



(43) International Publication Date
27 December 2012 (27.12.2012)

(51) International Patent Classification:

C12N 9/12 (2006.01) C12N 1/15 (2006.01)
C12N 15/54 (2006.01) C12N 5/10 (2006.01)
C12N 1/21 (2006.01) C12N 9/96 (2006.01)
C12N 1/19 (2006.01)

(21) International Application Number:

PCT/US2012/043246

(22) International Filing Date:

20 June 2012 (20.06.2012)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

13/165,373 21 June 2011 (21.06.2011) US

(71) Applicant (for all designated States except US): **BIO-RAD LABORATORIES, INC.** [US/US]; 1000 Alfred Nobel Drive, Hercules, California 94547 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **WANG, Yan** [CN/US]; 211 Juanita Way, San Francisco, California 94127 (US). **SULLIVAN, John** [US/US]; 81 Sheridan Road, Oakland, California 94618 (US).

(74) Agents: **HINSCH, Matthew E.** et al.; Kilpatrick Townsend & Stockton LLP, Two Embarcadero Center, 8th Floor, San Francisco, California 94111 (US).

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

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

- without international search report and to be republished upon receipt of that report (Rule 48.2(g))
- with sequence listing part of description (Rule 5.2(a))

(54) Title: HYBRID POLYMERASES HAVING THE ABILITY TO PRODUCE LONG AMPLICONS

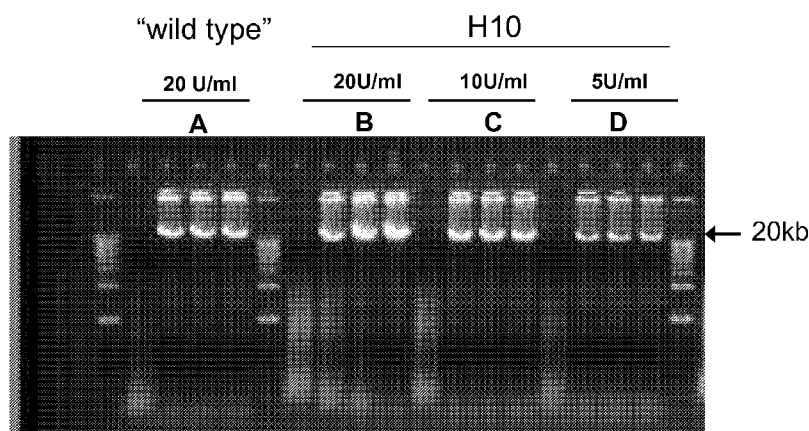


FIG. 1

(57) Abstract: The present invention provides DNA polymerases having increased efficiency of amplification of long amplicons. The present invention also provides for methods of amplifying target nucleic acid molecules with the DNA polymerases for increasing the efficiency of amplification of long amplicons.



HYBRID POLYMERASES HAVING THE ABILITY TO PRODUCE LONG AMPLICONS

CROSS-REFERENCE TO RELATED PATENT APPLICATIONS

5 [0001] The present patent application claims benefit of priority to US Patent Application No. 13/165,373, filed June 21, 2011, which is incorporated by reference for all purposes.

REFERENCE TO A "SEQUENCE LISTING," A TABLE, OR A COMPUTER PROGRAM LISTING APPENDIX SUBMITTED AS AN ASCII TEXT FILE

10

The Sequence Listing written in file -945-1PC.TXT, created on June 18, 2012, 53,248 bytes, machine format IBM-PC, MS-Windows operating system, is hereby incorporated by reference in its entirety for all purposes.

15

BACKGROUND OF THE INVENTION

[0002] Nucleic acid amplification reactions, such as polymerase chain reaction (PCR), are generally template-dependent reactions in which a desired nucleic acid sequence is amplified by treating separate complementary strands of a target nucleic acid with an excess of two oligonucleotide primers. The primers are extended to form complementary primer extension products which act as templates for synthesizing the desired nucleic acid sequence. In such processes, the nucleic acid sequence between the primers on the respective DNA strands is selectively amplified. However, the efficiency of a nucleic acid amplification reaction, particularly for long amplicon sequences, can be negatively affected by a number of factors.

20

25

BRIEF SUMMARY OF THE INVENTION

[0003] The present invention provides compositions and methods for amplifying a nucleic acid molecule. In one aspect, the present invention provides a polypeptide comprising a DNA polymerase having an amino acid sequence that is at least 95% (or at least 96%, 97%, 98%, or 99%) identical to SEQ ID NO:2.

[0004] In some embodiments, the DNA polymerase comprises SEQ ID NO:2.

[0005] In some embodiments, the polypeptide comprises a DNA polymerase having an amino acid sequence that is at least 95% (or at least 96%, 97%, 98%, or 99%) identical to SEQ ID NO:2, wherein the DNA polymerase is linked to a non-specific DNA binding

5 domain. In some embodiments, the non-specific DNA binding domain comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:4. In some embodiments, the non-specific DNA binding domain comprises SEQ ID NO:4.

10 [0006] In some embodiments, the DNA polymerase linked to the non-specific DNA binding domain has a higher efficiency for amplification of an amplicon of at least 7.5 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:8. In some embodiments, the DNA polymerase linked

15 to the non-specific DNA binding domain has a higher efficiency for amplification of an amplicon of at least 5 kb as compared to a DNA polymerase consisting of SEQ ID NO:8.

[0007] In some embodiments, the polypeptide comprises SEQ ID NO:5. In some embodiments, the polypeptide comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at

20 least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

[0008] In another aspect, the present invention provides for an isolated nucleic acid encoding a polypeptide of the present invention. In some embodiments, the isolated nucleic acid comprises a polynucleotide encoding a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, 97%, 98%, or 99%) identical to

25 SEQ ID NO:2. In some embodiments, the isolated nucleic acid comprises a polynucleotide encoding a polypeptide comprising a DNA polymerase comprising SEQ ID NO:2. In some embodiments, the isolated nucleic acid comprises a polynucleotide encoding a polypeptide having a higher efficiency for amplification of an amplicon of at least 5 kb, at least 7.5 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least

30 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:7.

[0009] In some embodiments, the isolated nucleic acid comprises a polynucleotide encoding a polypeptide comprising a DNA polymerase having an amino acid sequence at

least 95% (or at least 96%, 97%, 98%, or 99%) identical to SEQ ID NO:2, wherein the DNA polymerase is linked to a non-specific DNA binding domain. In some embodiments, the non-specific DNA binding domain comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%,
5 at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:4. In some embodiments, the non-specific DNA binding domain comprises SEQ ID NO:4. In some embodiments, the isolated nucleic acid comprises a polynucleotide encoding a DNA polymerase linked to the non-specific DNA binding domain that has a higher efficiency for amplification of an amplicon of at least 5 kb, at least 7.5 kb, at least 10 kb, at least 12 kb, at
10 least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:8.

[0010] In some embodiments, the isolated nucleic acid comprises a polynucleotide encoding a polypeptide comprising SEQ ID NO:5. In some embodiments, the isolated
15 nucleic acid comprises a polynucleotide encoding a polypeptide comprising an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

[0011] In some embodiments, the isolated nucleic acid further comprises a promoter
20 operably linked to the polynucleotide.

[0012] In another aspect, the present invention provides an isolated host cell comprising a heterologous expression cassette comprising a promoter operably linked to a polynucleotide encoding a polypeptide, the polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, 97%, 98%, or 99%) identical to SEQ ID NO:2.

25 [0013] In some embodiments, the DNA polymerase comprises SEQ ID NO:2. In some embodiments, the DNA polymerase, when linked to a non-specific DNA binding domain, has a higher efficiency for amplification of an amplicon of at least 5 kb, at least 7.5 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase
30 consisting of SEQ ID NO:7.

[0014] In some embodiments, the DNA polymerase is linked to a non-specific DNA binding domain. In some embodiments, the non-specific DNA binding domain comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least

92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:4. In some embodiments, the non-specific DNA binding domain comprises SEQ ID NO:4. In some embodiments, the DNA polymerase linked to the non-specific DNA binding domain has a higher efficiency for amplification of an amplicon of
5 at least 5 kb, at least 7.5 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:8.

[0015] In some embodiments, the isolated host cell comprises a heterologous expression cassette comprising a promoter operably linked to a polynucleotide encoding a polypeptide, the polypeptide comprising SEQ ID NO:5 or comprising an amino acid sequence that is at
10 at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

[0016] In yet another aspect, the present invention provides methods of amplifying a
15 nucleic acid molecule. In some embodiments, the method comprises incubating the target nucleic acid in a reaction mixture comprising at least one primer and a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2, wherein said incubating is under conditions that result in amplification of the target nucleic acid by the
20 polypeptide; thereby amplifying the target nucleic acid.

[0017] In some embodiments, the polypeptide has a higher amplification efficiency amplification of an amplicon of at least 5 kb, at least 7.5 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ
25 ID NO:7.

[0018] In some embodiments, the DNA polymerase comprises SEQ ID NO:2.

[0019] In some embodiments, the DNA polymerase is linked to a non-specific DNA binding domain. In some embodiments, the non-specific DNA binding domain comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least
30 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:4. In some embodiments, the non-specific DNA binding domain comprises SEQ ID NO:4. In some embodiments, the DNA polymerase linked to the non-specific DNA binding domain has a higher efficiency for amplification of an amplicon of

at least 5 kb, at least 7 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:8.

[0020] In some embodiments, the polypeptide comprises SEQ ID NO:5. In some
5 embodiments, the polypeptide comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

[0021] In still another aspect, the present invention provides reaction mixtures for
amplifying a nucleic acid molecule. In some embodiments, the reaction mixture comprises a
10 polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2. In some embodiments, the DNA polymerase comprises SEQ ID NO:2.

[0022] In yet another aspect, the present invention provides kits for amplifying a nucleic
acid molecule. In some embodiments, the kit comprises a polypeptide comprising a DNA
15 polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2. In some embodiments, the DNA polymerase comprises SEQ ID NO:2.

[0023] In some embodiments, the reaction mixture and/or kit comprises a polypeptide that
has a higher amplification efficiency amplification of an amplicon of at least 5 kb, at least 7
20 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:7.

[0024] In some embodiments, the reaction mixture and/or kit comprises a polypeptide
comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%,
25 97%, 98%, or 99%) identical to SEQ ID NO:2, wherein the DNA polymerase is linked to a non-specific DNA binding domain. In some embodiments, the non-specific DNA binding domain comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:4. In some embodiments, the non-
30 specific DNA binding domain comprises SEQ ID NO:4. In some embodiments, the DNA polymerase linked to the non-specific DNA binding domain has a higher efficiency for amplification of an amplicon of at least 5 kb, at least 7.5 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least

30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:8. In some embodiments, the polypeptide comprises SEQ ID NO:5. In some embodiments, the polypeptide comprises an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

DEFINITIONS

[0025] The term "polymerase" refers to an enzyme that performs template-directed synthesis of polynucleotides. The term encompasses both a full length polypeptide and a domain that has polymerase activity. DNA polymerases are well-known to those skilled in the art, including but not limited to DNA polymerases isolated or derived from *Pyrococcus furiosus*, *Thermococcus litoralis*, and *Thermotoga maritime*, or modified versions thereof. They include both DNA-dependent polymerases and RNA-dependent polymerases such as reverse transcriptase. At least five families of DNA-dependent DNA polymerases are known, although most fall into families A, B and C. There is little or no sequence similarity among the various families. Most family A polymerases are single chain proteins that can contain multiple enzymatic functions including polymerase, 3' to 5' exonuclease activity and 5' to 3' exonuclease activity. Family B polymerases typically have a single catalytic domain with polymerase and 3' to 5' exonuclease activity, as well as accessory factors. Family C polymerases are typically multi-subunit proteins with polymerizing and 3' to 5' exonuclease activity. In *E. coli*, three types of DNA polymerases have been found, DNA polymerases I (family A), II (family B), and III (family C). In eukaryotic cells, three different family B polymerases, DNA polymerases α , δ , and ϵ , are implicated in nuclear replication, and a family A polymerase, polymerase γ , is used for mitochondrial DNA replication. Other types of DNA polymerases include phage polymerases. Similarly, RNA polymerases typically include eukaryotic RNA polymerases I, II, and III, and bacterial RNA polymerases as well as phage and viral polymerases. RNA polymerases can be DNA-dependent and RNA-dependent. In some embodiments, a polymerase of the present invention is identical or substantially identical (*e.g.*, has at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% amino acid sequence identity) to SEQ ID NO:2 retaining at least one (and in some embodiments, each) of the amino acids in SEQ ID NO:2 that are at the positions that vary between SEQ ID NO:2 and SEQ ID NO:7, while varying at other positions.

