard molecular biology techniques to form the long nucleic acid (e.g., a gene) and variants thereof. When the long nucleic acid (e.g., a gene) and variants thereof are expressed in cells, a variant protein library is generated. Similarly, provided here are methods for synthesis of variant libraries encoding for RNA sequences (e.g., miRNA, shRNA, and mRNA) or DNA sequences (e.g., enhancer, promoter, UTR, and terminator regions). Also provided here are downstream applications for variants selected out of the libraries synthesized using methods described here. Downstream applications include identification of variant nucleic acid or protein sequences with

enhanced biologically relevant functions, e.g., biochemical affinity, enzymatic activity, changes in cellular activity, and for the treatment or prevention of a disease state.

[0079] *Synthesis followed by PCR Mutagenesis*

[0080] A first process for synthesis of a variant library of nucleic acids is for PCR mutagenesis (saturating or non-saturating) methods. In this workflow, a plurality of nucleic acids are synthesized, wherein each nucleic acid encodes for a predetermined sequence which is a predetermined variant of a reference nucleic acid sequence. Referring to the figures, an exemplary workflow is depicted in **FIGS. 1A-1D**, wherein nucleic acids are generated on a surface. **FIG. 1A** depicts an expansion view of a single cluster of a surface with 121 loci. Each nucleic acid depicted in **FIG. 1B** is a primer that can be used for amplification from a reference nucleic acid sequence to produce a library of variant long nucleic acids, **FIG. 1C**. The library of variant long nucleic acids is then, optionally, subject to transcription and or translation to generate a variant RNA or protein library, **FIG. 1D**. In this exemplary illustration, a device having a substantially planar surface is used for de novo synthesis of nucleic acids is depicted, **FIG. 1A**. In some instances, the device comprises a cluster of loci, wherein each locus is a site for nucleic acid extension. In some instances, a single cluster comprises all the nucleic acid variants needed to generate a desired variant sequence library. In an alternative arrangement, a plate comprises a field of loci which are not segregated into clusters.

[0081] In some instances, nucleic acids synthesized within a cluster (e.g., as seen in **FIG. 1A**) are amplified by PCR. Such an arrangement may provide for improved nucleic acid representation in compared to amplification of non-identical nucleic acids across an entire plate without a clustered arrangement. In some instances, amplification of nucleic acids synthesized on surfaces of loci within a cluster overcomes negative effects on representation due to repeated synthesis of large nucleic acid populations having nucleic acids with heavy GC content. In some instances, a cluster described herein, comprises about 50-1000, 75-900, 100-800, 125-700, 150-600, 200-500, 50-500 or 300-400 discrete loci. In some instances, a loci is a spot, well, microwell, channel, or post. In some instances, each cluster has at least 1X, 2X, 3X, 4X, 5X, 6X, 7X, 8X, 9X, 10X, or more redundancy of separate features supporting extension of nucleic acids having identical sequence.

[0082] A de novo synthesized nucleic acid library described herein may comprise a plurality of nucleic acids, each with at least one variant sequence at first position, position “x”, and each variant nucleic acid is used as a primer in a first round of PCR to generate a first extension product. In this example, position “x” in a first nucleic acid **220** encodes for a variant codon sequence, i.e., one of 19 possible variants from a reference sequence. See **FIG. 2A**. A second nucleic acid **225** comprising sequence overlapping that of the first nucleic acid is also used as a primer in a separate

round of PCR to generate a second extension product. In addition, outer primers **215**, **230** may be used for amplification of fragment from a long nucleic acid sequence. The resultant amplification products are fragments of the long nucleic acid sequence **235**, **240**. See **FIG. 2B**. The fragments of the long nucleic acid sequence **235**, **240** are then hybridized, and subject to an extension reaction to form a variant of the long nucleic acid **245**. See **FIG. 2C**. The overlapping ends of the first and second extension products may serve as primer of a second round of PCR, thereby generating a third extension product (**FIG. 2D**) that contains the variant. To increase the yield, the variant of the long nucleic acid is amplified in a reaction including a DNA polymerase, amplification reagents, and the outer primers **215**, **230**. In some instances, the second nucleic acid comprises sequence adjacent to, but not including, the variant site. In an alternative arrangement, a first nucleic acid is generated that has region that overlaps with a second nucleic acid. In this scenario, the first nucleic acid is synthesized with variation at a single codon for up to 19 variants. The second nucleic acid does not comprise a variant sequence. Optionally, a first population comprises the first nucleic acid variants and additional nucleic acids encoding for variants at a different codon site. Alternatively, the first nucleic acid and the second nucleic acid may be designed for blunt end ligation.

[0083] In alternative mutagenesis PCR method is depicted in **FIGS. 3A-3F**. In such a process, a template nucleic acid molecule **300** comprising a first and second strand **305**, **310** is amplified in a PCR reaction containing a first primer **315** and a second primer **320** (**FIG. 3A**). The amplification reaction includes uracil as a nucleotide reagent. A uracil-labeled extension product **325** (**FIG. 3B**) is generated, optionally purified, and serves as a template for a subsequent PCR reaction using a first nucleic acid **335** and a plurality of second nucleic acid **330** to generate first extension products **340** and **345** (**FIGS. 3C-3D**). In this process, plurality of second nucleic acid **330** comprises nucleic acids encoding for variant sequences (denoted as X, Y, and Z, in **FIG. 3C**). The uracil-labeled template nucleic acid is digested by a uracil-specific excision reagent, e.g., USER digest available commercially from New England Biolabs. Variant **335** and different codons **330** with variants X, Y, and Z are added and a limited PCR step is performed to generate **FIG. 3D**. After the uracil-containing template is digested, the overlapping ends of the extension products serve to prime a PCR reaction with the first extension products **340** and **345** acting as primers in combination with a first outer primer **350** and a second outer primer **355**, thereby generating a library of nucleic acid molecules **360** containing a plurality of variants X, Y, and Z at the variant site **FIG. 3F**.

[0084] *De novo synthesis of a population with variant and non-variant portions of a long nucleic acid*

[0085] In a second process for synthesis of a variant library, a surface is used for de novo synthesis of multiple fragments of a long nucleic acid, wherein at least one of the fragments is synthesized in multiple versions, each version being of a different variant sequence. In this arrangement, all of the fragments needed to assemble a library of variant long range nucleic acids are de novo synthesized. The synthesized fragments may have overlapping sequence such that, following synthesis, the fragment library is subject to hybridization. Following hybridization, an extension reaction may be performed to fill in any complementary gaps.

[0086] Alternatively, the synthesized fragments may be amplified with primers and then subject to either blunt end ligation or overlapping hybridization. In some instances, the device comprises a cluster of loci, wherein each locus is a site for nucleic acid extension. In some instances, a single cluster comprises all the nucleic acid variants and other fragment sequences of a predetermined long nucleic acid to generate a desired variant nucleic acid sequence library. The cluster may comprise about 50 to 500 loci. In some arrangements, a cluster comprises greater than 500 loci.

[0087] Each individual nucleic acid in the first nucleic acid population may be generated on a separate, individually addressable locus of a cluster. One nucleic acid variant may be represented by a plurality of individually addressable loci. Each variant in the first nucleic acid population may be represented 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more times. In some instances, each variant in the first nucleic acid population is represented at 3 or less loci. In some instances, each variant in the first nucleic acid population is represented at two loci. In some instances, each variant in the first nucleic acid population is represented at only a single locus.

[0088] Methods are provided herein to generate nucleic acid libraries with reduced redundancy. In some instances, variant polynucleotides may be generated without the need to synthesize the variant polynucleotide more than 1 time to obtain the desired variant polynucleotide. In some instances, the present disclosure provides methods to generate variant polynucleotides without the need to synthesize the variant polynucleotide more than 1, 2, 3, 4, 5 times, 6, 7, 8, 9, 10, or more times to generate the desired variant polynucleotide.

[0089] Variant polynucleotides may be generated without the need to synthesize the variant polynucleotide at more than 1 discrete site to obtain the desired variant polynucleotide. The present disclosure provides methods to generate variant polynucleotides without the need to synthesize the variant polynucleotide at more than 1 site, 2 sites, 3 sites, 4 sites, 5 sites, 6 sites, 7 sites, 8 sites, 9 sites, or 10 sites, to generate the desired variant polynucleotide. In some instances, a polynucleotide is synthesized in at most 6, 5, 4, 3, 2, or 1 discrete sites. The same polynucleotide may be synthesized in 1, 2, or 3 discrete loci on a surface.

[0090] In some instances, the amount of loci representing a single variant polynucleotide is a function of the amount of nucleic acid material required for downstream processing, e.g., an amplification reaction or cellular assay. In some instances, the amount of loci representing a single variant polynucleotide is a function of the available loci in a single cluster.

[0091] Provided herein are methods for generation of a library of nucleic acids comprising variant nucleic acids differing at a plurality of sites in a reference nucleic acid. In such cases, each variant library is generated on an individually addressable locus within a cluster of loci. It will be understood that the number of variant sites represented by the nucleic acid library will be determined by the number of individually addressable loci in the cluster and the number of desired variants at each site. In some instances, each cluster comprises about 50 to 500 loci. In some instances, each cluster comprises 100 to 150 loci.

[0092] In an exemplary arrangement, 19 variants are represented at a variant site corresponding to codons encoding for each of the 19 possible variant amino acids. In another exemplary case, 61 variants are represented at a variant site corresponding to triplets encoding for each of the 19 possible variant amino acids. In a non-limiting example, a cluster comprises 121 individually addressable loci. In this example, a nucleic acid population comprises 6 replicates each of a single-site variant (6 replicates $\times$ 1 variant site $\times$ 19 variants = 114 loci), 3 replicates each of a double-site variant (3 replicates $\times$ 2 variant sites $\times$ 19 variants = 114 loci), or 2 replicates each of a triple-site variant (2 replicates $\times$ 3 variant sites $\times$ 19 variants = 114 loci). In some instances, a nucleic acid population comprises variants at four, five, six or more than six variant sites.

[0093] *Codon variation*

[0094] Variant nucleic acid libraries described herein may comprise a plurality of nucleic acids, wherein each nucleic acid encodes for a variant codon sequence compared to a reference nucleic acid sequence. In some instances, each nucleic acid of a first nucleic acid population contains a variant at a single variant site. In some instances, the first nucleic acid population contains a plurality of variants at a single variant site such that the first nucleic acid population contains more than one variant at the same variant site. The first nucleic acid population may comprise nucleic acids collectively encoding multiple codon variants at the same variant site. The first nucleic acid population may comprise nucleic acids collectively encoding up to 19 or more codons at the same position. The first nucleic acid population may comprise nucleic acids collectively encoding up to 60 variant triplets at the same position, or the first nucleic acid population may comprise nucleic acids collectively encoding up to 61 different triplets of codons at the same position. Each variant may encode for a codon that results in a different amino acid during translation. **Table 5** provides a listing of each codon possible (and the representative amino acid) for a variant site.

Table 5. List of codons and amino acids

Amino Acids	One letter code	Three letter code	Codons					
Alanine	A	Ala	GCA	GCC	GCG	GCT		
Cysteine	C	Cys	TGC	TGT				
Aspartic acid	D	Asp	GAC	GAT				
Glutamic acid	E	Glu	GAA	GAG				
Phenylalanine	F	Phe	TTC	TTT				
Glycine	G	Gly	GGA	GGC	GGG	GGT		
Histidine	H	His	CAC	CAT				
Isoleucine	I	Iso	ATA	ATC	ATT			
Lysine	K	Lys	AAA	AAG				
Leucine	L	Leu	TTA	TTG	CTA	CTC	CTG	CTT
Methionine	M	Met	ATG					
Asparagine	N	Asn	AAC	AAT				
Proline	P	Pro	CCA	CCC	CCG	CCT		
Glutamine	Q	Gln	CAA	CAG				
Arginine	R	Arg	AGA	AGG	CGA	CGC	CGG	CGT
Serine	S	Ser	AGC	AGT	TCA	TCC	TCG	TCT
Threonine	T	Thr	ACA	ACC	ACG	ACT		
Valine	V	Val	GTA	GTC	GTG	GTT		
Tryptophan	W	Trp	TGG					

Tyrosine	Y	Tyr	TAC	TAT
----------	---	-----	-----	-----

[0095] A nucleic acid population may comprise varied nucleic acids collectively encoding up to 20 codon variations at multiple positions. In such cases, each nucleic acid in the population comprises variation for codons at more than one position in the same nucleic acid. In some instances, each nucleic acid in the population comprises variation for codons at 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more codons in a single nucleic acid. In some instances, each variant long nucleic acid comprises variation for codons at 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more codons in a single long nucleic acid. In some instances, the variant nucleic acid population comprises variation for codons at 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more codons in a single nucleic acid. In some instances, the variant nucleic acid population comprises variation for codons in at least about 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 or more codons in a single long nucleic acid.

[0096] Provided herein are processes where a second nucleic acid population is generated on a second cluster containing a plurality of individually addressable loci. The second nucleic acid population may comprise a plurality of second nucleic acids that are constant for each codon position (i.e., encode the same amino acid at each position). The second nucleic acid may overlap with at least a portion of the first nucleic acids. In some instances, the second nucleic acids do not contain the variant site represented on the first nucleic acids. Alternatively, the second nucleic acid population may comprise a plurality of second nucleic acids that variant for one or more codon positions.

[0097] Provided herein are methods for synthesizing a library of nucleic acids where a single population of nucleic acids is generated comprising variants at multiple codon positions. A first nucleic acid population may be generated on a first cluster containing a plurality of individually addressable loci. In such cases, the first nucleic acid population comprises variants at different codon positions. In some instances, the different sites are consecutive (i.e., encoding consecutive amino acids). A first polynucleotide acid population may comprise varied nucleic acids collectively encoding up to 19 codon variants at the same, or additional variant site. A first polynucleotide acid population may include a plurality of first nucleic acids that contains up to 19 variants at position x, up to 19 variants at position y, and up to 19 variants at position z. In such an arrangement, each variant encodes a different amino acid such that up to 19 amino acid variants are encoded at each of the different variant sites. In an additional instance, a second nucleic acid population is generated on a second cluster containing a plurality of individually addressable loci.