[0026] "Thermally stable polymerase," as used herein, refers to any enzyme that catalyzes polynucleotide synthesis by addition of nucleotide units to a nucleotide chain using DNA or RNA as a template and has an optimal activity at a temperature above 45°C.

[0027] The term "Sso7-like protein" or "Sso7," as used herein, refers to polypeptide variants, alleles, mutants, and interspecies homologs that: (1) have an amino acid sequence that has greater than about 60% amino acid sequence identity, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% or greater amino acid sequence identity to SEQ ID NO:4 or 10; or that (2) bind to antibodies, *e.g.*, polyclonal antibodies, raised against an immunogen comprising an amino acid sequence of SEQ ID NO:4 or 10. The term includes both full-length Sso7 polypeptides and fragments of the polypeptides that have sequence non-specific double-stranded binding activity. Sso7-like proteins include Sac7d, Sac7e, Ssh7b, and Sto7e.

[0028] A "domain" refers to a unit of a protein or protein complex, comprising a polypeptide subsequence, a complete polypeptide sequence, or a plurality of polypeptide sequences where that unit has a defined function. The function is understood to be broadly defined and can be ligand binding, catalytic activity, or can have a stabilizing effect on the structure of the protein.

[0029] The term "DNA binding domain" refers to a protein domain that binds DNA in a sequence non-specific manner. In some embodiments, the DNA binding domain is a protein domain which binds with significant affinity to DNA, for which there is no known nucleic acid which binds to the protein domain with more than 100-fold more affinity than another nucleic acid with the same nucleotide composition but a different nucleotide sequence.

[0030] The term "polymerase-Sso7 conjugate," as used herein, refers to a modified polymerase comprising at least one Sso7 DNA binding domain joined to a polymerase domain, or a catalytic subunit of the polymerase domain. A polymerase-Sso7 conjugate may comprise multiple Sso7 DNA binding domains.

[0031] The terms "join" or "link" refer to any method known in the art for functionally connecting protein domains, including without limitation recombinant fusion with or without intervening domains, intein-mediated fusion, non-covalent association, and covalent bonding, including disulfide bonding; hydrogen bonding; electrostatic bonding; and conformational bonding, *e.g.*, antibody-antigen, and biotin-avidin associations.

[0032] The term "nucleic acid amplification" or "amplification reaction" refers to any *in vitro* means for multiplying the copies of a target sequence of nucleic acid. Such methods

include but are not limited to polymerase chain reaction (PCR), DNA ligase chain reaction (see U.S. Pat. Nos. 4,683,195 and 4,683,202; PCR Protocols: A Guide to Methods and Applications (Innis *et al.*, eds, 1990)), (LCR), QBeta RNA replicase, and RNA transcription-based (such as TAS and 3SR) amplification reactions as well as others known to those of skill
5 in the art.

[0033] "Amplifying" refers to a step of submitting a solution to conditions sufficient to allow for amplification of a polynucleotide if all of the components of the reaction are intact. Components of an amplification reaction include, *e.g.*, primers, a polynucleotide template, polymerase, nucleotides, and the like. The term amplifying typically refers to an
10 "exponential" increase in target nucleic acid. However, amplifying as used herein can also refer to linear increases in the numbers of a select target sequence of nucleic acid, such as is obtained with cycle sequencing.

[0034] The term "amplification reaction mixture" or "amplification reaction composition" refers to an aqueous solution comprising the various reagents used to amplify a target nucleic
15 acid. These include enzymes, aqueous buffers, salts, amplification primers, target nucleic acid, and nucleoside triphosphates. As discussed further herein, amplification reaction mixtures may also further include stabilizers and other additives to optimize efficiency and specificity. Depending upon the context, the mixture can be either a complete or incomplete amplification reaction mixture.

[0035] "Polymerase chain reaction" or "PCR" refers to a method whereby a specific segment or subsequence of a target double-stranded DNA, is amplified in a geometric progression. PCR is well known to those of skill in the art; *see, e.g.*, U.S. Pat. Nos. 4,683,195 and 4,683,202; and PCR Protocols: A Guide to Methods and Applications, Innis *et al.*, eds, 1990. Exemplary PCR reaction conditions typically comprise either two or three step cycles.
20 Two step cycles have a denaturation step followed by a hybridization/elongation step. Three step cycles comprise a denaturation step followed by a hybridization step followed by a separate elongation step.

[0036] The term "increased efficiency" or "improved efficiency," as used with respect to nucleic acid amplification, refers to a detectable increase in the amount of amplification
30 products produced in a nucleic acid amplification during a specified number of rounds (*i.e.*, cycles) of amplification in the presence of a DNA polymerase of the present invention as compared to the amount of amplification products produced in a nucleic acid amplification during a specified number of rounds (*i.e.*, cycles) of amplification in the presence of a control

DNA polymerase. Efficiency of an amplification reaction can be measured according to any method, including but not limited to melt-curve analysis or gel analysis. In some embodiments, an amplification reaction in the presence of a DNA polymerase of the present invention will exhibit at least 10%, 15%, 20%, 25%, 30%, 40%, 50%, 100%, 2-fold (200%), 2.5-fold (250%), 3-fold (300%) or greater increase in the amount of amplification products produced in a given number of cycles as compared to the amount of amplification products produced in the same number of cycles using a control DNA polymerase (e.g., a DNA polymerase consisting of SEQ ID NO:7 or SEQ ID NO:8).

[0037] An "oligonucleotide primer" or "primer" refers to an oligonucleotide sequence that hybridizes to a sequence on a target nucleic acid and serves as a point of initiation of nucleic acid synthesis. Primers can be of a variety of lengths and are often less than 50 nucleotides in length, for example 12-30 nucleotides in length. The length and sequences of primers for use in PCR can be designed based on principles known to those of skill in the art; *see, e.g., Innis et al., supra*.

[0038] The term "amplicon" refers to a polynucleotide (e.g., DNA) fragment formed as the product of a natural or artificial amplification event (e.g., polymerase chain reaction). In some embodiments, an amplicon is a "long" amplicon having a length of at least 5 kb, or at least 6 kb, 7 kb, 8 kb, 9 kb, 10 kb, 11 kb, 12 kb, 13 kb, 14 kb, 15 kb, 16 kb, 17 kb, 18 kb, 19 kb, 20 kb, 21 kb, 22 kb, 23 kb, 24 kb, 25 kb, 26 kb, 27 kb, 28 kb, 29 kb, 30 kb, 31 kb, 32 kb, 33 kb, 34 kb, 35 kb, 36 kb, 37 kb, 38 kb, 39 kb, or 40 kb.

[0039] The terms "nucleic acid" and "polynucleotide" are used interchangeably herein to refer to deoxyribonucleotides or ribonucleotides and polymers thereof in either single- or double-stranded form. The term encompasses nucleic acids containing known nucleotide analogs or modified backbone residues or linkages, which are synthetic, naturally occurring, and non-naturally occurring, which have similar binding properties as the reference nucleic acid, and which are metabolized in a manner similar to the reference nucleotides. Examples of such analogs include, without limitation, phosphorothioates, phosphoramidates, methyl phosphonates, chiral-methyl phosphonates, 2-O-methyl ribonucleotides, and peptide nucleic acids (PNAs).

[0040] The terms "polypeptide," "peptide," and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical mimetic of a corresponding

naturally occurring amino acid, as well as to naturally occurring amino acid polymers and non-naturally occurring amino acid polymers.

[0041] The term "amino acid" refers to naturally occurring and synthetic amino acids, as well as amino acid analogs and amino acid mimetics that function in a manner similar to the naturally occurring amino acids. Naturally occurring amino acids are those encoded by the genetic code, as well as those amino acids that are later modified, *e.g.*, hydroxyproline, γ -carboxyglutamate, and O-phosphoserine. Amino acid analogs refers to compounds that have the same basic chemical structure as a naturally occurring amino acid, *i.e.*, an α carbon that is bound to a hydrogen, a carboxyl group, an amino group, and an R group, *e.g.*, homoserine, norleucine, methionine sulfoxide, methionine methyl sulfonium. Such analogs have modified R groups (*e.g.*, norleucine) or modified peptide backbones, but retain the same basic chemical structure as a naturally occurring amino acid. Amino acid mimetics refers to chemical compounds that have a structure that is different from the general chemical structure of an amino acid, but that functions in a manner similar to a naturally occurring amino acid.

[0042] Amino acids may be referred to herein by either their commonly known three letter symbols or by the one-letter symbols recommended by the IUPAC-IUB Biochemical Nomenclature Commission. Nucleotides, likewise, may be referred to by their commonly accepted single-letter codes.

[0043] "Conservatively modified variants" applies to both amino acid and nucleic acid sequences. With respect to particular nucleic acid sequences, conservatively modified variants refers to those nucleic acids which encode identical or essentially identical amino acid sequences, or where the nucleic acid does not encode an amino acid sequence, to essentially identical sequences. Because of the degeneracy of the genetic code, a large number of functionally identical nucleic acids encode any given protein. For instance, the codons GCA, GCC, GCG and GCU all encode the amino acid alanine. Thus, at every position where an alanine is specified by a codon, the codon can be altered to any of the corresponding codons described without altering the encoded polypeptide. Such nucleic acid variations are "silent variations," which are one species of conservatively modified variations. Every nucleic acid sequence herein which encodes a polypeptide also describes every possible silent variation of the nucleic acid. One of skill will recognize that each codon in a nucleic acid (except AUG, which is ordinarily the only codon for methionine, and TGG, which is ordinarily the only codon for tryptophan) can be modified to yield a functionally identical molecule. Accordingly, each silent variation of a nucleic acid which encodes a polypeptide is implicit in each described sequence.

[0044] As to amino acid sequences, one of skill will recognize that individual substitutions, deletions or additions to a nucleic acid, peptide, polypeptide, or protein sequence which alters, adds or deletes a single amino acid or a small percentage of amino acids in the encoded sequence is a "conservatively modified variant" where the alteration results in the substitution
5 of an amino acid with a chemically similar amino acid. Conservative substitution tables providing functionally similar amino acids are well known in the art. Such conservatively modified variants are in addition to and do not exclude polymorphic variants, interspecies homologs, and alleles of the invention.

[0045] The term "promoter" refers to regions or sequences located upstream and/or
10 downstream from the start of transcription and which are involved in recognition and binding of RNA polymerase and other proteins to initiate transcription.

[0046] The terms "identical" or percent "identity," in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same. Sequences are "substantially identical" to each other if they have a specified percentage
15 of nucleotides or amino acid residues that are the same (*e.g.*, at least 60%, 65%, 70%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% identity) over a specified region where indicated, or across the entire reference sequence if not otherwise indicated, when compared and aligned for maximum correspondence over a comparison
20 window, or designated region as measured using one of the following sequence comparison algorithms or by manual alignment and visual inspection. These definitions also refer to the complement of a test sequence.

[0047] For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if
25 necessary, and sequence algorithm program parameters are designated. Default program parameters are commonly used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities or similarities for the test sequences relative to the reference sequence, based on the program parameters.

[0048] A "comparison window," as used herein, includes reference to a segment of
30 contiguous positions, for example from 20 to 600 contiguous positions, about 50 to about 200, or about 100 to about 150, in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequences for comparison are well known in the art.

Optimal alignment of sequences for comparison can be conducted, for example, by the local homology algorithm of Smith and Waterman (*Adv. Appl. Math.* 2:482, 1970), by the homology alignment algorithm of Needleman and Wunsch (*J. Mol. Biol.* 48:443, 1970), by the search for similarity method of Pearson and Lipman (*Proc. Natl. Acad. Sci. USA* 85:2444, 5 1988), by computerized implementations of these algorithms (e.g., GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, Wis.), or by manual alignment and visual inspection (see, e.g., Ausubel *et al.*, *Current Protocols in Molecular Biology* (1995 supplement)).

[0049] Algorithms suitable for determining percent sequence identity and sequence 10 similarity are the BLAST and BLAST 2.0 algorithms, which are described in Altschul *et al.* (*Nuc. Acids Res.* 25:3389-402, 1977), and Altschul *et al.* (*J. Mol. Biol.* 215:403-10, 1990), respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short 15 words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the 20 cumulative alignment score can be increased. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W , T , and X determine the sensitivity and speed of the 25 alignment. The BLAST program uses as defaults a word length (W) of 11, the BLOSUM62 scoring matrix (see Henikoff & Henikoff, *Proc. Natl. Acad. Sci. USA* 89:10915 (1989)) alignments (B) of 50, expectation (E) of 10, $M=5$, $N=-4$, and a comparison of both strands.

[0050] The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, e.g., Karlin & Altschul, *Proc. Nat'l. Acad. Sci. USA* 90:5873-5787 30 (1993)). One measure of similarity provided by the BLAST algorithm is the smallest sum probability ($P(N)$), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence if the smallest sum probability in a comparison

of the test nucleic acid to the reference nucleic acid is less than about 0.2, more preferably less than about 0.01, and most preferably less than about 0.001.

BRIEF DESCRIPTION OF THE DRAWINGS

5 **[0051] Figure 1. H10 fusion polymerase has a higher efficiency than wild-type fusion polymerase in amplifying a 20 kb target from lambda DNA.** The increase in efficiency is reflected by the higher yield of the final product when equal units of enzyme were used (compare A and B), or the need for less enzyme to produce the same yield (compare A and C).

10 **[0052] Figure 2. Sequence alignment comparing the amino acid sequence of the DNA polymerase of SEQ ID NO:2 with the amino acid sequence of the DNA polymerase of SEQ ID NO:7.** Consensus = SEQ ID NO:17.

DETAILED DESCRIPTION

I. Introduction

15 **[0053]** The present invention provides methods and compositions for improving the efficiency of nucleic acid amplification, in particular in improving the efficiency of long amplicons. A DNA polymerase having improved activities has been discovered. The DNA polymerase was initially identified as a polymerase having a polymerase activity greater than its 3'-5' exonuclease activity. It was further discovered that the DNA polymerases has substantially better ability (as determined by amplicon yield) to amplify long (*e.g.*, 7.5 or 20
20 kb) amplicons as compared to SEQ ID NOs:7 or 8.

[0054] Thus, in one aspect, the present invention provides a polypeptide comprising a DNA polymerase having an amino acid sequence that is at least 95% identical to SEQ ID NO:2. In some embodiments, the polypeptide has higher efficiency for amplification of a long amplicon (*e.g.*, an amplicon of at least 5 kb, at least 7.5 kb, at least 10 kb, at least 15 kb, or at
25 least 20 kb) as compared to a DNA polymerase consisting of SEQ ID NO:7. In some embodiments, the polypeptide comprises a DNA polymerase having an amino acid sequence at least 95% identical to SEQ ID NO:2 linked to a non-specific DNA binding domain, *e.g.*, a non-specific DNA binding domain comprising an amino acid sequence that is at least 80%, 85%, 90%, 95%, 99%, or 100% identical to SEQ ID NO:4.

[0055] In another aspect, the present invention provides methods of amplifying a nucleic acid molecule. In some embodiments, the method comprises incubating the target nucleic acid in a reaction mixture comprising at least one primer and a polypeptide comprising a DNA polymerase having an amino acid sequence that is at least 95% identical to SEQ ID NO:2, wherein said incubating is under conditions that result in amplification of the target nucleic acid by the polymerase, thereby amplifying the target nucleic acid. In some embodiments, the polymerase comprises a DNA polymerase having an amino acid sequence at least 95% identical to SEQ ID NO:2 linked to a non-specific DNA binding domain, *e.g.*, a non-specific DNA binding domain comprising an amino acid sequence that is at least 80%, 85%, 90%, 95%, 99%, or 100% identical to SEQ ID NO:4.

[0056] In still another aspect, the present invention provides reaction mixtures and/or kits for performing nucleic acid amplification reactions that use polymerases. In some embodiments, the reaction mixtures and/or kits comprise a polypeptide comprising a DNA polymerase having an amino acid sequence that is at least 95% identical to SEQ ID NO:2. In some embodiments, the reaction mixtures and/or kits comprise a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% identical to SEQ ID NO:2 and linked to a non-specific DNA binding domain, *e.g.*, a non-specific DNA binding domain comprising an amino acid sequence that is at least 80%, 85%, 90%, 95%, 99%, or 100% identical to SEQ ID NO:4.

II. Polymerases of the invention

[0057] An exemplary polymerase of the invention comprises SEQ ID NO:2 or at least comprises a polypeptide substantially identical across the full-length of SEQ ID NO:2. These polymerases are synthetic, non-naturally-occurring polymerases.

[0058] Variants of SEQ ID NO:2 that substantially retain the polymerase activity of SEQ ID NO:2 can be generated, for example, by retaining each of the amino acids in SEQ ID NO:2 that are at the positions that vary between SEQ ID NO:2 and SEQ ID NO:7, while varying other positions. One of skill in the art can determine the positions that vary between SEQ ID NO:2 and SEQ ID NO:7, and thus determine positions which positions to retain and/or which positions to vary from SEQ ID NO:2, by reference to a sequence alignment, for example a sequence alignment shown in Figure 2. It will be further appreciated that some positions in the polymerase as highly conserved across different polymerases and thus variation in such positions are more likely to alter or harm polymerase activity than positions that are variable between polymerases. An example of a sequence conserved between SEQ

ID NO:2 and other polymerases is YGYGYAKARWYCKECAESVTAWGR (SEQ ID NO:9). *See, e.g.*, US Patent No. 7,560,260. Thus in some embodiments, the polymerases of the invention are substantially identical to SEQ ID NO:2 and comprise SEQ ID NO:9 as well as retaining each of the amino acids in SEQ ID NO:2 that are at the positions that vary
5 between SEQ ID NO:2 and SEQ ID NO:7. In some embodiments, the polymerases of the invention comprise a polypeptide at least 95% identical to SEQ ID NO:2.

[0059] In one aspect, the polymerases of the invention are characterized by having a high efficiency at amplifying long amplicons (*e.g.*, an amplicon of at least 5 kb, at least 7.5 kb, at least 10 kb, at least 15 kb, or at least 20 kb). For example, in some embodiments, the
10 polymerases of the invention are more efficient (*i.e.*, yield more product) in a nucleic acid amplification reaction than a polymerase comprising SEQ ID NO:7. In some embodiments, the polymerases of the invention are more efficient (*i.e.*, yield more product) in a nucleic acid amplification reaction than a polymerase comprising SEQ ID NO:8.

[0060] In some embodiments, polymerases of the invention comprise mutations that render
15 the polymerase exonuclease deficient. "Exonuclease deficient" as used herein means that the polymerase has a substantially reduced (*i.e.*, less than 10%, 5% or 1% of the 3'-5' exonuclease activity of Pfu DNA polymerase from *Pyrococcus furiosus*) or no exonuclease activity. For example, a double point mutation in the polymerase domain substituting an alanine at positions D141 and E143 can remove or eliminate 3'-5' exonuclease activity. *See,*
20 *e.g.*, Derbyshire *et al.*, *Methods in Enzymology*, Vol 262 (1995), pages 363-385. Hybrid polymerases comprising such double point mutations will generally show an increased specificity in nucleic acid amplification reactions, resulting in fewer amplification byproducts (such as amplification of primer-dimers) and increased efficiency in amplification of the desired target nucleic acids.

[0061] In some embodiments, the polymerases (including hybrid polymerases) of the
25 invention can be in isolated form. In some embodiments, the polymerases (including hybrid polymerases) are lyophilized. Alternatively, the polymerases (including hybrid polymerases) are in soluble form in an aqueous solution. In some embodiments, the polymerases (including hybrid polymerases) are produced recombinantly and thus are within an intact cell
30 or in a cell lysate.

Nucleic acid binding domains

[0062] In some embodiments, the polymerases of the invention are fused to a DNA binding domain. Such fusions are sometimes referred to herein as "hybrid polymerases." A DNA

binding domain is a protein, or a defined region of a protein, that binds to nucleic acid in a sequence-independent manner, e.g., binding does not exhibit a gross preference for a particular sequence. DNA binding domains may bind single or double stranded nucleic acids.

[0063] The DNA binding proteins of use in the invention are generally thermostable.

5 Examples of such proteins include, but are not limited to, the Archaeal small basic DNA binding proteins Sso7d and Sso7d-like proteins (see, e.g., Choli et al., *Biochimica et Biophysica Acta* 950:193-203, 1988; Baumann et al., *Structural Biol.* 1:808-819, 1994; and Gao et al, *Nature Struc. Biol.* 5:782-786, 1998), Archaeal HMF-like proteins (see, e.g., Starich et al., *J. Molec. Biol.* 255:187-203, 1996; Sandman et al., *Gene* 150:207-208, 1994), and
10 PCNA homologs (see, e.g., Cann et al., *J. Bacteriology* 181:6591-6599, 1999; Shamoo and Steitz, *Cell*:99, 155-166, 1999; De Felice et al., *J. Mol. Biol.* 291, 47-57, 1999; and Zhang et al., *Biochemistry* 34:10703-10712, 1995).

[0064] The HMF-like proteins are archaeal histones that share homology both in amino acid sequences and in structure with eukaryotic H4 histones, which are thought to interact directly
15 with DNA. The HMF family of proteins form stable dimers in solution, and several HMF homologs have been identified from thermostable species (e.g., *Methanothermus fervidus* and *Pyrococcus* strain GB-3a). The HMF family of proteins, once joined to Taq DNA polymerase or any DNA modifying enzyme with a low intrinsic processivity, can enhance the ability of the enzyme to slide along the DNA substrate and thus increase its processivity. For example,
20 the dimeric HMF-like protein can be covalently linked to the N terminus of Taq DNA polymerase, e.g., via chemical modification, and thus improve the processivity of the polymerase.

[0065] Certain helix-hairpin-helix motifs have been shown to bind DNA nonspecifically and enhance the processivity of a DNA polymerase to which it is fused (Pavlov et al., *Proc Natl Acad Sci USA.* 99:13510-5, 2002).
25

[0066] Sso7d and Sso7d-like proteins, Sac7d and Sac7d-like proteins, e.g., Sac7a, Sac7b, Sac7d, and Sac7e are small (about 7,000 kd MW), basic chromosomal proteins from the hyperthermophilic archaeobacteria *Sulfolobus solfataricus* and *S. acidocaldarius*, respectively. These proteins are lysine-rich and have high thermal, acid and chemical stability. They bind
30 DNA in a sequence-independent manner and when bound, increase the T_m of DNA by up to 40°C under some conditions (McAfee, *Biochemistry* 34:10063-10077, 1995; Gao et al., *Nat. Struct. Biol.* 5(9):782-786, 1998). These proteins and their homologs are typically believed to be involved in stabilizing genomic DNA at elevated temperatures. Suitable Sso7d-like

DNA binding domains for use in the invention can be modified based on their sequence homology to Sso7d. Typically, DNA binding domains that are identical to or substantially identical to a known DNA binding protein over a comparison window of about 25 amino acids, optionally about 50-100 amino acids, or the length of the entire protein, can be used in the invention. The sequence can be compared and aligned for maximum correspondence over a comparison window, or designated region as measured using one of the described comparison algorithms or by manual alignment and visual inspection. In some embodiments, the DNA binding domain comprises SEQ ID NO:4 or SEQ ID NO:10 or a substantially (*e.g.*, at least 60%, 70%, 80%, 90%, or 95%) identical sequence thereof. A variety of mutations in the Sso7 binding domain have been described in, *e.g.*, US Patent Application Nos. 2005/0048530 and 2007/0141591.

[0067] Additional DNA binding domains suitable for use in the invention can be identified by homology with known DNA binding proteins and/or by antibody crossreactivity, or may be found by means of a biochemical assay. DNA binding domains may be synthesized or isolated using the techniques described herein and known in the art.

[0068] Sequence non-specific double-stranded nucleic acid binding domains for use in the invention can also be identified by cross-reactivity using antibodies, including but not limited to polyclonal antibodies, that bind to known nucleic acid binding domains. Polyclonal antibodies are generated using methods well known to those of ordinary skill in the art (*see, e.g.*, Coligan, *Current Protocols in Immunology* (1991); Harlow & Lane, *Antibodies, A Laboratory Manual* (1988)). Those proteins that are immunologically cross-reactive binding proteins can then be detected by a variety of assay methods. For descriptions of various formats and conditions that can be used, *see, e.g.*, *Methods in Cell Biology: Antibodies in Cell Biology*, volume 37 (Asai, ed. 1993), Coligan, *supra*, and Harlow & Lane, *supra*.