The second nucleic acid population may comprise a plurality of second nucleic acids that are constant for each codon position (i.e., encode the same amino acid at each position). The second nucleic acids may overlap with at least a portion of the first nucleic acids. The second nucleic acids may not contain the variant site represented on the first nucleic acids.

[0098] Variant nucleic acid libraries generated by processes described herein provide for the generation of variant protein libraries. In a first exemplary arrangement, a template nucleic acid encodes for sequence that, when transcribed and translated, results in a reference amino acid sequence (**FIG. 4A**) having a number of codon positions, indicated by a single circle. Nucleic acid variants of the template can be generated using methods described herein. In some instances, a single variant is present in the nucleic acid, resulting in a single amino acid sequence (**FIG. 4B**). In some instances, more than one variant is present in the nucleic acid, wherein the variants are separated by one or more codons, resulting in a protein with spacing between variant residues (**FIG. 4C**). In some instances, more than one variant is present in the nucleic acid, wherein the variants are sequential and adjacent or consecutive to one another, resulting in spaced variant stretches of residues (**FIG. 4D**). In some instances, two stretches of variants are present in the nucleic acid, wherein each stretch of variants comprises sequential and adjacent or consecutive variants (**FIG. 4E**).

[0099] Provided herein are methods to generate a library of nucleic acid variants, wherein each variant comprises a single position codon variant. In one instance, a template nucleic acid has a number of codon positions wherein exemplary amino acid residues are indicated by circles with their respective one letter code protein codon, **FIG. 5A**. **FIG. 5B** depicts a library of amino acid variants encoded by a library of variant nucleic acids, wherein each variant comprises a single position variant, indicated by an "X", of located at a different single site. A first position variant has any codon to replace alanine, a second variant with any codon encoded by the library of variant nucleic acids to replace tryptophan, a third variant with any codon to replace isoleucine, a fourth variant with any codon to replace lysine, a fifth variant with any codon to replace arginine, a sixth variant with any codon to replace glutamic acid, and a seventh variant with any codon to replace glutamine. When all or less than all codon variants are encoded by the variant nucleic acid library, a resulting a corresponding population of amino acid sequence variants is generated following protein expression (i.e., standard cellular events of DNA transcription followed by translation and processing events).

[00100] In some arrangements, a library is generated with multiple sites of single position variants. As depicted in **FIG. 6A**, a wild-type template is provided. **FIG. 6B** depicts the resultant

amino acid sequence with two sites of single position codon variants, wherein each codon variant encoding for a different amino acid is indicated by differently patterned circles.

[00101] Provided herein are methods to generate a library having a stretch of multiple site, single position variants. Each stretch of nucleic acid may have 1, 2, 3, 4, 5, or more variants. Each stretch of nucleic acid may have at least 1 variants. Each stretch of nucleic acid may have at least 2 variants. Each stretch of nucleic acid may have at least 3 variants. For example, a stretch of 5 nucleic acids may have 1 variant. A stretch of 5 nucleic acids may have 2 variants. A stretch of 5 nucleic acids may have 3 variants. A stretch of 5 nucleic acids may have 4 variants. For example, a stretch of 4 nucleic acids may have 1 variant. A stretch of 4 nucleic acids may have 2 variants. A stretch of 4 nucleic acids may have 3 variants. A stretch of 4 nucleic acids may have 4 variants.

[00102] In some instances, single position variants may all encode for the same amino acid, e.g. a histidine. As depicted in **FIG. 7A**, a reference amino acid sequence is provided. In this arrangement, a stretch of a nucleic acid encodes for multiple sites of single position variants and, when expressed, results in an amino acid sequence having all single position variants encoding for a histidine, **FIG. 7B**. In some embodiments, a variant library synthesized by methods described herein does not encode for more than 4 histidine residues in a resultant amino acid sequence.

[00103] In some instances, a variant library of nucleic acids generated by methods described herein provides for expression of amino acid sequences have separate stretches of variation. A template amino acid sequence is depicted in **FIG. 8A**. A stretch of nucleic acids may have only 1 variant codon in two stretches and, when expressed, result in an amino acid sequence depicted in **FIG. 8B**. Variants are depicted in **FIG. 8B** by the differently patterned circles to indicate variation in amino acids are different position in a single stretch.

[00104] Provided herein are methods and devices to synthesize nucleic acid libraries with 1, 2, 3, or more codon variants, wherein the variant for each site is selectively controlled. The ratio of two amino acids for a single site variant may be about 1:100, 1:50, 1:10, 1:5, 1:3, 1:2, 1:1. The ratio of three amino acids for a single site variant may be about 1:1:100, 1:1:50, 1:1:20, 1:1:10, 1:1:5, 1:1:3, 1:1:2, 1:1:1, 1:10:10, 1:5:5, 1:3:3, or 1:2:2. **FIG. 9A** depicts a wild-type reference amino acid sequence encoded by a wild-type nucleic acid sequence. **FIG. 9B** depicts a library of amino acid variants, wherein each variant comprising a stretch of sequence (indicated by the patterned circles), wherein each position may have a certain ratio of amino acids in the resultant variant protein library. The resultant variant protein library is encoded by a variant nucleic acid library generated by methods described herein. In this illustration, 5 positions are varied: the first position **900** has a 50/50 K/R ratio; the second position **910** has a 50/25/25 V/L/S ratio, the third position **920** has a

50/25/25 Y/R/D ratio, the fourth position **930** has an equal ratio for all 20 amino acids, and the fifth position **940** has a 75/25 ratio for G/P. The ratios described herein are exemplary only.

[00105] *Variation in expression cassettes*

[00106] In some instances, a synthesized variant library is generated which encodes for a portion of an expression construct. Exemplary portions of an expression construct include the promoter, open reading frame, and termination region. In some instances, the expression construct encodes for one, two, three or more expression cassettes. A polynucleotide library may be generated, encoding for codon variation at a single site or multiple sites separate regions that make up portions of an expression construct cassette, as depicted in **FIG. 11**. To generate a two construct expressing cassette, variant nucleic acids were synthesized encoding at least a portion of a variant sequence of a first promoter **1110**, first open reading frame **1120**, first terminator **1130**, second promoter **1140**, second open reading frame **1150**, or second terminator sequence **1160**. After rounds of amplification, as described in previous examples, a library of 1,024 expression constructs was generated. **FIG. 11** provides but one example arrangement. In some instances, additional regulator sequences, such as untranslated regulatory region (UTR) or an enhancer region, is also included in an expression cassette referred to herein. An expression cassette may comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or more components for which variant sequences are generated by methods described herein. In some instances, the expression construct comprises more than one gene in a multicistronic vector. In one example, the synthesized DNA nucleic acids are inserted into viral vectors (e.g., a lentivirus) and then packaged for transduction into cells, or non-viral vectors for transfer into cells, followed by screening and analysis.

[00107] Expression vectors for inserting nucleic acids disclosed herein comprise mammalian cells, e.g., human, non-human primate, pig, rabbit and mouse. Exemplary expression vectors include, without limitation, mammalian expression vectors: pSF-CMV-NEO-NH2-PPT-3XFLAG, pSF-CMV-NEO-COOH-3XFLAG, pSF-CMV-PURO-NH2-GST-TEV, pSF-OXB20-COOH-TEV-FLAG(R)-6His, pCEP4 pDEST27, pSF-CMV-Ub-KrYFP, pSF-CMV-FMDV-daGFP, pEF1a-mCherry-N1 Vector, pEF1a-tdTomato Vector, pSF-CMV-FMDV-Hygro, pSF-CMV-PGK-Puro, pMCP-tag(m), and pSF-CMV-PURO-NH2-CMYC. Nucleic acids synthesized by methods described herein may be transferred into cells by various methods known in the art, including, without limitation, transfection, transduction, and electroporation. Exemplary cellular functions tested include, without limitation, changes in cellular proliferation, death, migration/adhesion, metabolic, and cell-signaling activity.

[00108] Highly Parallel Nucleic Acid Synthesis

[00109] Provided herein is a platform approach utilizing miniaturization, parallelization, and vertical integration of the end-to-end process from nucleic acid synthesis to gene assembly within nanowells on silicon to create a revolutionary synthesis platform. Devices described herein provide, with the same footprint as a 96-well plate, a silicon synthesis platform is capable of increasing throughput by a factor of up to 1,000 or more compared to traditional synthesis methods, with production of up to approximately 1,000,000 or more nucleic acids, or 10,000 or more genes in a single highly-parallelized run.

[00110] With the advent of next-generation sequencing, high resolution genomic data has become an important factor for studies that delve into the biological roles of various genes in both normal biology and disease pathogenesis. At the core of this research is the central dogma of molecular biology and the concept of “residue-by-residue transfer of sequential information.” Genomic information encoded in the DNA is transcribed into a message that is then translated into the protein that is the active product within a given biological pathway.

[00111] Another exciting area of study is on the discovery, development and manufacturing of therapeutic molecules focused on a highly-specific cellular target. High diversity DNA sequence libraries are at the core of development pipelines for targeted therapeutics. Gene mutants are used to express proteins in a design, build, and test protein engineering cycle that ideally culminates in an optimized gene for high expression of a protein with high affinity for its therapeutic target. As an example, consider the binding pocket of a receptor. The ability to test all sequence permutations of all residues within the binding pocket simultaneously will allow for a thorough exploration, increasing chances of success. Saturation mutagenesis, in which a researcher attempts to generate all possible mutations at a specific site within the receptor, represents one approach to this development challenge. Though costly and time and labor-intensive, it enables each variant to be introduced into each position. In contrast, combinatorial mutagenesis, where a few selected positions or short stretch of DNA may be modified extensively, generates an incomplete repertoire of variants with biased representation.

[00112] To accelerate the drug development pipeline, a library with the desired variants available at the intended frequency in the right position available for testing—in other words, a precision library, enables reduced costs as well as turnaround time for screening. Provided herein are methods for synthesizing nucleic acid synthetic variant libraries which provide for precise introduction of each intended variant at the desired frequency. To the end user, this translates to the ability to not only thoroughly sample sequence space but also be able to query these hypotheses in an efficient manner, reducing cost and screening time. Genome-wide editing can elucidate

important pathways, libraries where each variant and sequence permutation can be tested for optimal functionality, and thousands of genes can be used to reconstruct entire pathways and genomes to re-engineer biological systems for drug discovery.

[00113] In a first example, a drug itself can be optimized using methods described herein. For example, to improve a specified function of an antibody, a variant nucleic acid library encoding for a portion of the antibody is designed and synthesized. A variant nucleic acid library for the antibody can then be generated by processes described herein (e.g., PCR mutagenesis followed by insertion into a vector). The antibody is then expressed in a production cell line and screened for enhanced activity. Example screens include examining modulation in binding affinity to an antigen, stability, or effector function (e.g., ADCC, complement, or apoptosis). Exemplary regions to optimize the antibody include, without limitation, the Fc region, Fab region, variable region of the Fab region, constant region of the Fab region, variable domain of the heavy chain or light chain (V_H or V_L), and specific complementarity-determining regions (CDRs) of V_H or V_L .

[00114] Nucleic acid libraries synthesized by methods described herein may be expressed in various cells associated with a disease state. Cells associated with a disease state include cell lines, tissue samples, primary cells from a subject, cultured cells expanded from a subject, or cells in a model system. Exemplary model systems include, without limitation, plant and animal models of a disease state.

[00115] To identify a variant molecule associated with prevention, reduction or treatment of a disease state, a variant nucleic acid library described herein is expressed in a cell associated with a disease state, or one in which a cell a disease state can be induced. In some instances, an agent is used to induce a disease state in cells. Exemplary tools for disease state induction include, without limitation, a Cre/Lox recombination system, LPS inflammation induction, and streptozotocin to induce hypoglycemia. The cells associated with a disease state may be cells from a model system or cultured cells, as well as cells from a subject having a particular disease condition. Exemplary disease conditions include a bacterial, fungal, viral, autoimmune, or proliferative disorder (e.g., cancer). In some instances, the variant nucleic acid library is expressed in the model system, cell line, or primary cells derived from a subject, and screened for changes in at least one cellular activity. Exemplary cellular activities include, without limitation, proliferation, cycle progression, cell death, adhesion, migration, reproduction, cell signaling, energy production, oxygen utilization, metabolic activity, and aging, response to free radical damage, or any combination thereof.

[00116] Substrates

[00117] Provided herein are substrates comprising a plurality of clusters, wherein each cluster comprises a plurality of loci that support the attachment and synthesis of nucleic acids. The term

“locus” as used herein refers to a discrete region on a structure which provides support for polynucleotides encoding for a single predetermined sequence to extend from the surface. In some instances, a locus is on a two dimensional surface, *e.g.*, a substantially planar surface. In some instances, a locus refers to a discrete raised or lowered site on a surface *e.g.*, a well, microwell, channel, or post. In some instances, a surface of a locus comprises a material that is actively functionalized to attach to at least one nucleotide for nucleic acid synthesis, or preferably, a population of identical nucleotides for synthesis of a population of nucleic acids. In some instances, nucleic acid refers to a population of nucleic acids encoding for the same nucleic acid sequence. In some instances, a surface of a device is inclusive of one or a plurality of surfaces of a substrate.

[00118] Average error rates for nucleic acids synthesized within a library using the systems and methods provided may be less than 1 in 1000, less than 1 in 1250, less than 1 in 1500, less than 1 in 2000, less than 1 in 3000 or less often. In some instances, average error rates for nucleic acids synthesized within a library using the systems and methods provided are less than 1/500, 1/600, 1/700, 1/800, 1/900, 1/1000, 1/1100, 1/1200, 1/1250, 1/1300, 1/1400, 1/1500, 1/1600, 1/1700, 1/1800, 1/1900, 1/2000, 1/3000, or less. In some instances, average error rates for nucleic acids synthesized within a library using the systems and methods provided are less than 1/1000.

[00119] In some instances, aggregate error rates for nucleic acids synthesized within a library using the systems and methods provided are less than 1/500, 1/600, 1/700, 1/800, 1/900, 1/1000, 1/1100, 1/1200, 1/1250, 1/1300, 1/1400, 1/1500, 1/1600, 1/1700, 1/1800, 1/1900, 1/2000, 1/3000, or less compared to the predetermined sequences. In some instances, aggregate error rates for nucleic acids synthesized within a library using the systems and methods provided are less than 1/500, 1/600, 1/700, 1/800, 1/900, or 1/1000. In some instances, aggregate error rates for nucleic acids synthesized within a library using the systems and methods provided are less than 1/1000.

[00120] In some instances, an error correction enzyme may be used for nucleic acids synthesized within a library using the systems and methods provided can use. In some instances, aggregate error rates for nucleic acids with error correction can be less than 1/500, 1/600, 1/700, 1/800, 1/900, 1/1000, 1/1100, 1/1200, 1/1300, 1/1400, 1/1500, 1/1600, 1/1700, 1/1800, 1/1900, 1/2000, 1/3000, or less compared to the predetermined sequences. In some instances, aggregate error rates with error correction for nucleic acids synthesized within a library using the systems and methods provided can be less than 1/500, 1/600, 1/700, 1/800, 1/900, or 1/1000. In some instances, aggregate error rates with error correction for nucleic acids synthesized within a library using the systems and methods provided can be less than 1/1000.

[00121] Error rate may limit the value of gene synthesis for the production of libraries of gene variants. With an error rate of 1/300, about 0.7% of the clones in a 1500 base pair gene will be correct. As most of the errors from polynucleotide synthesis result in frame-shift mutations, over 99% of the clones in such a library will not produce a full-length protein. Reducing the error rate by 75% would increase the fraction of clones that are correct by a factor of 40. The methods and compositions of the disclosure allow for fast de novo synthesis of large polynucleotide and gene libraries with error rates that are lower than commonly observed gene synthesis methods both due to the improved quality of synthesis and the applicability of error correction methods that are enabled in a massively parallel and time-efficient manner. Accordingly, libraries may be synthesized with base insertion, deletion, substitution, or total error rates that are under 1/300, 1/400, 1/500, 1/600, 1/700, 1/800, 1/900, 1/1000, 1/1250, 1/1500, 1/2000, 1/2500, 1/3000, 1/4000, 1/5000, 1/6000, 1/7000, 1/8000, 1/9000, 1/10000, 1/12000, 1/15000, 1/20000, 1/25000, 1/30000, 1/40000, 1/50000, 1/60000, 1/70000, 1/80000, 1/90000, 1/100000, 1/125000, 1/150000, 1/200000, 1/300000, 1/400000, 1/500000, 1/600000, 1/700000, 1/800000, 1/900000, 1/1000000, or less, across the library, or across more than 80%, 85%, 90%, 93%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.8%, 99.9%, 99.95%, 99.98%, 99.99%, or more of the library. The methods and compositions of the disclosure further relate to large synthetic polynucleotide and gene libraries with low error rates associated with at least 30%, 40%, 50%, 60%, 70%, 75%, 80%, 85%, 90%, 93%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.8%, 99.9%, 99.95%, 99.98%, 99.99%, or more of the polynucleotides or genes in at least a subset of the library to relate to error free sequences in comparison to a predetermined/preselected sequence. In some instances, at least 30%, 40%, 50%, 60%, 70%, 75%, 80%, 85%, 90%, 93%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.8%, 99.9%, 99.95%, 99.98%, 99.99%, or more of the polynucleotides or genes in an isolated volume within the library have the same sequence. In some instances, at least 30%, 40%, 50%, 60%, 70%, 75%, 80%, 85%, 90%, 93%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.8%, 99.9%, 99.95%, 99.98%, 99.99%, or more of any polynucleotides or genes related with more than 95%, 96%, 97%, 98%, 99%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9% or more similarity or identity have the same sequence. In some instances, the error rate related to a specified locus on a polynucleotide or gene is optimized. Thus, a given locus or a plurality of selected loci of one or more polynucleotides or genes as part of a large library may each have an error rate that is less than 1/300, 1/400, 1/500, 1/600, 1/700, 1/800, 1/900, 1/1000, 1/1250, 1/1500, 1/2000, 1/2500, 1/3000, 1/4000, 1/5000, 1/6000, 1/7000, 1/8000, 1/9000, 1/10000, 1/12000, 1/15000, 1/20000, 1/25000, 1/30000, 1/40000, 1/50000, 1/60000, 1/70000, 1/80000, 1/90000, 1/100000, 1/125000, 1/150000, 1/200000, 1/300000, 1/400000, 1/500000, 1/600000, 1/700000, 1/800000, 1/900000, 1/1000000, or less. In various instances, such error

optimized loci may comprise at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 30000, 50000, 75000, 100000, 500000, 1000000, 2000000, 3000000 or more loci. The error optimized loci may be distributed to at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 30000, 75000, 100000, 500000, 1000000, 2000000, 3000000 or more polynucleotides or genes.

[00122] The error rates can be achieved with or without error correction. The error rates can be achieved across the library, or across more than 80%, 85%, 90%, 93%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.8%, 99.9%, 99.95%, 99.98%, 99.99%, or more of the library.

[00123] Provided herein are structures that may comprise a surface that supports the synthesis of a plurality of nucleic acids having different predetermined sequences at addressable locations on a common support. In some instances, a device provides support for the synthesis of more than 2,000; 5,000; 10,000; 20,000; 30,000; 50,000; 75,000; 100,000; 200,000; 300,000; 400,000; 500,000; 600,000; 700,000; 800,000; 900,000; 1,000,000; 1,200,000; 1,400,000; 1,600,000; 1,800,000; 2,000,000; 2,500,000; 3,000,000; 3,500,000; 4,000,000; 4,500,000; 5,000,000; 10,000,000 or more non-identical nucleic acids. In some instances, the device provides support for the synthesis of more than 2,000; 5,000; 10,000; 20,000; 30,000; 50,000; 75,000; 100,000; 200,000; 300,000; 400,000; 500,000; 600,000; 700,000; 800,000; 900,000; 1,000,000; 1,200,000; 1,400,000; 1,600,000; 1,800,000; 2,000,000; 2,500,000; 3,000,000; 3,500,000; 4,000,000; 4,500,000; 5,000,000; 10,000,000 or more nucleic acids encoding for distinct sequences. In some instances, at least a portion of the nucleic acids have an identical sequence or are configured to be synthesized with an identical sequence.

[00124] Provided herein are methods and devices for manufacture and growth of nucleic acids about 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, or 2000 bases in length. In some instances, the length of the nucleic acid formed is about 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, or 225 bases in length. A nucleic acid may be at least 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 bases in length. A nucleic acid may be from 10 to 225 bases in length, from 12 to 100 bases in length, from 20 to 150 bases in length, from 20 to 130 bases in length, or from 30 to 100 bases in length.

[00125] In some instances, nucleic acids are synthesized on distinct loci of a substrate, wherein each locus supports the synthesis of a population of nucleic acids. In some instances, each locus

supports the synthesis of a population of nucleic acids having a different sequence than a population of nucleic acids grown on another locus. In some instances, the loci of a device are located within a plurality of clusters. In some instances, a device comprises at least 10, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 11000, 12000, 13000, 14000, 15000, 20000, 30000, 40000, 50000 or more clusters. In some instances, a device comprises more than 2,000; 5,000; 10,000; 100,000; 200,000; 300,000; 400,000; 500,000; 600,000; 700,000; 800,000; 900,000; 1,000,000; 1,100,000; 1,200,000; 1,300,000; 1,400,000; 1,500,000; 1,600,000; 1,700,000; 1,800,000; 1,900,000; 2,000,000; 300,000; 400,000; 500,000; 600,000; 700,000; 800,000; 900,000; 1,000,000; 1,200,000; 1,400,000; 1,600,000; 1,800,000; 2,000,000; 2,500,000; 3,000,000; 3,500,000; 4,000,000; 4,500,000; 5,000,000; or 10,000,000 or more distinct loci. In some instances, a device comprises about 10,000 distinct loci. The amount of loci within a single cluster is varied in different instances. In some instances, each cluster includes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 130, 150, 200, 300, 400, 500 or more loci. In some instances, each cluster includes about 50-500 loci. In some instances, each cluster includes about 100-200 loci. In some instances, each cluster includes about 100-150 loci. In some instances, each cluster includes about 109, 121, 130 or 137 loci. In some instances, each cluster includes about 19, 20, 61, 64 or more loci.

[00126] The number of distinct nucleic acids synthesized on a device may be dependent on the number of distinct loci available in the substrate. In some instances, the density of loci within a cluster of a device is at least or about 1 locus per mm^2 , 10 loci per mm^2 , 25 loci per mm^2 , 50 loci per mm^2 , 65 loci per mm^2 , 75 loci per mm^2 , 100 loci per mm^2 , 130 loci per mm^2 , 150 loci per mm^2 , 175 loci per mm^2 , 200 loci per mm^2 , 300 loci per mm^2 , 400 loci per mm^2 , 500 loci per mm^2 , 1,000 loci per mm^2 or more. In some instances, a device comprises from about 10 loci per mm^2 to about 500 mm^2 , from about 25 loci per mm^2 to about 400 mm^2 , from about 50 loci per mm^2 to about 500 mm^2 , from about 100 loci per mm^2 to about 500 mm^2 , from about 150 loci per mm^2 to about 500 mm^2 , from about 10 loci per mm^2 to about 250 mm^2 , from about 50 loci per mm^2 to about 250 mm^2 , from about 10 loci per mm^2 to about 200 mm^2 , or from about 50 loci per mm^2 to about 200 mm^2 . In some instances, the distance from the centers of two adjacent loci within a cluster is from about 10 μm to about 500 μm , from about 10 μm to about 200 μm , or from about 10 μm to about 100 μm . In some instances, the distance from two centers of adjacent loci is greater than about 10 μm , 20 μm , 30 μm , 40 μm , 50 μm , 60 μm , 70 μm , 80 μm , 90 μm or 100 μm . In some instances, the distance from the centers of two adjacent loci is less than about 200 μm , 150 μm , 100 μm , 80 μm , 70 μm , 60 μm , 50 μm , 40 μm , 30 μm , 20 μm or 10 μm . In some instances, each locus has a width of about 0.5 μm , 1 μm , 2 μm , 3 μm , 4 μm , 5 μm , 6 μm , 7 μm , 8 μm , 9 μm , 10 μm , 20 μm , 30 μm , 40

um, 50 um, 60 um, 70 um, 80 um, 90 um or 100 um. In some instances, each locus has a width of about 0.5 um to 100um, about 0.5 um to 50 um, about 10 um to 75 um, or about 0.5 um to 50 um.

[00127] In some instances, the density of clusters within a device is at least or about 1 cluster per 100 mm², 1 cluster per 10 mm², 1 cluster per 5 mm², 1 cluster per 4 mm², 1 cluster per 3 mm², 1 cluster per 2 mm², 1 cluster per 1 mm², 2 clusters per 1 mm², 3 clusters per 1 mm², 4 clusters per 1 mm², 5 clusters per 1 mm², 10 clusters per 1 mm², 50 clusters per 1 mm² or more. In some instances, a device comprises from about 1 cluster per 10 mm² to about 10 clusters per 1 mm². In some instances, the distance from the centers of two adjacent clusters is less than about 50 um, 100 um, 200 um, 500 um, 1000 um, or 2000 um or 5000 um. In some instances, the distance from the centers of two adjacent clusters is from about 50 um and about 100 um, from about 50 um and about 200 um, from about 50 um and about 300 um, from about 50 um and about 500 um, and from about 100 um to about 2000 um. In some instances, the distance from the centers of two adjacent clusters is from about 0.05 mm to about 50 mm, from about 0.05 mm to about 10 mm, from about 0.05 mm and about 5 mm, from about 0.05 mm and about 4 mm, from about 0.05 mm and about 3 mm, from about 0.05 mm and about 2 mm, from about 0.1 mm and 10 mm, from about 0.2 mm and 10 mm, from about 0.3 mm and about 10 mm, from about 0.4 mm and about 10 mm, from about 0.5 mm and 10 mm, from about 0.5 mm and about 5 mm, or from about 0.5 mm and about 2 mm. In some instances, each cluster has a diameter or width along one dimension of about 0.5 to 2 mm, about 0.5 to 1 mm, or about 1 to 2 mm. In some instances, each cluster has a diameter or width along one dimension of about 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9 or 2 mm. In some instances, each cluster has an interior diameter or width along one dimension of about 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9 or 2 mm.

[00128] A device may be about the size of a standard 96 well plate, for example from about 100 and 200 mm by from about 50 and 150 mm. In some instances, a device has a diameter less than or equal to about 1000 mm, 500 mm, 450 mm, 400 mm, 300 mm, 250 mm, 200 mm, 150 mm, 100 mm or 50 mm. In some instances, the diameter of a device is from about 25 mm and 1000 mm, from about 25 mm and about 800 mm, from about 25 mm and about 600 mm, from about 25 mm and about 500 mm, from about 25 mm and about 400 mm, from about 25 mm and about 300 mm, or from about 25 mm and about 200. Non-limiting examples of device size include about 300 mm, 200 mm, 150 mm, 130 mm, 100 mm, 76 mm, 51 mm and 25 mm. In some instances, a device has a planar surface area of at least about 100 mm²; 200 mm²; 500 mm²; 1,000 mm²; 2,000 mm²; 5,000 mm²; 10,000 mm²; 12,000 mm²; 15,000 mm²; 20,000 mm²; 30,000 mm²; 40,000 mm²; 50,000 mm² or more. In some instances, the thickness of a device is from about 50 mm and about 2000 mm, from about 50 mm and about 1000 mm, from about 100 mm and about 1000 mm, from about 200

mm and about 1000 mm, or from about 250 mm and about 1000 mm. Non-limiting examples of device thickness include 275 mm, 375 mm, 525 mm, 625 mm, 675 mm, 725 mm, 775 mm and 925 mm. In some instances, the thickness of a device varies with diameter and depends on the composition of the substrate. For example, a device comprising materials other than silicon has a different thickness than a silicon device of the same diameter. Device thickness may be determined by the mechanical strength of the material used and the device must be thick enough to support its own weight without cracking during handling. In some instances, a structure comprises a plurality of devices described herein.