[0069] Specificity for binding to double-stranded nucleic acids can be tested using a variety of assays known to those of ordinary skill in the art. These include such assays as filter binding assays or gel-shift assays. For example, in a filter-binding assay the polypeptide to be assessed for binding activity to double-stranded DNA is pre-mixed with radio-labeled DNA, either double-stranded or single-stranded, in the appropriate buffer. The mixture is filtered through a membrane (*e.g.*, nitrocellulose) which retains the protein and the protein-DNA complex. The amount of DNA that is retained on the filter is indicative of the quantity that bound to the protein. Binding can be quantified by a competition analysis in which binding of labeled DNA is competed by the addition of increasing amounts of unlabelled DNA. A polypeptide that binds double-stranded DNA at a 10-fold or greater affinity than

single-stranded DNA is defined herein as a double-stranded DNA binding protein.

Alternatively, binding activity can be assessed by a gel shift assay in which radiolabeled DNA is incubated with the test polypeptide. The protein-DNA complex will migrate slower through the gel than unbound DNA, resulting in a shifted band. The amount of binding is assessed by incubating samples with increasing amounts of double-stranded or single-stranded unlabeled DNA, and quantifying the amount of radioactivity in the shifted band.

[0070] A binding domain suitable for use in the invention binds to double-stranded nucleic acids in a sequence-independent fashion, i.e., a binding domain of the invention binds double-stranded nucleic acids with a significant affinity, but, there is no known nucleic acid that binds to the domain with more than 100-fold more affinity than another nucleic acid with the same nucleotide composition, but a different nucleic acid sequence. Non-specific binding can be assayed using methodology similar to that described for determining double-stranded vs. single-stranded nucleic acid binding. Filter binding assays or gel mobility shift assays can be performed as above using competitor DNAs of the same nucleotide composition, but different nucleic acid sequences to determine specificity of binding.

[0071] Sequence non-specific double-stranded nucleic acid binding domains for use in the invention can also be assessed, for example, by assaying the ability of the double-stranded binding domain to increase processivity or efficiency of a modifying enzyme or to increase the stability of a nucleic acid duplex by at least 1°C can be determined.

[0072] A binding domain of the invention can also be identified by direct assessment of the ability of such a domain to stabilize a double-stranded nucleic acid conformation. For example, a melting curve of a primer-template construct can be obtained in the presence or absence of protein by monitoring the UV absorbance of the DNA at 260 nm. The T_m of the double-stranded substrate can be determined from the midpoint of the melting curve. The effect of the sequence-non-specific double-stranded nucleic-acid-binding protein on the T_m can then be determined by comparing the T_m obtained in the presence of the modified enzyme with that in the presence of the unmodified enzyme. (The protein does not significantly contribute to the UV absorbance because it has a much lower extinction coefficient at 260 nm than DNA). A domain that increases the T_m by 1°C, often by 5°C, 10°C or more, can then be selected for use in the invention.

[0073] Novel sequence non-specific double-stranded nucleic acid binding proteins of the invention can also be isolated by taking advantage of their DNA binding activity, for instance by purification on DNA-cellulose columns. The isolated proteins can then be further purified

by conventional means, sequenced, and the genes cloned by conventional means via PCR. Proteins overexpressed from these clones can then be tested by any of the means described above.

Conjugating polymerases to nucleic acid binding domains

5 [0074] A DNA polymerase of the present invention (e.g., SEQ ID NO2 or a variant thereof) can be joined to a sequence non-specific DNA binding domain by methods well known to those of skill in the art. These methods include both chemical and recombinant means.

[0075] Chemical linking of the DNA polymerase to the non-specific DNA binding domain can be performed, for example, as described in Bioconjugate Techniques, Hermanson, Ed.,
10 Academic Press (1996). Joining can include, for example, derivitization for the purpose of linking the two proteins to each other, either directly or through a linking compound, by methods that are well known in the art of protein chemistry. For example, in one chemical conjugation embodiment, the means of linking the catalytic domain and the nucleic acid binding domain comprises a heterobifunctional-coupling reagent which ultimately contributes
15 to formation of an intermolecular disulfide bond between the two moieties. Other types of coupling reagents that are useful in this capacity for the present invention are described, for example, in U.S. Patent 4,545,985. Alternatively, an intermolecular disulfide may conveniently be formed between cysteines in each moiety, which occur naturally or are inserted by genetic engineering. The means of linking moieties may also use thioether
20 linkages between heterobifunctional crosslinking reagents or specific low pH cleavable crosslinkers or specific protease cleavable linkers or other cleavable or noncleavable chemical linkages.

[0076] Linking the DNA polymerase to the non-specific DNA binding domain may also comprise a peptidyl bond formed between moieties that are separately synthesized by
25 standard peptide synthesis chemistry or recombinant means. The conjugate protein itself can also be produced using chemical methods to synthesize an amino acid sequence in whole or in part. For example, peptides can be synthesized by solid phase techniques, such as, e.g., the Merrifield solid phase synthesis method, in which amino acids are sequentially added to a growing chain of amino acids (see, Merrifield (1963) J. Am. Chem. Soc., 85:2149-2146).
30 Equipment for automated synthesis of polypeptides is commercially available from suppliers such as PE Corp. (Foster City, CA), and may generally be operated according to the manufacturer's instructions. The synthesized peptides can then be cleaved from the resin, and purified, e.g., by preparative high performance liquid chromatography (see Creighton,

Proteins Structures and Molecular Principles, 50-60 (1983)). The composition of the synthetic polypeptides or of subfragments of the polypeptide, may be confirmed by amino acid analysis or sequencing (e.g., the Edman degradation procedure; see Creighton, Proteins, Structures and Molecular Principles, pp. 34-49 (1983)).

5 [0077] In addition, nonclassical amino acids or chemical amino acid analogs can be introduced as a substitution or addition into the sequence. Non-classical amino acids include, but are not limited to, the D-isomers of the common amino acids, α -amino isobutyric acid, 4-aminobutyric acid, Abu, 2-amino butyric acid, γ -Abu, ϵ -Ahx, 6-amino hexanoic acid, Aib, 2-aminobutyric acid, 3-amino propionic acid, ornithine, norleucine, norvaline, hydroxy-
10 proline, sarcosine, citrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine, cyclohexylalanine, β - alanine, fluoro-amino acids, designer amino acids such as β -methyl amino acids, C α -methyl amino acids, N -methyl amino acids, and amino acid analogs in general. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

[0078] In another embodiment, the DNA polymerase and the non-specific DNA binding
15 domain are joined via a linking group. The linking group can be a chemical crosslinking agent, including, for example, succinimidyl-(N-maleimidomethyl)-cyclohexane-1-carboxylate (SMCC). The linking group can also be an additional amino acid sequence(s), including, for example, a polyalanine, polyglycine or similarly, linking group.

[0079] Alternatively, in some embodiments, the coding sequences of each polypeptide in
20 the hybrid polymerase are directly joined at their amino- or carboxy-terminus via a peptide bond in any order. Alternatively, an amino acid linker sequence may be employed to separate the first and second polypeptide components by a distance sufficient to ensure that each polypeptide folds into its secondary and tertiary structures. Such an amino acid linker sequence is incorporated into the fusion protein using standard techniques well known in the
25 art. Suitable peptide linker sequences may be chosen based on the following factors: (1) their ability to adopt a flexible extended conformation; (2) their inability to adopt a secondary structure that could interact with functional epitopes on the first and second polypeptides; and (3) the lack of hydrophobic or charged residues that might react with the polypeptide functional epitopes. Typical peptide linker sequences contain Gly, Ser, Val and Thr residues.
30 Other near neutral amino acids, such as Ala can also be used in the linker sequence. Amino acid sequences which may be usefully employed as linkers include those disclosed in Maratea et al. (1985) Gene 40:39-46; Murphy et al. (1986) Proc. Natl. Acad. Sci. USA 83:8258-8262; U.S. Patent Nos. 4,935,233 and 4,751,180. The linker sequence may generally be from 1 to about 50 amino acids in length, e.g., 3, 4, 6, or 10 amino acids in

length, but can be 100 or 200 amino acids in length. Linker sequences may not be required when the first and second polypeptides have non-essential N-terminal amino acid regions that can be used to separate the functional domains and prevent steric interference.

5 [0080] Other chemical linkers include carbohydrate linkers, lipid linkers, fatty acid linkers, polyether linkers, e.g., PEG, etc. For example, poly(ethylene glycol) linkers are available from Shearwater Polymers, Inc. Huntsville, Alabama. These linkers optionally have amide linkages, sulfhydryl linkages, or heterobifunctional linkages.

10 [0081] Other methods of linking the DNA polymerase and the non-specific DNA binding domain include ionic binding by expressing negative and positive tails and indirect binding through antibodies and streptavidin-biotin interactions. (See, e.g., *Bioconjugate Techniques*, supra). The domains may also be joined together through an intermediate interacting sequence. For example, an Sso7d-interacting sequence, i.e., a sequence that binds to Sso7d, can be joined to a polymerase. The resulting fusion protein can then be allowed to associate non-covalently with the Sso7d to generate an Sso7d-polymerase conjugate.

15 [0082] In some embodiments, a hybrid polymerase of the invention is produced by recombinant expression of a nucleic acid encoding the protein. Such a hybrid polymerase can be made by ligating the appropriate nucleic acid sequences encoding the desired amino acid sequences to each other by methods known in the art, in the proper coding frame, and expressing the product by methods known in the art.

20 [0083] Nucleic acids encoding the domains to be incorporated into the hybrid polymerases of the invention can be obtained using routine techniques in the field of recombinant genetics. Basic texts disclosing the general methods of use in this invention include Sambrook and Russell, *Molecular Cloning, A Laboratory Manual* (3rd ed. 2001); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology* (Ausubel et al., eds., 1994-1999).

25 [0084] Nucleic acid sequences that encode the DNA polymerase and the non-specific DNA binding domain polypeptides can be obtained using any of a variety of methods. In some embodiments, the nucleic acid sequences encoding the polypeptides are cloned from cDNA and genomic DNA libraries by hybridization with probes, or isolated using amplification techniques with oligonucleotide primers. More commonly, amplification techniques are used
30 to amplify and isolate the Sso7 and polymerase sequences using a DNA or RNA template (see, e.g., Dieffenbach & Dveksler, *PCR Primers: A Laboratory Manual* (1995)). Alternatively, overlapping oligonucleotides can be produced synthetically and joined to

produce one or more of the domains. Nucleic acids encoding catalytic or double-stranded nucleic acid binding domains can also be isolated from expression libraries using antibodies as probes.

[0085] In an example of obtaining a nucleic acid encoding a DNA polymerase or non-specific DNA binding domain using PCR, the nucleic acid sequence or subsequence is PCR amplified, using a sense primer containing one restriction site and an antisense primer containing another restriction site. This will produce a nucleic acid encoding the desired domain sequence or subsequence and having terminal restriction sites. This nucleic acid can then be ligated into a vector containing a nucleic acid encoding the second domain and having the appropriate corresponding restriction sites. The domains can be directly joined or may be separated by a linker, or other, protein sequence. Suitable PCR primers can be determined by one of skill in the art using the sequence information provided in GenBank or other sources. Appropriate restriction sites can also be added to the nucleic acid encoding the protein or protein subsequence by site-directed mutagenesis. The plasmid containing the domain-encoding nucleotide sequence or subsequence is cleaved with the appropriate restriction endonuclease and then ligated into an appropriate vector for amplification and/or expression according to standard methods.

[0086] Exemplary hybrid polymerases of the present invention include, for example, polymerases having the amino acid sequence of SEQ ID NO:5 or polymerases comprising an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

Assays to evaluate polymerase activity

[0087] Activity of a polymerase can be measured using a variety of assays that can be used to determine processivity or modification activity of a polymerase. Improvement in activity may include both increased processivity and increased efficiency.

[0088] The polymerases of the present invention, e.g., SEQ ID NO:5, exhibit polymerase activity, e.g., processivity, primer/template binding specificity, and 3' to 5' exonuclease activity. The activities can be measured using techniques that are standard in the art.