[00129] Surface Materials

[00130] Provided herein is a device comprising a surface, wherein the surface is modified to support polynucleotide synthesis at predetermined locations and with a resulting low error rate, a low dropout rate, a high yield, and a high oligo representation. In some embodiments, surfaces of a device for polynucleotide synthesis provided herein are fabricated from a variety of materials capable of modification to support a de novo polynucleotide synthesis reaction. In some cases, the devices are sufficiently conductive, *e.g.*, are able to form uniform electric fields across all or a portion of the device. A device described herein may comprise a flexible material. Exemplary flexible materials include, without limitation, modified nylon, unmodified nylon, nitrocellulose, and polypropylene. A device described herein may comprise a rigid material. Exemplary rigid materials include, without limitation, glass, fused silica, silicon, silicon dioxide, silicon nitride, plastics (for example, polytetrafluoroethylene, polypropylene, polystyrene, polycarbonate, and blends thereof, and metals (for example, gold, platinum). Device disclosed herein may be fabricated from a material comprising silicon, polystyrene, agarose, dextran, cellulosic polymers, polyacrylamides, polydimethylsiloxane (PDMS), glass, or any combination thereof. In some cases, a device disclosed herein is manufactured with a combination of materials listed herein or any other suitable material known in the art.

[00131] A listing of tensile strengths for exemplary materials described herein is provided as follows: nylon (70 MPa), nitrocellulose (1.5 MPa), polypropylene (40 MPa), silicon (268 MPa), polystyrene (40 MPa), agarose (1-10 MPa), polyacrylamide (1-10 MPa), polydimethylsiloxane (PDMS) (3.9-10.8 MPa). Solid supports described herein can have a tensile strength from 1 to 300, 1 to 40, 1 to 10, 1 to 5, or 3 to 11 MPa. Solid supports described herein can have a tensile strength of about 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 20, 25, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 270, or more MPa. In some instances, a device described herein comprises a solid support for polynucleotide synthesis that is in the form of a flexible material capable of being stored in a continuous loop or reel, such as a tape or flexible sheet.

[00132] Young's modulus measures the resistance of a material to elastic (recoverable) deformation under load. A listing of Young's modulus for stiffness of exemplary materials described herein is provided as follows: nylon (3 GPa), nitrocellulose (1.5 GPa), polypropylene (2 GPa), silicon (150 GPa), polystyrene (3 GPa), agarose (1-10 GPa), polyacrylamide (1-10 GPa), polydimethylsiloxane (PDMS) (1-10 GPa). Solid supports described herein can have a Young's moduli from 1 to 500, 1 to 40, 1 to 10, 1 to 5, or 3 to 11 GPa. Solid supports described herein can have a Young's moduli of about 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 20, 25, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 400, 500 GPa, or more. As the relationship between flexibility and stiffness are inverse to each other, a flexible material has a low Young's modulus and changes its shape considerably under load.

[00133] In some cases, a device disclosed herein comprises a silicon dioxide base and a surface layer of silicon oxide. Alternatively, the device may have a base of silicon oxide. Surface of the device provided here may be textured, resulting in an increase overall surface area for polynucleotide synthesis. Device disclosed herein may comprise at least 5 %, 10%, 25%, 50%, 80%, 90%, 95%, or 99% silicon. A device disclosed herein may be fabricated from a silicon on insulator (SOI) wafer.

[00134] Substrates, devices and reactors provided herein are fabricated from any variety of materials suitable for the methods and compositions described herein. In certain instances, device materials are fabricated to exhibit a low level of nucleotide binding. In some instances, device materials are modified to generate distinct surfaces that exhibit a high level of nucleotide binding. In some instances, device materials are transparent to visible and/or UV light. In some instances, device materials are sufficiently conductive, *e.g.*, are able to form uniform electric fields across all or a portion of a substrate. In some instances, conductive materials are connected to an electric ground. In some instances, the device is heat conductive or insulated. In some instances, the materials are chemical resistant and heat resistant to support chemical or biochemical reactions, for example nucleic acid synthesis reaction processes. In some instances, a device comprises flexible materials. Flexible materials include, without limitation, modified nylon, unmodified nylon, nitrocellulose, polypropylene, and the like. In some instances, a device comprises rigid materials. Rigid materials include, without limitation, glass, fused silica, silicon, silicon dioxide, silicon nitride, plastics (for example, polytetrafluoroethylene, polypropylene, polystyrene, polycarbonate, and blends thereof, and the like), and metals (for example, gold, platinum, and the like). In some instances, a device is fabricated from a material comprising silicon, polystyrene, agarose, dextran, cellulosic polymers, polyacrylamides, polydimethylsiloxane (PDMS), glass, or any combination

thereof. In some instances, a device is manufactured with a combination of materials listed herein or any other suitable material known in the art.

[00135] Surface Architecture

[00136] Provided herein are devices comprising raised and/or lowered features. One benefit of having such features is an increase in surface area to support nucleic acid synthesis. In some instances, a device having raised and/or lowered features is referred to as a three-dimensional substrate. In some instances, a three-dimensional device comprises one or more channels. In some instances, one or more loci comprise a channel. In some instances, the channels are accessible to reagent deposition via a deposition device such as a nucleic acid synthesizer. In some instances, reagents and/or fluids collect in a larger well in fluid communication one or more channels. For example, a device comprises a plurality of channels corresponding to a plurality of loci with a cluster, and the plurality of channels are in fluid communication with one well of the cluster. In some methods, a library of nucleic acids is synthesized in a plurality of loci of a cluster.

[00137] In some instances, the structure is configured to allow for controlled flow and mass transfer paths for nucleic acid synthesis on a surface. In some instances, the configuration of a device allows for the controlled and even distribution of mass transfer paths, chemical exposure times, and/or wash efficacy during nucleic acid synthesis. In some instances, the configuration of a device allows for increased sweep efficiency, for example by providing sufficient volume for a growing a nucleic acid such that the excluded volume by the growing nucleic acid does not take up more than 50, 45, 40, 35, 30, 25, 20, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1%, or less of the initially available volume that is available or suitable for growing the nucleic acid. In some instances, a three-dimensional structure allows for managed flow of fluid to allow for the rapid exchange of chemical exposure.

[00138] Provided herein are methods to synthesize an amount of DNA of 1 fM, 5 fM, 10 fM, 25 fM, 50 fM, 75 fM, 100 fM, 200 fM, 300 fM, 400 fM, 500 fM, 600 fM, 700 fM, 800 fM, 900 fM, 1 pM, 5 pM, 10 pM, 25 pM, 50 pM, 75 pM, 100 pM, 200 pM, 300 pM, 400 pM, 500 pM, 600 pM, 700 pM, 800 pM, 900 pM, or more. In some instances, a polynucleotide library may span the length of about 1 %, 2 %, 3 %, 4 %, 5 %, 10 %, 15 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 90 %, 95 %, or 100 % of a gene. A gene may be varied up to about 1 %, 2 %, 3 %, 4 %, 5 %, 10 %, 15 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 85%, 90 %, 95 %, or 100 %.

[00139] Non-identical nucleic acids may collectively encode a sequence for at least 1 %, 2 %, 3 %, 4 %, 5 %, 10 %, 15 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 %, 80 %, 85%, 90 %, 95 %, or 100 % of a gene. In some instances, a nucleic acid may encode a sequence of 50 %, 60 %, 70 %, 80 %, 85%, 90 %, 95 %, or 100 %.

85%, 90 %, 95 %, or more of a gene. In some instances, a nucleic acid may encode a sequence of 80 %, 85%, 90 %, 95 %, or more of a gene.

[00140] In some instances, segregation is achieved by physical structure. In some instances, segregation is achieved by differential functionalization of the surface generating active and passive regions for nucleic acid synthesis. Differential functionalization is also achieved by alternating the hydrophobicity across the device surface, thereby creating water contact angle effects that cause beading or wetting of the deposited reagents. Employing larger structures can decrease splashing and cross-contamination of distinct nucleic acid synthesis locations with reagents of the neighboring spots. In some instances, a device, such as a nucleic acid synthesizer, is used to deposit reagents to distinct nucleic acid synthesis locations. Substrates having three-dimensional features are configured in a manner that allows for the synthesis of a large number of nucleic acids (*e.g.*, more than about 10,000) with a low error rate (*e.g.*, less than about 1:500, 1:1000, 1:1500, 1:2,000; 1:3,000; 1:5,000; or 1:10,000). In some instances, a device comprises features with a density of about or greater than about 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 300, 400 or 500 features per mm².

[00141] A well of a device may have the same or different width, height, and/or volume as another well of the substrate. A channel of a device may have the same or different width, height, and/or volume as another channel of the substrate. In some instances, the width of a cluster is from about 0.05 mm to about 50 mm, from about 0.05 mm to about 10 mm, from about 0.05 mm and about 5 mm, from about 0.05 mm and about 4 mm, from about 0.05 mm and about 3 mm, from about 0.05 mm and about 2 mm, from about 0.05 mm and about 1 mm, from about 0.05 mm and about 0.5 mm, from about 0.05 mm and about 0.1 mm, from about 0.1 mm and 10 mm, from about 0.2 mm and 10 mm, from about 0.3 mm and about 10 mm, from about 0.4 mm and about 10 mm, from about 0.5 mm and 10 mm, from about 0.5 mm and about 5 mm, or from about 0.5 mm and about 2 mm. In some instances, the width of a well comprising a cluster is from about 0.05 mm to about 50 mm, from about 0.05 mm to about 10 mm, from about 0.05 mm and about 5 mm, from about 0.05 mm and about 4 mm, from about 0.05 mm and about 3 mm, from about 0.05 mm and about 2 mm, from about 0.05 mm and about 1 mm, from about 0.05 mm and about 0.5 mm, from about 0.05 mm and about 0.1 mm, from about 0.1 mm and 10 mm, from about 0.2 mm and 10 mm, from about 0.3 mm and about 10 mm, from about 0.4 mm and about 10 mm, from about 0.5 mm and 10 mm, from about 0.5 mm and about 5 mm, or from about 0.5 mm and about 2 mm. In some instances, the width of a cluster is less than or about 5 mm, 4 mm, 3 mm, 2 mm, 1 mm, 0.5 mm, 0.1 mm, 0.09 mm, 0.08 mm, 0.07 mm, 0.06 mm or 0.05 mm. In some instances, the width of a cluster is from about 1.0 and 1.3 mm. In some instances, the width of a cluster is about 1.150 mm. In

some instances, the width of a well is less than or about 5 mm, 4 mm, 3 mm, 2 mm, 1 mm, 0.5 mm, 0.1 mm, 0.09 mm, 0.08 mm, 0.07 mm, 0.06 mm or 0.05 mm. In some instances, the width of a well is from about 1.0 and 1.3 mm. In some instances, the width of a well is about 1.150 mm. In some instances, the width of a cluster is about 0.08 mm. In some instances, the width of a well is about 0.08 mm. The width of a cluster may refer to clusters within a two-dimensional or three-dimensional substrate.

[00142] In some instances, the height of a well is from about 20 μm to about 1000 μm , from about 50 μm to about 1000 μm , from about 100 μm to about 1000 μm , from about 200 μm to about 1000 μm , from about 300 μm to about 1000 μm , from about 400 μm to about 1000 μm , or from about 500 μm to about 1000 μm . In some instances, the height of a well is less than about 1000 μm , less than about 900 μm , less than about 800 μm , less than about 700 μm , or less than about 600 μm .

[00143] In some instances, a device comprises a plurality of channels corresponding to a plurality of loci within a cluster, wherein the height or depth of a channel is from about 5 μm to about 500 μm , from about 5 μm to about 400 μm , from about 5 μm to about 300 μm , from about 5 μm to about 200 μm , from about 5 μm to about 100 μm , from about 5 μm to about 50 μm , or from about 10 μm to about 50 μm . In some instances, the height of a channel is less than 100 μm , less than 80 μm , less than 60 μm , less than 40 μm or less than 20 μm .

[00144] In some instances, the diameter of a channel, locus (*e.g.*, in a substantially planar substrate) or both channel and locus (*e.g.*, in a three-dimensional device wherein a locus corresponds to a channel) is from about 1 μm to about 1000 μm , from about 1 μm to about 500 μm , from about 1 μm to about 200 μm , from about 1 μm to about 100 μm , from about 5 μm to about 100 μm , or from about 10 μm to about 100 μm , for example, about 90 μm , 80 μm , 70 μm , 60 μm , 50 μm , 40 μm , 30 μm , 20 μm or 10 μm . In some instances, the diameter of a channel, locus, or both channel and locus is less than about 100 μm , 90 μm , 80 μm , 70 μm , 60 μm , 50 μm , 40 μm , 30 μm , 20 μm or 10 μm . In some instances, the distance from the center of two adjacent channels, loci, or channels and loci is from about 1 μm to about 500 μm , from about 1 μm to about 200 μm , from about 1 μm to about 100 μm , from about 5 μm to about 200 μm , from about 5 μm to about 100 μm , from about 5 μm to about 50 μm , or from about 5 μm to about 30 μm , for example, about 20 μm .

[00145] Surface Modifications

[00146] In various instances, surface modifications are employed for the chemical and/or physical alteration of a surface by an additive or subtractive process to change one or more chemical and/or physical properties of a device surface or a selected site or region of a device surface. For example, surface modifications include, without limitation, (1) changing the wetting properties of a surface, (2) functionalizing a surface, *i.e.*, providing, modifying or substituting

surface functional groups, (3) defunctionalizing a surface, i.e., removing surface functional groups, (4) otherwise altering the chemical composition of a surface, *e.g.*, through etching, (5) increasing or decreasing surface roughness, (6) providing a coating on a surface, *e.g.*, a coating that exhibits wetting properties that are different from the wetting properties of the surface, and/or (7) depositing particulates on a surface.

[00147] In some instances, the addition of a chemical layer on top of a surface (referred to as adhesion promoter) facilitates structured patterning of loci on a surface of a substrate. Exemplary surfaces for application of adhesion promotion include, without limitation, glass, silicon, silicon dioxide and silicon nitride. In some instances, the adhesion promoter is a chemical with a high surface energy. In some instances, a second chemical layer is deposited on a surface of a substrate. In some instances, the second chemical layer has a low surface energy. In some instances, surface energy of a chemical layer coated on a surface supports localization of droplets on the surface. Depending on the patterning arrangement selected, the proximity of loci and/or area of fluid contact at the loci are alterable.

[00148] In some instances, a device surface, or resolved loci, onto which nucleic acids or other moieties are deposited, *e.g.*, for nucleic acid synthesis, are smooth or substantially planar (*e.g.*, two-dimensional) or have irregularities, such as raised or lowered features (*e.g.*, three-dimensional features). In some instances, a device surface is modified with one or more different layers of compounds. Such modification layers of interest include, without limitation, inorganic and organic layers such as metals, metal oxides, polymers, small organic molecules and the like. Non-limiting polymeric layers include peptides, proteins, nucleic acids or mimetics thereof (*e.g.*, peptide nucleic acids and the like), polysaccharides, phospholipids, polyurethanes, polyesters, polycarbonates, polyureas, polyamides, polyethyleneamines, polyarylene sulfides, polysiloxanes, polyimides, polyacetates, and any other suitable compounds described herein or otherwise known in the art. In some instances, polymers are heteropolymeric. In some instances, polymers are homopolymeric. In some instances, polymers comprise functional moieties or are conjugated.

[00149] In some instances, resolved loci of a device are functionalized with one or more moieties that increase and/or decrease surface energy. In some instances, a moiety is chemically inert. In some instances, a moiety is configured to support a desired chemical reaction, for example, one or more processes in a nucleic acid synthesis reaction. The surface energy, or hydrophobicity, of a surface is a factor for determining the affinity of a nucleotide to attach onto the surface. In some instances, a method for device functionalization may comprise: (a) providing a device having a surface that comprises silicon dioxide; and (b) silanizing the surface using, a

suitable silanizing agent described herein or otherwise known in the art, for example, an organofunctional alkoxysilane molecule.

[00150] In some instances, the organofunctional alkoxysilane molecule comprises dimethylchloro-octadecyl-silane, methylchloro-octadecyl-silane, trichloro-octadecyl-silane, trimethyl-octadecyl-silane, triethyl-octadecyl-silane, or any combination thereof. In some instances, a device surface comprises functionalized with polyethylene/polypropylene (functionalized by gamma irradiation or chromic acid oxidation, and reduction to hydroxyalkyl surface), highly crosslinked polystyrene-divinylbenzene (derivatized by chloromethylation, and aminated to benzylamine functional surface), nylon (the terminal aminohexyl groups are directly reactive), or etched with reduced polytetrafluoroethylene. Other methods and functionalizing agents are described in U.S. Patent No. 5474796, which is herein incorporated by reference in its entirety.

[00151] In some instances, a device surface is functionalized by contact with a derivatizing composition that contains a mixture of silanes, under reaction conditions effective to couple the silanes to the device surface, typically via reactive hydrophilic moieties present on the device surface. Silanization generally covers a surface through self-assembly with organofunctional alkoxysilane molecules.

[00152] A variety of siloxane functionalizing reagents can further be used as currently known in the art, *e.g.*, for lowering or increasing surface energy. The organofunctional alkoxysilanes can be classified according to their organic functions.

[00153] Provided herein are devices that may contain patterning of agents capable of coupling to a nucleoside. In some instances, a device may be coated with an active agent. In some instances, a device may be coated with a passive agent. Exemplary active agents for inclusion in coating materials described herein includes, without limitation, N-(3-triethoxysilylpropyl)-4-hydroxybutyramide (HAPS), 11-acetoxundecyltriethoxysilane, n-decyltriethoxysilane, (3-aminopropyl)trimethoxysilane, (3-aminopropyl)triethoxysilane, 3-glycidoxypropyltrimethoxysilane (GOPS), 3-iodo-propyltrimethoxysilane, butyl-aldehyde-trimethoxysilane, dimeric secondary aminoalkyl siloxanes, (3-aminopropyl)-diethoxy-methylsilane, (3-aminopropyl)-dimethyl-ethoxysilane, and (3-aminopropyl)-trimethoxysilane, (3-glycidoxypropyl)-dimethyl-ethoxysilane, glycidoxy-trimethoxysilane, (3-mercaptopropyl)-trimethoxysilane, 3-4 epoxycyclohexyl-ethyltrimethoxysilane, and (3-mercaptopropyl)-methyl-dimethoxysilane, allyl trichlorochlorosilane, 7-oct-1-enyl trichlorochlorosilane, or bis (3-trimethoxysilylpropyl) amine.

[00154] Exemplary passive agents for inclusion in a coating material described herein includes, without limitation, perfluorooctyltrichlorosilane; tridecafluoro-1,1,2,2-

tetrahydrooctyl)trichlorosilane; 1H, 1H, 2H, 2H-fluorooctyltriethoxysilane (FOS); trichloro(1H, 1H, 2H, 2H - perfluorooctyl)silane; tert-butyl-[5-fluoro-4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)indol-1-yl]-dimethyl-silane; CYTOP™; Fluorinert™; perfluorooctyltrichlorosilane (PFOTCS); perfluorooctyldimethylchlorosilane (PFODCS); perfluorodecyltriethoxysilane (PFDTES); pentafluorophenyl-dimethylpropylchloro-silane (PFPTES); perfluorooctyltriethoxysilane; perfluorooctyltrimethoxysilane; octylchlorosilane; dimethylchloro-octodecyl-silane; methyldichloro-octodecyl-silane; trichloro-octodecyl-silane; trimethyl-octodecyl-silane; triethyl-octodecyl-silane; or octadecyltrichlorosilane.

[00155] In some instances, a functionalization agent comprises a hydrocarbon silane such as octadecyltrichlorosilane. In some instances, the functionalizing agent comprises 11-acetoxyundecyltriethoxysilane, n-decyltriethoxysilane, (3-aminopropyl)trimethoxysilane, (3-aminopropyl)triethoxysilane, glycidyloxypropyl/trimethoxysilane and N-(3-triethoxysilylpropyl)-4-hydroxybutyramide.

[00156] Polynucleotide Synthesis

[00157] Methods of the current disclosure for nucleic acid synthesis may include processes involving phosphoramidite chemistry. In some instances, nucleic acid synthesis comprises coupling a base with phosphoramidite. Nucleic acid synthesis may comprise coupling a base by deposition of phosphoramidite under coupling conditions, wherein the same base is optionally deposited with phosphoramidite more than once, i.e., double coupling. Nucleic acid synthesis may comprise capping of unreacted sites. In some instances, capping is optional. Nucleic acid synthesis may also comprise oxidation or an oxidation step or oxidation steps. Nucleic acid synthesis may comprise deblocking, detritylation, and sulfurization. In some instances, nucleic acid synthesis comprises either oxidation or sulfurization. In some instances, between one or each step during a nucleic acid synthesis reaction, the device is washed, for example, using tetrazole or acetonitrile. Time frames for any one step in a phosphoramidite synthesis method may be less than about 2 min, 1 min, 50 sec, 40 sec, 30 sec, 20 sec and 10 sec.

[00158] Nucleic acid synthesis using a phosphoramidite method may comprise a subsequent addition of a phosphoramidite building block (*e.g.*, nucleoside phosphoramidite) to a growing nucleic acid chain for the formation of a phosphite triester linkage. Phosphoramidite nucleic acid synthesis proceeds in the 3' to 5' direction. Phosphoramidite nucleic acid synthesis allows for the controlled addition of one nucleotide to a growing nucleic acid chain per synthesis cycle. In some instances, each synthesis cycle comprises a coupling step. Phosphoramidite coupling involves the formation of a phosphite triester linkage between an activated nucleoside phosphoramidite and a nucleoside bound to the substrate, for example, via a linker. In some instances, the nucleoside

phosphoramidite is provided to the device activated. In some instances, the nucleoside phosphoramidite is provided to the device with an activator. In some instances, nucleoside phosphoramidites are provided to the device in a 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100-fold excess or more over the substrate-bound nucleosides. In some instances, the addition of nucleoside phosphoramidite is performed in an anhydrous environment, for example, in anhydrous acetonitrile. Following addition of a nucleoside phosphoramidite, the device is optionally washed. In some instances, the coupling step is repeated one or more additional times, optionally with a wash step between nucleoside phosphoramidite additions to the substrate. In some instances, a nucleic acid synthesis method used herein comprises 1, 2, 3 or more sequential coupling steps. Prior to coupling, in many cases, the nucleoside bound to the device is de-protected by removal of a protecting group, where the protecting group functions to prevent polymerization. A common protecting group is 4,4'-dimethoxytrityl (DMT).

[00159] Following coupling, phosphoramidite nucleic acid synthesis methods optionally comprise a capping step. In a capping step, the growing nucleic acid is treated with a capping agent. A capping step is useful to block unreacted substrate-bound 5'-OH groups after coupling from further chain elongation, preventing the formation of nucleic acids with internal base deletions. Further, phosphoramidites activated with 1H-tetrazole may react, to a small extent, with the O6 position of guanosine. Without being bound by theory, upon oxidation with I₂/water, this side product, possibly via O6-N7 migration, may undergo depurination. The apurinic sites may end up being cleaved in the course of the final deprotection of the polynucleotide thus reducing the yield of the full-length product. The O6 modifications may be removed by treatment with the capping reagent prior to oxidation with I₂/water. In some instances, inclusion of a capping step during nucleic acid synthesis decreases the error rate as compared to synthesis without capping. As an example, the capping step comprises treating the substrate-bound nucleic acid with a mixture of acetic anhydride and 1-methylimidazole. Following a capping step, the device is optionally washed.

[00160] In some instances, following addition of a nucleoside phosphoramidite, and optionally after capping and one or more wash steps, the device bound growing nucleic acid is oxidized. The oxidation step comprises the phosphite triester is oxidized into a tetracoordinated phosphate triester, a protected precursor of the naturally occurring phosphate diester internucleoside linkage. In some instances, oxidation of the growing nucleic acid is achieved by treatment with iodine and water, optionally in the presence of a weak base (*e.g.*, pyridine, lutidine, collidine). Oxidation may be carried out under anhydrous conditions using, *e.g.* tert-Butyl hydroperoxide or (1S)-(+)-(10-

camphorsulfonyl)-oxaziridine (CSO). In some methods, a capping step is performed following oxidation. A second capping step allows for device drying, as residual water from oxidation that may persist can inhibit subsequent coupling. Following oxidation, the device and growing nucleic acid is optionally washed. In some instances, the step of oxidation is substituted with a sulfurization step to obtain polynucleotide phosphorothioates, wherein any capping steps can be performed after the sulfurization. Many reagents are capable of the efficient sulfur transfer, including but not limited to 3-(Dimethylaminomethylidene)amino)-3H-1,2,4-dithiazole-3-thione, DDTT, 3H-1,2-benzodithiol-3-one 1,1-dioxide, also known as Beaucage reagent, and N,N,N'N'-Tetraethylthiuram disulfide (TETD).

[00161] In order for a subsequent cycle of nucleoside incorporation to occur through coupling, the protected 5' end of the device bound growing nucleic acid is removed so that the primary hydroxyl group is reactive with a next nucleoside phosphoramidite. In some instances, the protecting group is DMT and deblocking occurs with trichloroacetic acid in dichloromethane. Conducting detritylation for an extended time or with stronger than recommended solutions of acids may lead to increased depurination of solid support-bound polynucleotide and thus reduces the yield of the desired full-length product. Methods and compositions of the disclosure described herein provide for controlled deblocking conditions limiting undesired depurination reactions. In some instances, the device bound nucleic acid is washed after deblocking. In some instances, efficient washing after deblocking contributes to synthesized nucleic acids having a low error rate.

[00162] Methods for the synthesis of nucleic acids typically involve an iterating sequence of the following steps: application of a protected monomer to an actively functionalized surface (*e.g.*, locus) to link with either the activated surface, a linker or with a previously deprotected monomer; deprotection of the applied monomer so that it is reactive with a subsequently applied protected monomer; and application of another protected monomer for linking. One or more intermediate steps include oxidation or sulfurization. In some instances, one or more wash steps precede or follow one or all of the steps.

[00163] Methods for phosphoramidite-based nucleic acid synthesis comprise a series of chemical steps. In some instances, one or more steps of a synthesis method involve reagent cycling, where one or more steps of the method comprise application to the device of a reagent useful for the step. For example, reagents are cycled by a series of liquid deposition and vacuum drying steps. For substrates comprising three-dimensional features such as wells, microwells, channels and the like, reagents are optionally passed through one or more regions of the device via the wells and/or channels.

[00164] Methods and systems described herein relate to polynucleotide synthesis devices for the synthesis of polynucleotides. The synthesis may be in parallel. For example, at least or about at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 850, 900, 1000, 10000, 50000, 75000, 100000 or more polynucleotides can be synthesized in parallel. The total number nucleic acids that may be synthesized in parallel may be from 2-100000, 3-50000, 4-10000, 5-1000, 6-900, 7-850, 8-800, 9-750, 10-700, 11-650, 12-600, 13-550, 14-500, 15-450, 16-400, 17-350, 18-300, 19-250, 20-200, 21-150, 22-100, 23-50, 24-45, 25-40, 30-35. Those of skill in the art appreciate that the total number of polynucleotides synthesized in parallel may fall within any range bound by any of these values, for example 25-100. The total number of polynucleotides synthesized in parallel may fall within any range defined by any of the values serving as endpoints of the range. Total molar mass of polynucleotides synthesized within the device or the molar mass of each of the polynucleotides may be at least or at least about 10, 20, 30, 40, 50, 100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 25000, 50000, 75000, 100000 picomoles, or more. The length of each of the polynucleotides or average length of the polynucleotides within the device may be at least or about at least 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 150, 200, 300, 400, 500 nucleotides, or more. The length of each of the polynucleotides or average length of the polynucleotides within the device may be at most or about at most 500, 400, 300, 200, 150, 100, 50, 45, 35, 30, 25, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10 nucleotides, or less. The length of each of the polynucleotides or average length of the polynucleotides within the device may fall from 10-500, 9-400, 11-300, 12-200, 13-150, 14-100, 15-50, 16-45, 17-40, 18-35, 19-25. Those of skill in the art appreciate that the length of each of the polynucleotides or average length of the polynucleotides within the device may fall within any range bound by any of these values, for example 100-300. The length of each of the polynucleotides or average length of the polynucleotides within the device may fall within any range defined by any of the values serving as endpoints of the range.