[0089] For example, polymerase processivity can be measured by a variety of methods known to those of ordinary skill in the art. Polymerase processivity is generally defined as the number of nucleotides incorporated during a single binding event of a modifying enzyme to a primed template. For example, a 5' FAM-labeled primer is annealed to circular or

linearized ssM13mp18 DNA to form a primed template. In measuring processivity, the primed template usually is present in significant molar excess to the polymerase so that the chance of any primed template being extended more than once by the polymerase is minimized. The primed template is therefore mixed with the polymerase at a ratio such as approximately 4000:1 (primed DNA:DNA polymerase) in the presence of buffer and dNTPs. MgCl₂ is added to initiate DNA synthesis. Samples are quenched at various times after initiation, and analyzed on a sequencing gel. At a polymerase concentration where the median product length does not change with time or polymerase concentration, the length corresponds to the processivity of the enzyme. The processivity of a protein of the invention, e.g., SEQ ID NO:2 or SEQ ID NO:5, is then compared to the processivity of a wild type enzyme.

[0090] Efficiency can be demonstrated by measuring the ability of an enzyme to produce product. Increased efficiency can be demonstrated by measuring the increased ability of an enzyme to produce product. Such an analysis measures the stability of the double-stranded nucleic acid duplex indirectly by determining the amount of product obtained in a reaction. For example, a PCR assay can be used to measure the amount of PCR product obtained with a short, e.g., 12 nucleotide in length, primer annealed at an elevated temperature, e.g., 50°C. In this analysis, enhanced efficiency is shown by the ability of a polymerase to produce more product in a PCR reaction using the 12 nucleotide primer annealed at 50°C.

[0091] Efficiency can also be measured, e.g., in a real-time PCR. The Ct value represents the number of cycles required to generate a detectable amount of DNA (a "detectable" amount of DNA is typically 2X, usually 5X, 10X, 100X or more above background). An efficient polymerase may be able to produce a detectable amount of DNA in a smaller number of cycles by more closely approaching the theoretical maximum amplification efficiency of PCR. Accordingly, a lower Ct value reflects a greater amplification efficiency for the enzyme.

[0092] Long PCR may be used as another method of demonstrating enhanced efficiency. For example, an enzyme with enhanced efficiency typically allows the amplification of a long amplicon (> 5 kb) in a shorter extension time compared to an enzyme with relatively lower efficiency.

[0093] Assays such as salt sensitivity can also be used to demonstrate improvement in efficiency or equivalent efficiency of a polymerase of the invention. A polymerase of the present invention may exhibit increased tolerance to high salt concentrations, i.e., a

processive enzyme with increased processivity can produce more product in higher salt concentrations. For example, a PCR analysis can be performed to determine the amount of product obtained in a reaction using a polymerase of the present invention compared to a wild type polymerase in reaction conditions with high salt, *e.g.*, 80 mM.

5 [0094] Other methods of assessing efficiency of the polymerases of the invention can be determined by those of ordinary skill in the art using standard assays of the enzymatic activity of a given modification enzyme.

[0095] Primer/template specificity is the ability of an enzyme to discriminate between matched primer/template duplexes and mismatched primer/template duplexes. Specificity
10 can be determined, for example, by comparing the relative yield of two reactions, one of which employs a matched primer, and one of which employs a mismatched primer. An enzyme with increased discrimination will have a higher relative yield with the matched primer than with the mismatched primer, *i.e.*, the ratio of the yield in the reaction using the matched primer vs. the reaction using the mismatched primer is about 1 or above. This ratio
15 can then be compared to the yield obtained in a parallel set of reactions employing a wild type polymerase.

Producing polymerases

[0096] Polymerases of the invention can be produced using techniques known in the art. Nucleic acids encoding the polymerase, and optionally linked to a DNA binding domain can
20 be obtained using routine techniques in the field of recombinant genetics. An exemplary nucleic acid encoding SEQ ID NO:2 is provided as SEQ ID NO:1. Depending on the host cell in which the polymerase is to be expressed, codon optimization can be employed to optimize expression of the polymerase in a particular host cell. Basic texts disclosing the general methods of use in this invention include Sambrook and Russell, *Molecular Cloning, A Laboratory Manual* (3rd ed. 2001); Kriegler, *Gene Transfer and Expression: A Laboratory Manual* (1990); and *Current Protocols in Molecular Biology* (Ausubel et al., eds., 1994-
25 1999). Such nucleic acids may also be obtained through *in vitro* amplification methods such as those described herein and in Berger, Sambrook, and Ausubel, as well as Mullis et al., (1987) U.S. Pat. No. 4,683,202; *PCR Protocols A Guide to Methods and Applications* (Innis et al., eds) Academic Press Inc. San Diego, Calif. (1990) (Innis); Arnheim & Levinson (Oct.
30 1, 1990) *C&EN* 36-47; *The Journal Of NIH Research* (1991) 3: 81-94; Kwoh et al. (1989) *Proc. Natl. Acad. Sci. USA* 86: 1173; Guatelli et al. (1990) *Proc. Natl. Acad. Sci. USA* 87, 1874; Lomell et al. (1989) *J. Clin. Chem.*, 35: 1826; Landegren et al., (1988) *Science* 241:

1077-1080; Van Brunt (1990) *Biotechnology* 8: 291-294; Wu and Wallace (1989) *Gene* 4: 560; and Barringer et al. (1990) *Gene* 89: 117, each of which is incorporated by reference in its entirety for all purposes and in particular for all teachings related to amplification methods.

5 [0097] One of skill will recognize that modifications can additionally be made to the polymerases of the present invention without diminishing their biological activity. Some modifications may be made to facilitate the cloning, expression, or incorporation of a domain into a fusion protein. Such modifications are well known to those of skill in the art and include, for example, the addition of codons at either terminus of the polynucleotide that
10 encodes the binding domain to provide, for example, a methionine added at the amino terminus to provide an initiation site, or additional amino acids (e.g., poly His) placed on either terminus to create conveniently located restriction sites or termination codons or purification sequences.

[0098] The polymerases of the present invention can be expressed in a variety of host cells,
15 including *E. coli*, other bacterial hosts, yeasts, filamentous fungi, and various higher eukaryotic cells such as the COS, CHO and HeLa cells lines and myeloma cell lines. Techniques for gene expression in microorganisms are described in, for example, Smith, *Gene Expression in Recombinant Microorganisms (Bioprocess Technology, Vol. 22)*, Marcel Dekker, 1994. Examples of bacteria that are useful for expression include, but are not limited
20 to, *Escherichia*, *Enterobacter*, *Azotobacter*, *Erwinia*, *Bacillus*, *Pseudomonas*, *Klebsiella*, *Proteus*, *Salmonella*, *Serratia*, *Shigella*, *Rhizobia*, *Vitreoscilla*, and *Paracoccus*. Filamentous fungi that are useful as expression hosts include, for example, the following genera: *Aspergillus*, *Trichoderma*, *Neurospora*, *Penicillium*, *Cephalosporium*, *Achlya*, *Podospora*, *Mucor*, *Cochliobolus*, and *Pyricularia*. See, e.g., U.S. Pat. No. 5,679,543 and Stahl and
25 Tudzynski, Eds., *Molecular Biology in Filamentous Fungi*, John Wiley & Sons, 1992. Synthesis of heterologous proteins in yeast is well known and described in the literature. *Methods in Yeast Genetics*, Sherman, F., et al., Cold Spring Harbor Laboratory, (1982) is a well recognized work describing the various methods available to produce the enzymes in yeast.

30 [0099] There are many expression systems for producing the polymerase polypeptides of the present invention that are well known to those of ordinary skill in the art. (See, e.g., *Gene Expression Systems*, Fernandex and Hoeffler, Eds. Academic Press, 1999; Sambrook and Russell, supra; and Ausubel et al, supra.) Typically, the polynucleotide encoding the polymerase is placed under the control of a promoter that is functional in the desired host

cell. Many different promoters are available and known to one of skill in the art, and can be used in the expression vectors of the invention, depending on the particular application. Ordinarily, the promoter selected depends upon the cell in which the promoter is to be active. Other expression control sequences such as ribosome binding sites, transcription termination sites and the like are also optionally included. Constructs that include one or more of these control sequences are termed "expression cassettes." Accordingly, the nucleic acids that encode the joined polypeptides are incorporated for high level expression in a desired host cell.

[0100] Expression control sequences that are suitable for use in a particular host cell are often obtained by cloning a gene that is expressed in that cell. Commonly used prokaryotic control sequences, which are defined herein to include promoters for transcription initiation, optionally with an operator, along with ribosome binding site sequences, include such commonly used promoters as the beta-lactamase (penicillinase) and lactose (lac) promoter systems (Change et al., Nature (1977) 198: 1056), the tryptophan (trp) promoter system (Goeddel et al., Nucleic Acids Res. (1980) 8: 4057), the tac promoter (DeBoer, et al., Proc. Natl. Acad. Sci. U.S.A. (1983) 80:21-25); and the lambda-derived PL promoter and N-gene ribosome binding site (Shimatake et al., Nature (1981) 292: 128). The particular promoter system is not critical to the invention, any available promoter that functions in prokaryotes can be used. Standard bacterial expression vectors include plasmids such as pBR322-based plasmids, e.g., pBLUESCRIPT™, pSKF, pET23D, lambda-phage derived vectors, and fusion expression systems such as GST and LacZ. Epitope tags can also be added to recombinant proteins to provide convenient methods of isolation, e.g., c-myc, HA-tag, 6-His (SEQ ID NO:15) tag, maltose binding protein, VSV-G tag, anti-DYKDDDDK (SEQ ID NO:16) tag, or any such tag, a large number of which are well known to those of skill in the art.

[0101] For expression in prokaryotic cells other than *E. coli*, a promoter that functions in the particular prokaryotic species is required. Such promoters can be obtained from genes that have been cloned from the species, or heterologous promoters can be used. For example, the hybrid trp-lac promoter functions in *Bacillus* sp. in addition to *E. coli*. These and other suitable bacterial promoters are well known in the art and are described, e.g., in Sambrook et al. and Ausubel et al. Bacterial expression systems for expressing the proteins of the invention are available in, e.g., *E. coli*, *Bacillus* sp., and *Salmonella* (Palva et al., Gene 22:229-235 (1983); Mosbach et al., Nature 302:543-545 (1983). Kits for such expression systems are commercially available.

[0102] Eukaryotic expression systems for mammalian cells, yeast, and insect cells are well known in the art and are also commercially available. In yeast, vectors include Yeast Integrating plasmids (e.g., YIp5) and Yeast Replicating plasmids (the YRp series plasmids) and pGPD-2. Expression vectors containing regulatory elements from eukaryotic viruses are typically used in eukaryotic expression vectors, e.g., SV40 vectors, papilloma virus vectors, and vectors derived from Epstein-Barr virus. Other exemplary eukaryotic vectors include pMSG, pAV009/A+, pMTO10/A+, pMAMneo-5, baculovirus pDSVE, and any other vector allowing expression of proteins under the direction of the CMV promoter, SV40 early promoter, SV40 later promoter, metallothionein promoter, murine mammary tumor virus promoter, Rous sarcoma virus promoter, polyhedrin promoter, or other promoters shown effective for expression in eukaryotic cells.

[0103] Either constitutive or regulated promoters can be used in the present invention. Regulated promoters can be advantageous because the host cells can be grown to high densities before expression of the fusion polypeptides is induced. High level expression of heterologous proteins slows cell growth in some situations. An inducible promoter is a promoter that directs expression of a gene where the level of expression is alterable by environmental or developmental factors such as, for example, temperature, pH, anaerobic or aerobic conditions, light, transcription factors and chemicals.

[0104] For *E. coli* and other bacterial host cells, inducible promoters are known to those of skill in the art. These include, for example, the lac promoter, the bacteriophage lambda P_L promoter, the hybrid trp-lac promoter (Amann et al. (1983) Gene 25: 167; de Boer et al. (1983) Proc. Nat'l. Acad. Sci. USA 80: 21), and the bacteriophage T7 promoter (Studier et al. (1986) J. Mol. Biol.; Tabor et al. (1985) Proc. Nat'l Acad. Sci. USA 82: 1074-8). These promoters and their use are also discussed in Sambrook et al., supra.

[0105] Translational coupling may be used to enhance expression. The strategy uses a short upstream open reading frame derived from a highly expressed gene native to the translational system, which is placed downstream of the promoter, and a ribosome binding site followed after a few amino acid codons by a termination codon. Just prior to the termination codon is a second ribosome binding site, and following the termination codon is a start codon for the initiation of translation. The system dissolves secondary structure in the RNA, allowing for the efficient initiation of translation. See Squires, et. al. (1988), J. Biol. Chem. 263: 16297-16302.

[0106] The construction of polynucleotide constructs can involve the use of vectors able to replicate in bacteria. Such vectors are commonly used in the art. A plethora of kits are commercially available for the purification of plasmids from bacteria (for example, EasyPrep™, FlexiPrep™, from Pharmacia Biotech; StrataClean™, from Stratagene; and, 5 QIAexpress® Expression System, Qiagen). The isolated and purified plasmids can then be further manipulated to produce other plasmids, and used to transform cells.

[0107] The polypeptides of the invention can be expressed intracellularly, or can be secreted from the cell. Intracellular expression often results in high yields. If necessary, the amount of soluble, active fusion polypeptide may be increased by performing refolding 10 procedures (see, e.g., Sambrook et al., supra.; Marston et al., Bio/Technology (1984) 2: 800; Schoner et al., Bio/Technology (1985) 3: 151). Polypeptides of the invention can be expressed in a variety of host cells, including E. coli, other bacterial hosts, yeast, and various higher eukaryotic cells such as the COS, CHO and HeLa cells lines and myeloma cell lines. The host cells can be mammalian cells, insect cells, or microorganisms, such as, for example, 15 yeast cells, bacterial cells, or fungal cells.

[0108] Once expressed, the polypeptides can be purified according to standard procedures of the art, including ammonium sulfate precipitation, affinity columns, column chromatography, gel electrophoresis and the like (see, generally, R. Scopes, Protein Purification, Springer-Verlag, N.Y. (1982), Deutscher, Methods in Enzymology Vol. 182: 20 Guide to Protein Purification., Academic Press, Inc. N.Y. (1990)). In some embodiments, substantially pure compositions of at least about 90 to 95% homogeneity are preferred, and 98 to 99% or more homogeneity are most preferred. Once purified, partially or to homogeneity as desired, the polypeptides may then be used (e.g., in DNA amplification).

[0109] To facilitate purification of the polypeptides of the invention, the nucleic acids that 25 encode the polypeptides can also include a coding sequence for an epitope or "tag" for which an affinity binding reagent is available. Examples of suitable epitopes include the myc and V-5 reporter genes; expression vectors useful for recombinant production of fusion polypeptides having these epitopes are commercially available (e.g., Invitrogen (Carlsbad, CA) vectors pcDNA3.1/Myc-His and pcDNA3.1/V5-His are suitable for expression in 30 mammalian cells). Additional expression vectors suitable for attaching a tag to the fusion proteins of the invention, and corresponding detection systems are known to those of skill in the art, and several are commercially available (e.g., FLAG" (Kodak, Rochester, NY)). Another example of a suitable tag is a polyhistidine sequence, which is capable of binding to metal chelate affinity ligands. Typically, six adjacent histidines are used, although one can

use more or less than six. Suitable metal chelate affinity ligands that can serve as the binding moiety for a polyhistidine tag include nitrilo-tri-acetic acid (NTA) (Hochuli, E. (1990) "Purification of recombinant proteins with metal chelating adsorbents" In Genetic Engineering: Principles and Methods, J. K. Setlow, Ed., Plenum Press, N.Y.; commercially available from Qiagen (Santa Clarita, CA)).

[0110] One of skill in the art will recognize that after biological expression or purification, the polymerase conjugates(s) may possess a conformation substantially different than the native conformations of the constituent polypeptides. In this case, it may be necessary or desirable to denature and reduce the polypeptide and then to cause the polypeptide to re-fold into the preferred conformation. Methods of reducing and denaturing proteins and inducing re-folding are well known to those of skill in the art (See, Debinski et al. (1993) J. Biol. Chem. 268: 14065-14070; Kreitman and Pastan (1993) Bioconjug. Chem. 4: 581-585; and Buchner et al. (1992) Anal. Biochem. 205: 263-270). Debinski et al., for example, describe the denaturation and reduction of inclusion body proteins in guanidine-DTE. The protein is then refolded in a redox buffer containing oxidized glutathione and L-arginine.

III. Methods

[0111] In another aspect, the present invention provides for methods of amplifying a target nucleic acid using a polymerase of the invention (*e.g.*, a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2 or a polymerase linked to a sequence non-specific DNA binding domain). Such amplification reactions include without limitation the polymerase chain reaction (PCR). Polymerase chain reactions that can be conducted using the compositions described herein include, without limitation, reverse-transcription PCR (rt-PCR) and quantitative PCR (qPCR).

[0112] In some embodiments, the methods of the present invention are useful for amplifying a long amplicon, *e.g.*, an amplicon of at least 5 kb, at least 7 kb, at least 10 kb, at least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb. In some embodiments, the methods of amplifying a target nucleic acid using a polymerase of the invention (*e.g.*, a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2 or a polymerase linked to a sequence non-specific DNA binding domain) result in a higher efficiency for amplification of an amplicon of at least 5 kb, at least 7 kb, at least 10 kb, at

least 12 kb, at least 15 kb, at least 17 kb, at least 20 kb, at least 22 kb, at least 25 kb, at least 27 kb, at least 30 kb, at least 35 kb, or at least 40 kb as compared to a DNA polymerase consisting of SEQ ID NO:7 or SEQ ID NO:8.

[0113] The amplification methods of the present invention using polymerases (*e.g.*, a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2 or a polymerase linked to a sequence non-specific DNA binding domain) are carried out using reaction mixtures that are sufficient for amplifying a nucleic acid molecule. In some embodiments, an amplification reaction mixture comprises, in addition to the DNA polymerase or hybrid polymerase, one or more of the following components: nucleotide triphosphates, one or more oligonucleotide primers, salt, buffer, water, stabilizer, and DNA-binding dye.

[0114] In some embodiments, the amplification methods of the present invention comprise using polymerases (*e.g.*, a polypeptide comprising a polypeptide DNA polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2 or a polymerase linked to a sequence non-specific DNA binding domain) with an agent that improves amplification specificity, for example an agent selected from arginine, spermidine, and spermine. In some embodiments, an amplification reaction mixture comprises, in addition to the DNA polymerase or hybrid polymerase, one or more of the following components: nucleotide triphosphates, one or more oligonucleotide primers, salt, buffer, water, stabilizer, and DNA-binding dye; and an agent selected from arginine, spermidine, or spermine.

[0115] In some embodiments, an amplification reaction mixture of the present invention comprises: a polymerase (*e.g.*, a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2 or a polymerase linked to a sequence non-specific DNA binding domain) as described herein at a concentration of about 1 U/ml to about 75 U/ml (*e.g.*, about 1 U/ml, 5 U/ml, 10 U/ml, 15 U/ml, 20 U/ml, 25 U/ml, 30 U/ml, 35 U/ml, 40 U/ml, 45 U/ml, 50 U/ml, 55 U/ml, 60 U/ml, 65 U/ml, 70 U/ml, or 75 U/ml); dNTPs at a concentration of about 0.1 mM to about 10 mM (*e.g.*, about 0.1 mM, 0.2 mM, 0.3 mM, 0.4 mM, 0.5 mM, 0.6 mM, 0.7 mM, 0.8 mM, 0.9 mM, 1 mM, 2 mM, 3 mM, 4 mM, 5 mM, 6 mM, 7 mM, 8 mM, 9 mM, or 10 mM); magnesium, *e.g.*, MgCl₂, at a concentration of about 1 mM to about 20 mM (*e.g.*, about 1 mM, 2 mM, 3 mM, 4 mM, 5 mM, 6 mM, 7 mM, 8 mM, 9 mM, 10 mM, 11 mM, 12 mM, 13 mM, 14 mM, 15 mM, 16 mM, 17 mM, 18 mM, 19 mM, or 20 mM); (NH₄)₂SO₄ at

a concentration of about 10 mM to about 100 mM (*e.g.*, about 10 mM, 20 mM, 30 mM, 40 mM, 50 mM, 60 mM, 70 mM, 80 mM, 90 mM, or 100 mM); potassium, *e.g.*, KCl, at a concentration of about 50 mM to about 200 mM (*e.g.*, about 50 mM, 60 mM, 70 mM, 80 mM, 90 mM, 100 mM, 110 mM, 120 mM, 130 mM, 140 mM, 150 mM, 160 mM, 170 mM, 180 mM, 190 mM, or 200 mM); a buffer, *e.g.*, Tris pH 8.5-9.5 at a concentration of about 50 mM to about 200 mM (*e.g.*, about 50 mM, 60 mM, 70 mM, 80 mM, 90 mM, 100 mM, 110 mM, 120 mM, 130 mM, 140 mM, 150 mM, 160 mM, 170 mM, 180 mM, 190 mM, or 200 mM); a disaccharide, *e.g.*, trehalose, at a concentration of about 100 mM to about 500 mM (*e.g.*, about 100 mM, 125 mM, 150 mM, 175 mM, 200 mM, 225 mM, 250 mM, 275 mM, 300 mM, 325 mM, 350 mM, 375 mM, 400 mM, 425 mM, 450 mM, 475 mM, or 500 mM); one or more osmolytes, *e.g.*, sarcosine, trimethylamine N-oxide (TMAO), dimethylsulfoniopropionate, and trimethylglycine, at a concentration of about 50 mM to about 200 mM (*e.g.*, about 50 mM, 60 mM, 70 mM, 80 mM, 90 mM, 100 mM, 110 mM, 120 mM, 130 mM, 140 mM, 150 mM, 160 mM, 170 mM, 180 mM, 190 mM, or 200 mM); Tween-20 at a concentration of about 0.1% to about 0.5% (*e.g.*, about 0.1%, 0.2%, 0.3%, 0.4%, or 0.5%); glycerol at a concentration of about 1% to about 5% (*e.g.*, about 1%, 2%, 3%, 4%, or 5%); DMSO at a concentration of about 1% to about 10% (*e.g.*, about 1%, 1.5%, 2%, 2.5%, 3%, 3.5%, 4%, 4.5%, 5%, 5.5%, 6%, 6.5%, 7%, 7.5%, 8%, 8.5%, 9%, 9.5%, or 10%); fluorescein at a concentration of about 0.001% to about 0.01% (*e.g.*, about 0.001%, 0.002%, 0.003%, 0.004%, 0.005%, 0.006%, 0.007%, 0.008%, 0.009%, or 0.01%); DNA binding dye (*e.g.*, cyanine dye) at a concentration of about 0.5X to about 5X (*e.g.*, about 0.5X, 0.6X, 0.7X, 0.8X, 0.9X, 1X, 1.5X, 2X, 2.5X, 3X, 3.5X, 4X, 4.5X, or 5X); and optionally arginine, spermidine, or spermine or a salt thereof at a concentration of about 1 mM to about 100 mM (*e.g.*, about 1 mM, 5 mM, 10 mM, 15 mM, 20 mM, 25 mM, 30 mM, 35 mM, 40 mM, 45 mM, 50 mM, 55 mM, 60 mM, 65 mM, 70 mM, 75 mM, 80 mM, 85 mM, 90 mM, 95 mM, or 100 mM).

[0116] Improvements in efficiency and specificity due to certain aspects of the present invention can be identified and quantified using assays known in the art and described in further detail below.

[0117] In some embodiments, dye-based qPCR detection methods are used to monitor amplification reactions utilizing components of the invention. Such detection methods generally rely on monitoring the increase in fluorescence signal due to the binding of DNA-binding dye to the amplified DNA. For example, SYBR Green I, a commonly used fluorescent DNA binding dye, binds all double-stranded DNA and detection is monitored by

measuring the increase in fluorescence throughout the cycle. SYBR Green I has an excitation and emission maxima of 494 nm and 521 nm, respectively.

[0118] In other embodiments, probe-based qPCR detection methods are used to monitor amplification reactions utilizing components of the invention. Such detection methods
5 generally rely on the sequence-specific detection of a desired PCR product. Unlike dye-based qPCR methods that detect all double-stranded DNA, probe-based qPCR utilizes a fluorescent-labeled target-specific probe, which detects specific sequences in the amplified DNA.