[00165] Methods for nucleic acid synthesis on a surface provided herein allow for synthesis at a fast rate. As an example, at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, 100, 125, 150, 175, 200 nucleotides per hour, or more are synthesized. Nucleotides include adenine, guanine, thymine, cytosine, uridine building blocks, or analogs/modified versions thereof. In some instances, libraries of nucleic acids are synthesized in parallel on substrate. For example, a device comprising about or at least about 100; 1,000; 10,000; 30,000; 75,000; 100,000; 1,000,000; 2,000,000; 3,000,000; 4,000,000; or 5,000,000 resolved loci is able to support the synthesis of at least the same number of

distinct nucleic acids, wherein nucleic acid encoding a distinct sequence is synthesized on a resolved locus. In some instances, a library of nucleic acids is synthesized on a device with low error rates described herein in less than about three months, two months, one month, three weeks, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2 days, 24 hours or less. In some instances, larger nucleic acids assembled from a nucleic acid library synthesized with low error rate using the substrates and methods described herein are prepared in less than about three months, two months, one month, three weeks, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2 days, 24 hours or less.

[00166] In some instances, methods described herein provide for generation of a library of nucleic acids comprising variant nucleic acids differing at a plurality of codon sites. In some instances, a nucleic acid may have 1 site, 2 sites, 3 sites, 4 sites, 5 sites, 6 sites, 7 sites, 8 sites, 9 sites, 10 sites, 11 sites, 12 sites, 13 sites, 14 sites, 15 sites, 16 sites, 17 sites, 18 sites, 19 sites, 20 sites, 30 sites, 40 sites, 50 sites, or more of variant codon sites.

[00167] In some instances, the one or more sites of variant codon sites may be adjacent. In some instances, the one or more sites of variant codon sites may not be adjacent and separated by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more codons.

[00168] In some instances, a nucleic acid may comprise multiple sites of variant codon sites, wherein all the variant codon sites are adjacent to one another, forming a stretch of variant codon sites. In some instances, a nucleic acid may comprise multiple sites of variant codon sites, wherein none the variant codon sites are adjacent to one another. In some instances, a nucleic acid may comprise multiple sites of variant codon sites, wherein some the variant codon sites are adjacent to one another, forming a stretch of variant codon sites, and some of the variant codon sites are not adjacent to one another.

[00169] Referring to the Figures, **FIG. 12** illustrates an exemplary process workflow for synthesis of nucleic acids (e.g., genes) from shorter nucleic acids. The workflow is divided generally into phases: (1) de novo synthesis of a single stranded nucleic acid library, (2) joining nucleic acids to form larger fragments, (3) error correction, (4) quality control, and (5) shipment. Prior to de novo synthesis, an intended nucleic acid sequence or group of nucleic acid sequences is preselected. For example, a group of genes is preselected for generation.

[00170] Once large nucleic acids for generation are selected, a predetermined library of nucleic acids is designed for de novo synthesis. Various suitable methods are known for generating high density nucleic acid arrays. In the workflow example, a device surface layer **1201** is provided. In the example, chemistry of the surface is altered in order to improve the nucleic acid synthesis process. Areas of low surface energy are generated to repel liquid while areas of high surface energy are generated to attract liquids. The surface itself may be in the form of a planar surface or

contain variations in shape, such as protrusions or microwells which increase surface area. In the workflow example, high surface energy molecules selected serve a dual function of supporting DNA chemistry, as disclosed in International Patent Application Publication WO/2015/021080, which is herein incorporated by reference in its entirety.

[00171] In situ preparation of nucleic acid arrays is generated on a solid support and utilizes single nucleotide extension process to extend multiple oligomers in parallel. A material deposition device, such as a nucleic acid synthesizer, is designed to release reagents in a step wise fashion such that multiple nucleic acids extend, in parallel, one residue at a time to generate oligomers with a predetermined nucleic acid sequence **1202**. In some instances, nucleic acids are cleaved from the surface at this stage. Cleavage includes gas cleavage, *e.g.*, with ammonia or methylamine.

[00172] The generated nucleic acid libraries are placed in a reaction chamber. In this exemplary workflow, the reaction chamber (also referred to as “nanoreactor”) is a silicon coated well, containing PCR reagents and lowered onto the nucleic acid library **1203**. Prior to or after the sealing **1204** of the nucleic acids, a reagent is added to release the nucleic acids from the substrate. In the exemplary workflow, the nucleic acids are released subsequent to sealing of the nanoreactor **1205**. Once released, fragments of single stranded nucleic acids hybridize in order to span an entire long range sequence of DNA. Partial hybridization **1205** is possible because each synthesized nucleic acid is designed to have a small portion overlapping with at least one other nucleic acid in the pool.

[00173] After hybridization, a PCA reaction is commenced. During the polymerase cycles, the nucleic acids anneal to complementary fragments and gaps are filled in by a polymerase. Each cycle increases the length of various fragments randomly depending on which nucleic acids find each other. Complementarity amongst the fragments allows for forming a complete large span of double stranded DNA **1206**.

[00174] After PCA is complete, the nanoreactor is separated from the device **1207** and positioned for interaction with a device having primers for PCR **1208**. After sealing, the nanoreactor is subject to PCR **1209** and the larger nucleic acids are amplified. After PCR **1210**, the nanochamber is opened **1211**, error correction reagents are added **1212**, the chamber is sealed **1213** and an error correction reaction occurs to remove mismatched base pairs and/or strands with poor complementarity from the double stranded PCR amplification products **1214**. The nanoreactor is opened and separated **1215**. Error corrected product is next subject to additional processing steps, such as PCR and molecular bar coding, and then packaged **1222** for shipment **1223**.

[00175] In some instances, quality control measures are taken. After error correction, quality control steps include for example interaction with a wafer having sequencing primers for

amplification of the error corrected product **1216**, sealing the wafer to a chamber containing error corrected amplification product **1217**, and performing an additional round of amplification **1218**. The nanoreactor is opened **1219** and the products are pooled **1220** and sequenced **1221**. After an acceptable quality control determination is made, the packaged product **1222** is approved for shipment **1223**.

[00176] In some instances, a nucleic acid generate by a workflow such as that in **FIG. 12** is subject to mutagenesis using overlapping primers disclosed herein. In some instances, a library of primers are generated by in situ preparation on a solid support and utilize single nucleotide extension process to extend multiple oligomers in parallel. A deposition device, such as a nucleic acid synthesizer, is designed to release reagents in a step wise fashion such that multiple nucleic acids extend, in parallel, one residue at a time to generate oligomers with a predetermined nucleic acid sequence **1202**.

[00177] Computer systems

[00178] Any of the systems described herein, may be operably linked to a computer and may be automated through a computer either locally or remotely. In various instances, the methods and systems of the disclosure may further comprise software programs on computer systems and use thereof. Accordingly, computerized control for the synchronization of the dispense/vacuum/refill functions such as orchestrating and synchronizing the material deposition device movement, dispense action and vacuum actuation are within the bounds of the disclosure. The computer systems may be programmed to interface between the user specified base sequence and the position of a material deposition device to deliver the correct reagents to specified regions of the substrate.

[00179] The computer system **1300** illustrated in **FIG. 13** may be understood as a logical apparatus that can read instructions from media **1311** and/or a network port **1305**, which can optionally be connected to server **1309** having fixed media **1312**. The system, such as shown in **FIG. 13** can include a CPU **1301**, disk drives **1303**, optional input devices such as keyboard **1315** and/or mouse **1316** and optional monitor **1307**. Data communication can be achieved through the indicated communication medium to a server at a local or a remote location. The communication medium can include any means of transmitting and/or receiving data. For example, the communication medium can be a network connection, a wireless connection or an internet connection. Such a connection can provide for communication over the World Wide Web. It is envisioned that data relating to the present disclosure can be transmitted over such networks or connections for reception and/or review by a party **1322** as illustrated in **FIG. 13**.

[00180] **FIG. 14** is a block diagram illustrating a first example architecture of a computer system **1400** that can be used in connection with example instances of the present disclosure. As depicted

in **FIG. 14**, the example computer system can include a processor **1402** for processing instructions. Non-limiting examples of processors include: Intel Xeon™ processor, AMD Opteron™ processor, Samsung 32-bit RISC ARM 1176JZ(F)-S v1.0™ processor, ARM Cortex-A8 Samsung S5PC100™ processor, ARM Cortex-A8 Apple A4™ processor, Marvell PXA 930™ processor, or a functionally-equivalent processor. Multiple threads of execution can be used for parallel processing. In some instances, multiple processors or processors with multiple cores can also be used, whether in a single computer system, in a cluster, or distributed across systems over a network comprising a plurality of computers, cell phones, and/or personal data assistant devices.

[00181] As illustrated in **FIG. 14**, a high speed cache **1404** can be connected to, or incorporated in, the processor **1402** to provide a high speed memory for instructions or data that have been recently, or are frequently, used by processor **1402**. The processor **1402** is connected to a north bridge **1406** by a processor bus **1408**. The north bridge **1406** is connected to random access memory (RAM) **1410** by a memory bus **1412** and manages access to the RAM **1410** by the processor **1402**. The north bridge **1406** is also connected to a south bridge **1414** by a chipset bus **1416**. The south bridge **1414** is, in turn, connected to a peripheral bus **1418**. The peripheral bus can be, for example, PCI, PCI-X, PCI Express, or other peripheral bus. The north bridge and south bridge are often referred to as a processor chipset and manage data transfer between the processor, RAM, and peripheral components on the peripheral bus **1418**. In some alternative architectures, the functionality of the north bridge can be incorporated into the processor instead of using a separate north bridge chip. In some instances, system **1400** can include an accelerator card **1422** attached to the peripheral bus **1418**. The accelerator can include field programmable gate arrays (FPGAs) or other hardware for accelerating certain processing. For example, an accelerator can be used for adaptive data restructuring or to evaluate algebraic expressions used in extended set processing.

[00182] Software and data are stored in external storage **1424** and can be loaded into RAM **1410** and/or cache **1404** for use by the processor. The system **1400** includes an operating system for managing system resources; non-limiting examples of operating systems include: Linux, Windows™, MACOS™, BlackBerry OST™, iOS™, and other functionally-equivalent operating systems, as well as application software running on top of the operating system for managing data storage and optimization in accordance with example instances of the present disclosure. In this example, system **1400** also includes network interface cards (NICs) **1420** and **1421** connected to the peripheral bus for providing network interfaces to external storage, such as Network Attached Storage (NAS) and other computer systems that can be used for distributed parallel processing.

[00183] FIG. 15 is a diagram showing a network 1500 with a plurality of computer systems 1502a, and 1502b, a plurality of cell phones and personal data assistants 1502c, and Network Attached Storage (NAS) 1504a, and 1504b. In example instances, systems 1502a, 1502b, and 1502c can manage data storage and optimize data access for data stored in Network Attached Storage (NAS) 1504a and 1504b. A mathematical model can be used for the data and be evaluated using distributed parallel processing across computer systems 1502a, and 1502b, and cell phone and personal data assistant systems 1502c. Computer systems 1502a, and 1502b, and cell phone and personal data assistant systems 1502c can also provide parallel processing for adaptive data restructuring of the data stored in Network Attached Storage (NAS) 1504a and 1504b. FIG. 15 illustrates an example only, and a wide variety of other computer architectures and systems can be used in conjunction with the various instances of the present disclosure. For example, a blade server can be used to provide parallel processing. Processor blades can be connected through a back plane to provide parallel processing. Storage can also be connected to the back plane or as Network Attached Storage (NAS) through a separate network interface. In some example instances, processors can maintain separate memory spaces and transmit data through network interfaces, back plane or other connectors for parallel processing by other processors. In other instances, some or all of the processors can use a shared virtual address memory space.

[00184] FIG. 16 is a block diagram of a multiprocessor computer system 1600 using a shared virtual address memory space in accordance with an example instance. The system includes a plurality of processors 1602a-f that can access a shared memory subsystem 1604. The system incorporates a plurality of programmable hardware memory algorithm processors (MAPs) 1606a-f in the memory subsystem 1604. Each MAP 1606a-f can comprise a memory 1608a-f and one or more field programmable gate arrays (FPGAs) 1610a-f. The MAP provides a configurable functional unit and particular algorithms or portions of algorithms can be provided to the FPGAs 1610a-f for processing in close coordination with a respective processor. For example, the MAPs can be used to evaluate algebraic expressions regarding the data model and to perform adaptive data restructuring in example instances. In this example, each MAP is globally accessible by all of the processors for these purposes. In one configuration, each MAP can use Direct Memory Access (DMA) to access an associated memory 1608a-f, allowing it to execute tasks independently of, and asynchronously from the respective microprocessor 1602a-f. In this configuration, a MAP can feed results directly to another MAP for pipelining and parallel execution of algorithms.

[00185] The above computer architectures and systems are examples only, and a wide variety of other computer, cell phone, and personal data assistant architectures and systems can be used in connection with example instances, including systems using any combination of general

processors, co-processors, FPGAs and other programmable logic devices, system on chips (SOCs), application specific integrated circuits (ASICs), and other processing and logic elements. In some instances, all or part of the computer system can be implemented in software or hardware. Any variety of data storage media can be used in connection with example instances, including random access memory, hard drives, flash memory, tape drives, disk arrays, Network Attached Storage (NAS) and other local or distributed data storage devices and systems.

[00186] In example instances, the computer system can be implemented using software modules executing on any of the above or other computer architectures and systems. In other instances, the functions of the system can be implemented partially or completely in firmware, programmable logic devices such as field programmable gate arrays (FPGAs) as referenced in **FIG. 13**, system on chips (SOCs), application specific integrated circuits (ASICs), or other processing and logic elements. For example, the Set Processor and Optimizer can be implemented with hardware acceleration through the use of a hardware accelerator card, such as accelerator card **1322** illustrated in **FIG. 13**.