[0119] In certain aspects, it may be desirable to include an additional compound as an
10 additive to improve efficiency in amplification reactions, including but not limited to qPCR. In some embodiments, inclusion of the additive is sufficient to increase efficiency of the polymerase conjugate by at least 5, 10, 15, 20, 25, 35, 40, or 50% or more compared to a control mixture lacking the additive.

[0120] In some embodiments, a polymerase conjugate of the invention exhibits low
15 efficiency for certain targets when used in a formulation that includes certain binding dyes (such as, for example, an EvaGreen DNA binding dye). Such low efficiency may in some embodiments result in a delay of Ct values associated with low input DNA concentrations. Methods for measuring efficiency of a particular reaction are known in the art and described in further detail below.

[0121] In some embodiments, the additive is an osmolyte included in an amplification
20 reaction of the invention to improve efficiency. Members of the osmolyte family have been shown to improve the thermal stability of proteins (Santoro, Biochemistry, 1992) as well as decrease DNA double helix stability (Chadalavada, FEBS Letters, 1997). In some embodiments, osmolytes are small molecules or compounds which are produced by living
25 organisms in response to environmental stresses such as extreme temperatures, dehydration, or salinity and which protect their cellular components and help to maintain optimal cytosolic conditions. Osmolytes of use in the present invention may include without limitation sarcosine, trimethylamine N-oxide (TMAO), dimethylsulfoniopropionate, and trimethylglycine. Sarcosine is chemically similar to betaine, a chemical which has been
30 shown to improve conventional PCR (Henke, Nucleic Acids Research, 1997).

[0122] In conventional uses of osmolytes, the stabilizing effects of such compounds are generally observed at relatively high concentrations (> 1M). However, in methods of the present invention, millimolar concentrations of osmolytes have been found to be effective for

improving the reaction efficiency of amplification reactions such as qPCR. Without being bound by a mechanism of action, it is possible that the improvement in efficiency is the result of improving the accessibility of the DNA polymerase to the targeted region of the DNA template for reactions that contain low concentrations of input DNA sample. In some
5 embodiments, concentrations of about 100 to about 1000 mM of osmolytes are used in methods and kits of the present invention. In still further embodiments, concentrations of about 50 to about 700, about 100 to about 600, about 150 to about 500, about 200 to about 400 mM, and about 300 to about 350 mM osmolytes are used in methods and kits of the invention. In some embodiments, the osmolyte used in methods, reaction mixtures, and kits
10 of the invention is sarcosine (optionally at the above-listed concentrations).

IV. Reaction Mixtures

[0123] In another aspect, the present invention provides reaction mixtures comprising a polymerase of the invention. In some embodiments, a reaction mixture comprises a DNA polymerase having the amino acid sequence of SEQ ID NO:2 or a DNA polymerase
15 comprising an amino acid sequence that is at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2. In some embodiments, a reaction mixture comprises a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% identical to SEQ ID NO:2 linked to a non-specific DNA binding domain, e.g., a non-specific DNA binding domain comprising an amino acid sequence that is at least 80%
20 identical to SEQ ID NO:4 or SEQ ID NO:10. In some embodiments, a reaction mixture comprises a polypeptide having the amino acid sequence of SEQ ID NO:5 or a polypeptide comprising an amino acid sequence that is at least 80% (or at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

25 [0124] The reaction mixture can further comprise one or more reagents for amplification of a target nucleic acid, including but not limited to a target nucleic acid, one or more oligonucleotides, buffers, nucleotide triphosphates, salts, stabilizers, one or more additives for improving efficiency, nuclease-free water, and/or a double stranded DNA binding dye.

[0125] In some embodiments, the reaction mixture can comprise a sufficient amount of an
30 agent to improve the specificity of nucleic acid amplification. For example, in some embodiments, the agent is selected from arginine (*e.g.*, L-arginine or D-arginine), spermidine, and spermine, or a salt thereof. In some embodiments, the reaction mixture can comprise an additional compound as an additive to improve efficiency in amplification reactions. In some

embodiments, the additive is an osmolyte, including but not limited to sarcosine, trimethylamine N-oxide (TMAO), dimethylsulfoniopropionate, or trimethylglycine.

V. Kits

[0126] In another aspect, the present invention provides kits for conducting nucleic acid
5 amplification reactions. The kits include a polymerase of the invention (*e.g.*, a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% identical to SEQ ID NO:2, or a polypeptide comprising a DNA polymerase having an amino acid sequence at least 95% identical to SEQ ID NO:2 linked to a non-specific DNA binding domain). In some
10 embodiments, a kit comprises a DNA polymerase having the amino acid sequence of SEQ ID NO:2 or a DNA polymerase comprising an amino acid sequence that is at least 95% (or at least 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:2. In some
embodiments, a kit comprises a polypeptide having the amino acid sequence of SEQ ID NO:5 or a polypeptide comprising an amino acid sequence that is at least 80% (or at least
85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least
15 96%, at least 97%, at least 98%, or at least 99%) identical to SEQ ID NO:5.

[0127] Optionally, the kits comprise one or more dNTPs, at least one buffer, and/or a double stranded DNA binding dye. Such kits may also include stabilizers and other additives (*e.g.*, sarcosine) to increase the efficiency of the amplification reactions. Such kits may also include one or more primers as well as instructions for conducting nucleic acid amplification
20 reactions using the components of the kits. Optionally, the kit can further comprise a sufficient amount of an agent to improve the specificity of nucleic acid amplification. For example, in some embodiments, the agent is selected from arginine (*e.g.*, L-arginine or D-arginine), spermidine, and spermine, or a salt thereof.

EXAMPLES

25 [0128] The following examples are offered to illustrate, but not to limit the claimed invention.

Example 1

[0129] Lysates from 96 clones from a library of hybrid polymerase variants were tested for DNA polymerase activity, 3'-5' exonuclease activity and level of protein expression. Clones
30 exhibiting >1 ratio of polymerase activity over exonuclease activity were selected for further purification and characterizations. Among the several clones tested in PCR using equal units of purified enzymes, a variant designated "H10" exhibited improved ability to amplify long

amplicons (20kb). *See*, Figure 1. Similar results were observed for 7.5 kb amplicons, with variant H10 producing a noticeably more 7.5kb product than a control hybrid polymerase (SEQ ID NO:8).

5 [0130] A polymerase fidelity test was performed on H10 along with Taq polymerase and a control hybrid polymerase (SEQ ID NO:8). The results show that H10 has significantly higher fidelity compare to Taq DNA polymerase, and has comparable or slightly higher fidelity than a previously characterized hybrid polymerase clone.

10 [0131] The amino acid sequence of H10 (SEQ ID NO:5) is made up of a polymerase sequence (SEQ ID NO:2) and a sequence non-specific DNA binding domain (SEQ ID NO:4). Figure 2 illustrates an alignment of the polymerase sequence of H10 with the polymerase sequence of the control hybrid polymerase. Without intending to limit the scope of the invention, one or more of the differences in the amino acid sequences of the two polymerases are believed to result in the increased efficiency observed in the H10 variant compared to the control.

15 [0132] It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims. All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all
20 purposes.

SEQUENCE LISTING

SEQ ID NO:1 (coding sequence for SEQ ID NO:2)

ATGATCCTGGATGCTGACTACATCACTGAAGAAGGCCAAACCGGTTATCCGTCTGT
 TCAAAAAGAGAACGGCAAATTTAAGATTGAGCATGATCGCACCTTTTCGTCCAT
 5 ACATTTACGCTCTGCTGAAAGATGATTCTAAGATTGATGAAGTTAGAAAAATCAC
 TGGTGAGCGCCATGGCAAAGATTGTTTCGTATCGTTGATGCGGAAAAGGTAGAAAA
 GAAATTTCTGGGCAGACCAATCGAGGTGTGGAAACTGTATCTCGAACATCCACA
 AGATGTTCCGGCTATTCGCGATAAAGTTCGCGAACATCCTGCAGTTGTTGACATC
 10 TTCGAATACGATATTCCATTTGCAAAGCGTTACCTCATCGACAAAGGCCTGATAC
 CAATGGAGGGCGAGGAAGAACTCAAGATCCTGGCGTTCGATATAGAAACCCTCT
 ATCACGAAGGCGAAGAGTTTGGTAAAGGCCCAATTATAATGATCAGCTATGCAG
 ATGAAAACGAAGCAAAGGTGATTACTTGGAAAAACATAGATCTCCATACGTTG
 AGGTTGTATCTTCCGAGCGCGAGATGATTAAGCGCTTTCTCAGAGTTATCCGCGA
 GAAGGATCCGGACGTTATCATTACTTATAACGGCGACTCTTTTGACTTCCCATAT
 15 CTGGTGAAACGCGCAGAAAACTCGGTATTAAGTACTGACTATCGGCCGTGATGGT
 TCCGAGCCGAAGATGCAGCGTCTCGGCATATGACCGCTGTAGAAATTAAGGGT
 CGTATCCATTTTCGACCTGTATCATGTAATTCGTCTACTATTAACCTCCCGACTTA
 CACTCTCGAGGCTGTATATGAAGCAATTTTTGGTAAGCCGAAGGAGAAGGTATA
 CGCCGATGAGATTGCAGAGGCGTGGGAATCCGGTGAGGGCCTCGAGCGTGTGTC
 20 AAAATACTCCATGGAAGATGCAAAGGTGACTTATGAACTCGGCCAAAGAATTCCT
 CCAATGGAAATCCAGCTCTCTCGCCTGGTTGGCCAACCACTGTGGGATGTTTCT
 CGTTCTTCCACCGGTAACCTCGTAGAGTGGTTTCTCCTGCGCAAAGCGTACGAAC
 GCAACGAAGTGGCTCCGAACAAGCCATCTGAAGAAGAGTATGAACGCCGTCTCC
 GCGAGTCTTACGCTGGTGGCTATGTTAAAGAGCCAGAAAAGGGCCTCTGGGAAA
 25 ACATCGTGTCCCTCGATTTTCGCTCTCTGTATCCGTCTATTATCATTACCCACAAC
 GTGTCTCCGGATACTCTCAACCGCGAGGGCTGCAGAGAGTATGATATTGCTCCGG
 AAGTAGGCCACAAGTTCTGCAAGGACTTCCCGGGCTTTATTCCGTCTCTCCTGAA
 ACATCTGCTCGAGGAACGCCAAAAGATTAAGACTAAAATGAAGGAGTCCAGGA
 TCCGATTGAAAAAAAATGCTCGACTATCGCCAAAGAGCGATTAATAATCCTCGC
 30 AACTCTTTTTACGGCTATTATGGCTATGCAAAGCACGCTGGTACTGTAAGGAG
 TGTGCTGAGTCCGTTACTGCTTGGGGTCGCAAATACATCGAGTTCGTGTGGAAGG
 AGCTCGAAGAAAAGTTTGGCTTTAAAGTTCTCTACATTGACACTGATGGTCTCTA
 TCGACTATTCCGGGTGGTGAGCCTGAGGAAATTAAGAAAAGGCTCTAGAATT
 TGTGAAATACATTAAGCTCGAAGCTCCCGGGTCTCCTGGAGCTCGAATATGAAGGC
 35 TTTTATGTTTCGCGGCTTCTTCGTTACCAAGAAGAGATATGCGGTGATTGATGAAG
 AAGGCAAAGTTACTCGTGGTCTCGAGATTGTGCGCCGTGATTGGAGCGAAAT
 TGCGAAAGAACTCAAGCTAGAGTTCTCGAGACTATTCTCAAACACGGCAACGT
 TGAAGAAGCTGTGAAAATTGTAAGAAGTAACCCAAAAGCTCGCTAACTATGA
 AATTCCGCCAGAGAAGCTCGCGATTTATGAGCAGATTACTCGCCCGCTGCATGAG
 40 TATAAGGCGATTGGTCCGCACGTGGCTGTTGCAAAGAACTGGCTGCTAGAGGC
 GTGAAAATTAACCGGGTATGGTAATTGGCTACATTGTACTCCGCGGCGATGGTC
 CGATTAGCAAACGTGCAATTCTAGCTGAGGAATTCGATCCGAGAAAGCACAAAGT
 ATGACGCAGAATATTACATTGAGAACCAGGTGCTCCCGGCGGTACTCCGTATTCT
 GGAGGGTTTTGGCTACCGTAAGGAAGACCTCCGCTGGCAAAGACTAAACAGGT
 45 TGGCCTCACTTCTTGGCTCAACATTAATAAAATCCGG