[00187] The following examples are set forth to illustrate more clearly the principle and practice of embodiments disclosed herein to those skilled in the art and are not to be construed as limiting the scope of any claimed embodiments. Unless otherwise stated, all parts and percentages are on a weight basis.

EXAMPLES

[00188] The following examples are given for the purpose of illustrating various embodiments of the disclosure and are not meant to limit the present disclosure in any fashion. The present examples, along with the methods described herein are presently representative of preferred embodiments, are exemplary, and are not intended as limitations on the scope of the disclosure. Changes therein and other uses which are encompassed within the spirit of the disclosure as defined by the scope of the claims will occur to those skilled in the art.

[00189] Example 1: Functionalization of a device surface

[00190] A device was functionalized to support the attachment and synthesis of a library of nucleic acids. The device surface was first wet cleaned using a piranha solution comprising 90% H_2SO_4 and 10% H_2O_2 for 20 minutes. The device was rinsed in several beakers with DI water, held under a DI water gooseneck faucet for 5 min, and dried with N_2 . The device was subsequently soaked in NH_4OH (1:100; 3 mL:300 mL) for 5 min, rinsed with DI water using a handgun, soaked in three successive beakers with DI water for 1 min each, and then rinsed again with DI water using the handgun. The device was then plasma cleaned by exposing the device surface to O_2 . A

SAMCO PC-300 instrument was used to plasma etch O₂ at 250 watts for 1 min in downstream mode.

[00191] The cleaned device surface was actively functionalized with a solution comprising N-(3-triethoxysilylpropyl)-4-hydroxybutyramide using a YES-1224P vapor deposition oven system with the following parameters: 0.5 to 1 torr, 60 min, 70 °C, 135 °C vaporizer. The device surface was resist coated using a Brewer Science 200X spin coater. SPRTM 3612 photoresist was spin coated on the device at 2500 rpm for 40 sec. The device was pre-baked for 30 min at 90 °C on a Brewer hot plate. The device was subjected to photolithography using a Karl Suss MA6 mask aligner instrument. The device was exposed for 2.2 sec and developed for 1 min in MSF 26A. Remaining developer was rinsed with the handgun and the device soaked in water for 5 min. The device was baked for 30 min at 100 °C in the oven, followed by visual inspection for lithography defects using a Nikon L200. A descum process was used to remove residual resist using the SAMCO PC-300 instrument to O₂ plasma etch at 250 watts for 1 min.

[00192] The device surface was passively functionalized with a 100 µL solution of perfluorooctyltrichlorosilane mixed with 10 µL light mineral oil. The device was placed in a chamber, pumped for 10 min, and then the valve was closed to the pump and left to stand for 10 min. The chamber was vented to air. The device was resist stripped by performing two soaks for 5 min in 500 mL NMP at 70 °C with ultrasonication at maximum power (9 on Crest system). The device was then soaked for 5 min in 500 mL isopropanol at room temperature with ultrasonication at maximum power. The device was dipped in 300 mL of 200 proof ethanol and blown dry with N₂. The functionalized surface was activated to serve as a support for nucleic acid synthesis.

[00193] Example 2: Synthesis of a 50-mer sequence on a polynucleotide synthesis device

[00194] A two dimensional polynucleotide synthesis device was assembled into a flowcell, which was connected to a flowcell (Applied Biosystems (ABI394 DNA Synthesizer"). The two-dimensional polynucleotide synthesis device was uniformly functionalized with N-(3-TRIETHOXYSYLYLPROPYL)-4-HYDROXYBUTYRAMIDE (Gelest) was used to synthesize an exemplary polynucleotide of 50 bp ("50-mer polynucleotide") using polynucleotide synthesis methods described herein.

[00195] The sequence of the 50-mer was as described in SEQ ID NO.: 20.

5'AGACAATCAACCATTTGGGGTGGACAGCCTTGACCTCTAGACTTCGGCAT##TTTTTTT TTT3' (SEQ ID NO.: 20), where # denotes Thymidine-succinyl hexamide CED phosphoramidite (CLP-2244 from ChemGenes), which is a cleavable linker enabling the release of oligos from the surface during deprotection.

[00196] The synthesis was done using standard DNA synthesis chemistry (coupling, capping, oxidation, and deblocking) according to the protocol in **Table 6** and an ABI synthesizer.

Table 6. Synthesis protocol

Table 6		
General DNA Synthesis Process Name	Process Step	Time (sec)
WASH (Acetonitrile Wash Flow)	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	23
	N2 System Flush	4
	Acetonitrile System Flush	4
DNA BASE ADDITION (Phosphoramidite + Activator Flow)	Activator Manifold Flush	2
	Activator to Flowcell	6
	Activator + Phosphoramidite to Flowcell	6
	Activator to Flowcell	0.5
	Activator + Phosphoramidite to Flowcell	5
	Activator to Flowcell	0.5
	Activator + Phosphoramidite to Flowcell	5
	Activator to Flowcell	0.5
	Activator + Phosphoramidite to Flowcell	5
	Incubate for 25sec	25
WASH (Acetonitrile Wash Flow)	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	15
	N2 System Flush	4
	Acetonitrile System Flush	4
DNA BASE ADDITION (Phosphoramidite + Activator Flow)	Activator Manifold Flush	2
	Activator to Flowcell	5
	Activator + Phosphoramidite to Flowcell	18
	Incubate for 25sec	25
WASH (Acetonitrile Wash Flow)	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	15
	N2 System Flush	4
	Acetonitrile System Flush	4
CAPPING (CapA+B, 1:1, Flow)	CapA+B to Flowcell	15
WASH (Acetonitrile Wash Flow)	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	15
	Acetonitrile System Flush	4
OXIDATION (Oxidizer)	Oxidizer to Flowcell	18

Table 6		
General DNA Synthesis Process Name	Process Step	Time (sec)
Flow)		
WASH (Acetonitrile Wash Flow)	Acetonitrile System Flush	4
	N2 System Flush	4
	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	15
	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	15
	N2 System Flush	4
	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	23
	N2 System Flush	4
	Acetonitrile System Flush	4
DEBLOCKING (Deblock Flow)	Deblock to Flowcell	36
WASH (Acetonitrile Wash Flow)	Acetonitrile System Flush	4
	N2 System Flush	4
	Acetonitrile System Flush	4
	Acetonitrile to Flowcell	18
	N2 System Flush	4.13
	Acetonitrile System Flush	4.13
	Acetonitrile to Flowcell	15

[00197] The phosphoramidite/activator combination was delivered similar to the delivery of bulk reagents through the flowcell. No drying steps were performed as the environment stays "wet" with reagent the entire time.

[00198] The flow restrictor was removed from the ABI 394 synthesizer to enable faster flow. Without flow restrictor, flow rates for amidites (0.1M in ACN), Activator, (0.25M Benzoylthiotetrazole ("BTT"; 30-3070-xx from GlenResearch) in ACN), and Ox (0.02M I2 in 20% pyridine, 10% water, and 70% THF) were roughly ~100uL/sec, for acetonitrile ("ACN") and capping reagents (1:1 mix of CapA and CapB, wherein CapA is acetic anhydride in THF/Pyridine and CapB is 16% 1-methylimidazole in THF), roughly ~200uL/sec, and for Deblock (3% dichloroacetic acid in toluene), roughly ~300uL/sec (compared to ~50uL/sec for all reagents with flow restrictor). The time to completely push out Oxidizer was observed, the timing for chemical flow times was adjusted accordingly and an extra ACN wash was introduced between different chemicals. After polynucleotide synthesis, the chip was deprotected in gaseous ammonia overnight at 75 psi. Five drops of water were applied to the surface to recover nucleic acids. The recovered nucleic acids were then analyzed on a BioAnalyzer small RNA chip (data not shown).

[00199] **Example 3: Synthesis of a 100-mer sequence on a polynucleotide synthesis device**

[00200] The same process as described in Example 2 for the synthesis of the 50-mer sequence was used for the synthesis of a 100-mer polynucleotide ("100-mer polynucleotide"; 5' CGGGATCCTTATCGTCATCGTCGTACAGATCCCGACCCATTTGCTGTCCACCAGTCATGCTAGCCATACCATGATGATGATGATGATGAGAACCCCGCAT##TTTTTTTTTTT3', where # denotes Thymidine-succinyl hexamide CED phosphoramidite (CLP-2244 from ChemGenes); SEQ ID NO.: 21) on two different silicon chips, the first one uniformly functionalized with N-(3-TRIETHOXY-SILYL-PROPYL)-4-HYDROXYBUTYRAMIDE and the second one functionalized with 5/95 mix of 11-acetoxyundecyltriethoxysilane and n-decyltriethoxysilane, and the nucleic acids extracted from the surface were analyzed on a BioAnalyzer instrument (data not shown).

[00201] All ten samples from the two chips were further PCR amplified using a forward (5'ATGCGGGGTTCTCATCATC3'; SEQ ID NO.: 22) and a reverse (5'CGGGATCCTTATCGTCATCG3'; SEQ ID NO.: 23) primer in a 50uL PCR mix (25uL NEB Q5 mastermix, 2.5uL 10uM Forward primer, 2.5uL 10uM Reverse primer, 1uL nucleic acid extracted from the surface, and water up to 50uL) using the following thermalcycling program:

98 C, 30 sec

98 C, 10 sec; 63C, 10 sec; 72C, 10 sec; repeat 12 cycles

72C, 2min

[00202] The PCR products were also run on a BioAnalyzer (data not shown), demonstrating sharp peaks at the 100-mer position. Next, the PCR amplified samples were cloned, and Sanger sequenced. **Table 7** summarizes the results from the Sanger sequencing for samples taken from spots 1-5 from chip 1 and for samples taken from spots 6-10 from chip 2.

Table 7. Sequencing results

Spot	Error rate	Cycle efficiency
1	1/763 bp	99.87%
2	1/824 bp	99.88%
3	1/780 bp	99.87%
4	1/429 bp	99.77%
5	1/1525 bp	99.93%
6	1/1615 bp	99.94%
7	1/531 bp	99.81%
8	1/1769 bp	99.94%
9	1/854 bp	99.88%
10	1/1451 bp	99.93%

[00203] Thus, the high quality and uniformity of the synthesized polynucleotides were repeated on two chips with different surface chemistries. Overall, 89%, corresponding to 233 out of 262 of the 100-mers that were sequenced were perfect sequences with no errors.

[00204] Finally, **Table 8** summarizes error characteristics for the sequences obtained from the polynucleotides samples from spots 1-10.

Table 8. Error characteristics

Sample ID/Spot no.	OSA_0 046/1	OSA_0 047/2	OSA_0 048/3	OSA_0 049/4	OSA_0 050/5	OSA_0 051/6	OSA_0 052/7	OSA_0 053/8	OSA_0 054/9	OSA_00 55/10
Total Sequences	32	32	32	32	32	32	32	32	32	32
Sequencing Quality	25 of 28	27 of 27	26 of 30	21 of 23	25 of 26	29 of 30	27 of 31	29 of 31	28 of 29	25 of 28
Oligo Quality	23 of 25	25 of 27	22 of 26	18 of 21	24 of 25	25 of 29	22 of 27	28 of 29	26 of 28	20 of 25
ROI Match Count	2500	2698	2561	2122	2499	2666	2625	2899	2798	2348
ROI Mutation	2	2	1	3	1	0	2	1	2	1
ROI Multi Base Deletion	0	0	0	0	0	0	0	0	0	0
ROI Small Insertion	1	0	0	0	0	0	0	0	0	0
ROI Single Base Deletion	0	0	0	0	0	0	0	0	0	0
Large Deletion Count	0	0	1	0	0	1	1	0	0	0
Mutation: G>A	2	2	1	2	1	0	2	1	2	1
Mutation: T>C	0	0	0	1	0	0	0	0	0	0
ROI Error Count	3	2	2	3	1	1	3	1	2	1
ROI Error Rate	Err: ~1 in 834	Err: ~1 in 1350	Err: ~1 in 1282	Err: ~1 in 708	Err: ~1 in 2500	Err: ~1 in 2667	Err: ~1 in 876	Err: ~1 in 2900	Err: ~1 in 1400	Err: ~1 in 2349
ROI Minus Primer Error Rate	MP Err: ~1 in 763	MP Err: ~1 in 824	MP Err: ~1 in 780	MP Err: ~1 in 429	MP Err: ~1 in 1525	MP Err: ~1 in 1615	MP Err: ~1 in 531	MP Err: ~1 in 1769	MP Err: ~1 in 854	MP Err: ~1 in 1451

[00205] Example 4: Generation of a nucleic acid library by single-site, single position mutagenesis

[00206] Nucleic acid primers were de novo synthesized for use in a series of PCR reactions to generate a library of nucleic acid variants of a template nucleic acid, see **FIGS. 2A-2D**. Four types of primers were generated in **FIG. 2A**: an outer 5' primer **215**, an outer 3' primer **230**, an inner 5' primer **225**, and an inner 3' primer **220**. The inner 5' primer / first nucleic acid **220** and an inner 3' primer / second nucleic acid **225** were generated using a nucleic acid synthesis method as generally outlined in **Table 6**. The inner 5' primer / first nucleic acid **220** represents a set of up to 19 primers of predetermined sequence, where each primer in the set differs from another at a single codon, in a single site of the sequence.

[00207] Nucleic acid synthesis was performed on a device having at least two clusters, each cluster having 121 individually addressable loci.

[00208] The inner 5' primer **225** and the inner 3' primer **220** were synthesized in separate clusters. The inner 5' primer **225** was replicated 121 times, extending on 121 loci within a single cluster. For inner 3' primer **220**, each of the 19 primers of variant sequences were each extended on 6 different loci, resulting in the extension of 114 nucleic acids on 114 different loci.

[00209] Synthesized nucleic acids were cleaved from the surface of the device and transferred to a plastic vial. A first PCR reaction was performed, using fragments of the long nucleic acid sequence **235, 240** to amplify the template nucleic acid, as illustrated in **FIG. 2B**. A second PCR reaction was performed using primer combination and the products of the first PCR reaction as a template, as illustrated in **FIGs. 2C-2D**. Analysis of the second PCR products was conducted on a BioAnalyzer, as shown in the trace of **FIG. 17**.

[00210] Example 5: Generation of a nucleic acid library comprising 96 different sets of single position variants

[00211] Four sets of primers, as generally shown in **FIG. 2A** and addressed in Example 2, were generated using de novo nucleic acid synthesis. For the inner 5' primer **220**, 96 different sets of primers were generated, each set of primers targeting a different single codon positioned within a single site of the template nucleic acid. For each set of primers, 19 different variants were generated, each variant comprising a codon encoding for a different amino acid at the single site. Two rounds of PCR were performed using the generated primers, as generally shown in **FIGs. 2A-2D** and described in Example 2. The 96 sets of amplification products were visualized in an electropherogram (**FIG. 18**), which was used to calculate a 100% amplification success rate.

[00212] Example 6: Generation of a nucleic acid library comprising 500 different sets of single position variants

[00213] Four sets of primers, as generally shown in **FIG. 2A** and addressed in Example 2, were generated using de novo nucleic acid synthesis. For the inner 5' primer **220**, 500 different sets of primers were generated, each set of primers targeting a different single codon positioned within a single site of the template nucleic acid. For each set of primers, 19 different variants were generated, each variant comprising a codon encoding for a different amino acid at the single site. Two rounds of PCR were performed using the generated primers, as generally shown in **FIG. 2A** and described in Example 2. Electropherograms display each of the 500 sets of PCR products having a population of nucleic acids with 19 variants at a different single site (data not shown). A comprehensive sequencing analysis of the library showed a greater than 99% success rate across preselected codon mutations (sequence trace and analysis data not shown).

[00214] Example 7: Single-site mutagenesis primers for 1 position

[00215] An example of codon variation design is provided in **Table 9** for Yellow Fluorescent Protein. In this case, a single codon from a 50-mer of the sequence is varied 19 times. Variant nucleic acid sequence is indicated by bold letters. The wild type primer sequence is:

ATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT (SEQ ID NO.: 1).

In this case, the wild type codon encodes for valine, indicated by underline in SEQ ID NO.: 1.

Therefore the 19 variants below exclude a codon encoding for valine. In an alternative example, if all triplets are to be considered, then all 60 variants would be generated, including alternative sequence for the wild type codon.

Table 9. Variant sequences

SEQ ID NO.	Variant sequence	Variant codon
2	atgTTTAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	F
3	atgTTAAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	L
4	atgATTAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	I
5	atgTCTAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	S
6	atgCCTAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	P
7	atgACTAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	T
8	atgGCTAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	A
9	atgTATAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	Y
10	atgCATAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	H
11	atgCAAAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	Q
12	atgAATAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCCCAT	N

13	atg AAA AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	K
14	atg GAT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	D
15	atg GAA AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	E
16	atg TGT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	C
17	atg TGG AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	W
18	atg CGT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	R
19	atg GGT AGCAAGGGCGAGGAGCTGTTACACGGGGTGGTGCCCAT	G

[00216] Example 8: Single site, dual position nucleic acid variants

[00217] De novo nucleic acid synthesis was performed under conditions similar to those described in Example 2. A single cluster on a device was generated which contained synthesized predetermined variants of a nucleic acid for 2 consecutive codon positions at a single site, each position being a codon encoding for an amino acid. In this arrangement, 19 variants/per position were generated for 2 positions with 3 replicates of each nucleic acid, resulting in 114 nucleic acids synthesized.

[00218] Example 9: Multiple site, dual position nucleic acid variants

[00219] De novo nucleic acid synthesis was performed under conditions similar to those described in Example 2. A single cluster on a device was generated which contained synthesized predetermined variants of a nucleic acid for 2 non-consecutive codon positions, each position being a codon encoding for an amino acid. In this arrangement, 19 variants/per position were generated for 2 positions.

[00220] Example 10: Single stretch, triple position nucleic acid variants

[00221] De novo nucleic acid synthesis was performed under conditions similar to those described in Example 2. A single cluster on a device was generated which contained synthesized predetermined variants of a reference nucleic acid for 3 consecutive codon positions. In the 3 consecutive codon position arrangement, 19 variants/ per position were generated for 3 positions with 2 replicates of each nucleic acid, and resulted in 114 nucleic acids synthesized.

[00222] Example 11: Multiple site, triple position nucleic acid variants

[00223] De novo nucleic acid synthesis was performed under conditions similar to those described in Example 2. A single cluster on a device was generated which contains synthesized predetermined variants of a reference nucleic acid for at least 3 non-consecutive codon positions. Within a predetermined region, the location of codons encoding for 3 histidine residues were varied.

[00224] Example 12: Multiple site, multiple position nucleic acid variants

[00225] De novo nucleic acid synthesis was performed under conditions similar to those described in Example 2. A single cluster on a device was generated which contained synthesized predetermined variants of a reference nucleic acid for 1 or more codon positions in 1 or more stretches. Five positions were varied in the library. The first position encoded codons for a resultant 50/50 K/R ratio in the expressed protein; the second position encoded codons for a resultant 50/25/25 V/L/S ratio in the expressed protein, the third position encoded codons for a resultant a 50/25/25 Y/R/D ratio in the expressed protein, the fourth position encoded codons for a resultant an equal ratio for all amino acids in the expressed protein, and the fifth position encoded codons for a resultant a 75/25 G/P ratio in the expressed protein.

[00226] **Example 13: Modular plasmid components for expressing diverse peptides**

[00227] A polynucleotide library is generated as in Examples 4-6 and 8-12, encoding for codon variation at a single site or multiple sites for each of separate regions that make up portions of an expression construct cassette, as depicted in **FIG. 11**. To generate a two construct expressing cassette, variant nucleic acids were synthesized encoding at least a portion of a variant sequence of a first promoter **1110**, first open reading frame **1120**, first terminator **1130**, second promoter **1140**, second open reading frame **1150**, or second terminator sequence **1160**. After rounds of amplification, as described in previous examples, a library of 1,024 expression constructs is generated.

[00228] **Example 14: Multiple site, single position variants**

[00229] A polynucleotide library is generated as in Examples 4-6 and 8-12, encoding for codon variation at a single site or multiple sites in a region encoding for at least a portion of nucleic acid. A library of nucleic acid variants is generated, wherein the library consists of multiple site, single position variants. See, for example, **FIG. 6B**.

[00230] **Example 15: Variant library synthesis**

[00231] De novo nucleic acid synthesis is performed under conditions similar to those described in Example 2. At least 30,000 non-identical nucleic acids are de novo synthesized, wherein each of the non-identical nucleic acids encodes for a different codon variant of an amino acid sequence. The synthesized at least 30,000 non-identical nucleic acids have an aggregate error rate of less than 1 in 1:000 bases compared to predetermined sequences for each of the at least 30,000 non-identical nucleic acids. The library is used for PCR mutagenesis of a long nucleic acid and at least 30,000 non-identical variant nucleic acids are formed.

[00232] **Example 16: Cluster-based Variant library synthesis**

[00233] De novo nucleic acid synthesis is performed under conditions similar to those described in Example 2. A single cluster on a device is generated which contained synthesized predetermined

variants of a reference nucleic acid for 2 codon positions. In the 2 consecutive codon position arrangement, 19 variants/ per position were generated for the 2 positions with 2 replicates of each nucleic acid, and resulted in 38 nucleic acids synthesized. Each variant sequence is 40 bases in length. In the same cluster, additional non-variant nucleic acids sequence are generated, where the additional non-variant nucleic acids and the variant nucleic acids collective encode for 38 variants of the coding sequence of a gene. Each of the nucleic acids has at least one region reverse complementary to another of the nucleic acids. The nucleic acids in the cluster are released by gaseous ammonia cleavage. A pin comprising water contacts the cluster, picks up the nucleic acids, and moves the nucleic acids to a small vial. The vial also contains DNA polymerase reagents for a polymerase cycling assembly (PCA) reaction. The nucleic acids anneal, gaps are filled in by an extension reaction, and resultant double-stranded DNA molecules are formed, forming a variant nucleic acid library. The variant nucleic acid library is, optionally, subjected to restriction enzyme is then ligated into expression vectors.

[00234] Example 17: Generation of a variant nucleic acid TCR library

[00235] The following steps are taken for generation of a TCR library with diversity of 10^8 . Peripheral blood mononuclear cells are obtained from 10 human donor subjects and T cells are isolated. All 10 subjects are HLA mapped to have a matched allele, e.g., HLA-A*2:01. Of the 10 human donor subjects, 7 are healthy donors and 3 are donors with a form of cancer. Genes involved in the TCR complex are sequenced from the T cells of each donor, and the 1000 most common TCR sequences are identified. Each of the common TCR sequences will have a span of between 600-900 base pairs.

[00236] 10,000 predetermined variant nucleic acid sequences for each of the 1000 most common TCR sequences are designed, resulting in a DNA library with at least 10^8 total diversity. Each variant sequence is inserted into a mammalian expression vector for gene expression.

[00237] Example 18: Expression and screening of a variant TCR protein complex library

[00238] The library of variant TCR genes is transferred into mammalian cells to generate a library of cells populations, each cell population expressing a different TCR variant protein. The protein library is screened for TCR complexes with improved affinity (measure of the strength of interaction between an epitope and an antibody's antigen binding site) for a tumor antigen, such as MAGE A3 or NY-ESO-1. Additional functional considerations, such as variant gene expression, avidity (measure of the overall strength of an antibody-antigen complex), stability, and target specificity are also assessed.

[00239] Example 19: Manufacturing and delivery of engineered T cells

[00240] T cells are harvested from a subject diagnosed with cancer and are genetically engineered with a new T cell receptor (TCR) selected after performing analysis in Example 18. After a brief period of *in vitro* expansion and passing of product-specific release criteria, the T-cell product is administered to the same subject. *See* **FIG. 10C**.

[00241] Example 20: Variant TCR libraries

[00242] De novo nucleic acid synthesis is performed under conditions similar to those described in Example 2 to generate nucleic acids encoding for variant TCRs. The variant TCRs comprise variants in the antigen binding interface between the TCR and the tumor antigen, such as MAGE A3 or NY-ESO-1. Variants are generated at a single site across the variant TCR library for up to 30 residues, resulting in a first library with about 10^9 total diversity. Residues for variation are in the variable domains of the TCR alpha chain and the TCR beta chain.

[00243] The first library comprising variant TCRs are then expressed in T cells. The variant TCRs are screened *in vitro* against tumor antigens for specificity of the variant TCRs to the tumor antigen. Variant TCRs that are highly specific for the tumor antigen are then further variegated to generate a second library. The second library comprises variation in residues located in the constant domains of the TCR alpha chain and the TCR beta chain. The second library is expressed in T cells and screened *in vitro* for a second improvement, for example, avidity, stability, affinity, or expression.

[00244] Select TCR genes or gene fragment variants having desired features after screening products from the first and/or second variant libraries are selected for development of a potential therapeutic, optionally in the form of a soluble TCR protein, lacking a region to anchor the protein to a T cell, or in the form of a gene for expression in an engineered T cell as part of an immunotherapy involving the reprogramming of T cells (either from a subject or HLA-matched donor) to target cancer cells in a subject suffering from cancer.

[00245] While preferred embodiments of the present disclosure have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the disclosure. It should be understood that various alternatives to the embodiments of the disclosure described herein may be employed in practicing the disclosure. It is intended that the following claims define the scope of the disclosure and that methods and structures within the scope of these claims and their equivalents be covered thereby.

CLAIMS

WHAT IS CLAIMED IS:

1. A method of synthesizing a nucleic acid library, comprising:
 - a. providing a first set of preselected polynucleotide sequences encoding for at least 3000 variant sequences of a TCR gene or gene fragment, wherein each variant sequence comprises at least one variation at a preselected codon for an amino acid residue in an antigen contacting interface;
 - b. synthesizing the first set of preselected polynucleotide sequences;
 - c. screening binding activity for proteins encoded by the first set of polynucleotide sequences;
 - d. providing a second set of preselected polynucleotide sequences encoding for at least one variant sequence of a TCR gene or gene fragment, where each variant sequence comprises at least one variation at a preselected codon for an amino acid residue in the TCR gene or gene fragment in a region encoding a constant domain;
 - e. synthesizing the second set of preselected polynucleotide sequences; and
 - f. screening a second activity for proteins encoded by the second set of polynucleotide sequences.
2. The method of claim 1, wherein the at least one variation is in a variable domain coding region of the TCR gene or gene fragment.
3. The method of claim 2, wherein the variable domain is a variable domain of TCR alpha, TCR beta, TCR gamma, or TCR delta.
4. The method of claim 1, wherein the at least one variation is in a constant domain coding region of the TCR gene or gene fragment.
5. The method of claim 4, wherein the constant domain is a constant domain of TCR alpha, TCR beta, TCR gamma, or TCR delta.
6. The method of claim 1, wherein the antigen is a cancer antigen.
7. The method of claim 6, wherein the cancer antigen is MAGE A3, MAGE A12, MAGE A2, MAGE A6, NY-ESO-1, or CEA.
8. The method of claim 4, wherein each variant sequence comprises up to 100 variations at preselected codons for amino acid residues in the antigen contacting interface.
9. The method of claim 4, wherein each variant sequence comprises up to 30 variations at preselected codons for amino acid residues in the antigen contacting interface.

10. The method of claim 1, wherein the second activity is cancer cell killing, protein expression, or protein stability.

11. The method of claim 4, wherein each variant sequence comprises up to 5 variations at preselected codons for amino acid residues in the antigen contacting interface.

12. The method of claim 1, wherein the variant allows for generation of a TCR protein having increased specificity, avidity, affinity, stability, or expression.

13. The method of claim 1, wherein each variant nucleic acid further comprises at least one variation at a preselected codon for an amino acid residue within a constant domain of the T cell receptor (TCR) protein or fragment thereof.

14. The method of claim 1, wherein a region selected for variation spans up to 1000 bases in length.