SEQ ID NO:2 (polymerase sequence)

MILDADYITEEGKPVIRLFKKEKFKIEHDRTFRPYIYALLKDDSKIDEVRKITGERH
 GKIVRIVDAEKVEKKFLGRPIEVWKLYLEHPQDVP AIRDKVREHPAVVDIFEYDIPFA
 50 KRYLIDKGLIPMEGEEELKILAFDIETLYHEGEEFVGKGPIMISYADENEAKVITWKNID
 LPYVEVVSSSEREMIKRFLRVIREKDPDVIITYNGDSFDFPYLVKRAEKLGIKLTIGRDG

SEPQMQRDLGDMTAVEIKGRIHFDLYHVIRRTINLPTYTLEAVYEAIFGKPKKVKYADE
 IAEAWESGEGLERVAKYSMEDAKVTYELGKEFLPMEIQLSRLVGQPLWDVSRSSSTG
 NLVEWFLLRKAYERNEVAPNKPSEEEYERRLRESYAGGYVKEPEKGLWENIVSLDF
 RSLYPSIIITHNVSPDTLNREGCREYDIAPEVGHKFKDFPGFIPSLKHLLEERQKIKT
 5 KMKESQDPIEKKMLDYRQRAIKILANSFYGYGYAKARWYCKECAESVTAWGRKY
 IEFVWKELEEKFGFKVLYIDTDGLYATIPGGEPEEIKKALEFVKYINSLKPLGLELEY
 EGFYVRGFFVTKKRYAVIDEEGKVITRGLIVRRDWSEIAKETQARVLETILKHGNVE
 EAVKIVKEVTQKLANYEIPPEKLAIYEQITRPLHEYKAIGPHVAVAKKLAARGVKIKP
 GMVIGYIVLRGDGPISKRAILAEFDPKHKYDAEYYIENQVLPVLRILEGFGYRKE
 10 DLRWQKTKQVGLTSLWLNKKS

SEQ ID NO:3

Linker

gtgggg

15

SEQ ID NO:4

Exemplary non-specific DNA binding domain

Atvkfkykgeekevdiskikkvwrvgkmisftydegggktrgavsekdpkellqmlkqkk

SEQ ID NO:5

Exemplary hybrid polymerase comprising non-specific DNA binding domain

MILDADYITEEGKPVIRLFKENGKFKIEHDRTFRPYIYALLKDDSKIDEVRKITGERH
 GKIVRIVDAEKVEKKFLGRPIEVWKLYLEHPQDVPVPAIRDKVREHPAVVDIFEYDIPFA
 KRYLIDKGLIPMEGEEELKILAFDIETLYHEGEEFGKGPIMISYADENEAKVITWKNID
 25 LPYVEVVSSEREMIKRFLRVIREKDPDVIITYNGDSFDFPYLVKRAEKLGIKLTIGRDG
 SEPQMQRDLGDMTAVEIKGRIHFDLYHVIRRTINLPTYTLEAVYEAIFGKPKKVKYADE
 IAEAWESGEGLERVAKYSMEDAKVTYELGKEFLPMEIQLSRLVGQPLWDVSRSSSTG
 NLVEWFLLRKAYERNEVAPNKPSEEEYERRLRESYAGGYVKEPEKGLWENIVSLDF
 RSLYPSIIITHNVSPDTLNREGCREYDIAPEVGHKFKDFPGFIPSLKHLLEERQKIKT
 30 KMKESQDPIEKKMLDYRQRAIKILANSFYGYGYAKARWYCKECAESVTAWGRKY
 IEFVWKELEEKFGFKVLYIDTDGLYATIPGGEPEEIKKALEFVKYINSLKPLGLELEY
 EGFYVRGFFVTKKRYAVIDEEGKVITRGLIVRRDWSEIAKETQARVLETILKHGNVE
 EAVKIVKEVTQKLANYEIPPEKLAIYEQITRPLHEYKAIGPHVAVAKKLAARGVKIKP
 GMVIGYIVLRGDGPISKRAILAEFDPKHKYDAEYYIENQVLPVLRILEGFGYRKE
 35 DLRWQKTKQVGLTSLWLNKKSgtggggatvkfkykgeekevdiskikkvwrvgkmisftydegggktrgav
 sekdpkellqmlkqkk

SEQ ID NO:6

Nucleic acid sequence encoding SEQ ID NO:5

40 ATGATCCTGGATGCTGACTACATCACTGAAGAAGGCAAACCGGTTATCCGTCTGT
 TCAAAAAGAGAACGGCAAATTTAAGATTGAGCATGATCGCACCTTTTCGTCCAT
 ACATTTACGCTCTGCTGAAAGATGATTCTAAGATTGATGAAGTTAGAAAAATCAC
 TGGTGAGCGCCATGGCAAGATTGTTTCGTATCGTTGATGCGGAAAAGGTAGAAA
 GAAATTTCTGGGCAGACCAATCGAGGTGTGGAAACTGTATCTCGAACATCCACA
 45 AGATGTTCCGGCTATTCGCGATAAAGTTCGCGAACATCCTGCAGTTGTTGACATC
 TTCGAATACGATATTCCATTTGCAAAGCGTTACCTCATCGACAAAGGCCTGATAC
 CAATGGAGGGCGAGGAAGAACTCAAGATCCTGGCGTTCGATATAGAAACCCTCT
 ATCACGAAGGCGAAGAGTTTGGTAAAGGCCCAATTATAATGATCAGCTATGCAG
 ATGAAAACGAAGCAAAGGTGATTACTTGGAAAAACATAGATCTCCCATACGTTG
 50 AGGTTGTATCTTCCGAGCGGAGATGATTAAGCGCTTTCTCAGAGTTATCCGCGA
 GAAGGATCCGGACGTTATCATTACTTATAACGGCGACTCTTTTGACTTCCCATAT
 CTGGTGAAACGCGCAGAAAACTCGGTATTAAGTACTGACTATCGGCCGTGATGTT

TCCGAGCCGAAGATGCAGCGTCTCGGCGATATGACCGCTGTAGAAATTAAGGGT
 CGTATCCATTTTCGACCTGTATCATGTAATTCGTCTACTATTanAACCTCCC GACTT
 AACTCTCGAGGCTGTATATGAAGCAATTTTTGGTAAGCCGAAGGAGAAGGTAT
 ACGCCGATGAGATTGCAGAGGCGTGGGAATCCGGTGAGGGCCTCGAGCGTGTTG
 5 CAAAATACTCCATGGAAGATGCAAAGGTGACTTATGAACTCGGCAAAGAATTCC
 TCCCAATGGAATCCAGCTCTCTCGCCTGGTTGGCCAACCACTGTGGGATGTTTC
 TCGTTCTTCCACCGGTAACCTCGTAGAGTGGTTTCTCCTGCGCAAAGCGTACGAA
 CGCAACGAAGTGGCTCCGAACAAGCCATCTGAAGAAGAGTATGAACGCCGTCTC
 10 CGCGAGTCTTACGCTGGTGGCTATGTTAAAGAGCCAGAAAAGGGCCTCTGGGAA
 AACATCGTGTCCCTCGATTTTCGCTCTCTGTATCCGTCTATTATCATTACCCACAA
 CGTGTCTCCGGATACTCTCAACCGCGAGGGGCTGCAGAGAGTATGATATTGCTCCG
 GAAGTAGGCCACAAGTTCTGCAAGGACTTCCCGGGCTTTATTCCGTCTCTCCTGA
 AACATCTGCTCGAGGAACGCCAAAAGATTAAGACTAAAATGAAGGAGTCCCAGG
 ATCCGATTGAAAAAAAATGCTCGACTATCGCCAAAGAGCGATTA AAAATCCTCG
 15 CAAACTCTTTTTACGGCTATTATGGCTATGCAAAGCACGCTGGTACTGTAAGGA
 GTGTGCTGAGTCCGTTACTGCTTGGGGTTCGCAAATACATCGAGTTCGTGTGGAAG
 GAGCTCGAAGAAAAGTTTGGCTTTAAAGTTCTCTACATTGACACTGATGGTCTCT
 ATGCGACTATTCCGGGTGGTGAGCCTGAGGAAATTAAGAAAAAGGCTCTAGAAT
 TTGTGAAATACATTA ACTCGAAGCTCCCGGGTCTCCTGGAGCTCGAATATGAAGG
 20 CTTTTATGTTTCGCGGCTTCTTCGTTACCAAGAAGAGATATGCGGTGATTGATGAA
 GAAGGCAAAGTTACTACTCGTGGTCTCGAGATTGTGCGCCGTGATTGGAGCGAA
 ATTGCGAAAGAACTCAAGCTAGAGTTCTCGAGACTATTCTCAAACACGGCAAC
 GTTGAAGAAGCTGTGAAAATTGTA AAAAGAAGTAACCCAAAAGCTCGCTAACTAT
 GAAATTCGCCAGAGAAGCTCGCGATTTATGAGCAGATTACTCGCCCGCTGCATG
 25 AGTATAAGGCGATTGGTCCGCACGTGGCTGTTGCAAAGAAACTGGCTGCTAGAG
 GCGTGAAAATTAACCGGGTATGGTAATTGGCTACATTGTA CTCCGCGGCGATGG
 TCCGATTAGCAAACGTGCAATTCTAGCTGAGGAATTCGATCCGAGAAAGCACAA
 GTATGACGCAGAATATTACATTGAGAACCAGGTGCTCCCGGCGGTACTCCGTATT
 CTGGAGGGTTTTGGCTACCGTAAGGAAGACCTCCGCTGGCAA AAGACTAAACAG
 30 GTTGGCCTCACTTCTTGGCTCAACATTA AAAAATCCGGTACCGGCGGTGGCGGTG
 CAACCGTAAAGTTCAAGTACAAAGGCGAAGAAAAAGAGGTAGACATCTCCAAG
 ATCAAGAAAGTATGGCGTGTGGGCAAGATGATCTCCTTACCTACGACGAGGGC
 GGTGGCAAGACCGGCCGTGGTGC GGTAAGCGAAAAGGACGCGCCGAAGGAGCT
 GCTGCAGATGCTGGAGAAGCAGAAAAAGTGAGCTAGC

35 **SEQ ID NO:7** (Control polymerase)
 MILDADYITEEGKPVIRLFKKENGEFKIEHRTFRPYIYALLKDDSKIEEVKKITAERH
 GKIVRIVDAEKVEKKFLGRPITVWRLYFEHPQDVPTIREKIREHSAVVDIFEYDIPFAK
 RYLIDKGLIPMEGDEELKLLAFDIETLYHEGEEFGKGPIMISYADEEEAKVITWKKID
 40 LPYVEVVSSEREMIKRFLKIIREKDPDIIITYNGDSFDLPYLAKRAEKLGIKLTIGRDGS
 EPKMQRIGDMTAVEVKGRIHFDLYHVIRRTINLPTYTLEAVYEAIFGKPKEKVYADEI
 AKAWETGEGLERVAKYSMEDAKATYELGKEFFPMEAQLSRLVGQPLWDVSRSTG
 NLVEWFLLRKAYERNELAPNKPDEREYERRLRRESYAGGFVKEPEKGLWENIVSLDFR
 ALYPSIIITHNVSPDTLNREGCRNYDVAPEVGHKFCCKDFPGFIPSLKRLLDERQKIKT
 45 KMKASQDPIEKIMLDYRQRAIKILANSYYGYGYAKARWYCKECAESVTAWGREYI
 EFVWKELEEKFGFKVLYIDTDGLYATIPGGKSEEIKKALEFVDYINAKLPGLLELEY
 EGFYKRGGFFVTKKXYALIDEEGKIITRGLIVRRDWSEIAKETQARVLEAILKHGNVE
 EAVRIVKEVTQKLSKYEIPPEKLAIYEQITRPLHEYKAIGPHVAVAKRLAAKGVKIKP
 GMVIGYIVLRGDGPISNRILAEEYDPRKHKYDAEYYIENQVLPVLRILEGFGYRKE
 50 DLRWQKTKQTGLTSWLNKKS

SEQ ID NO:8 (Control polymerase with Sso7D DNA binding domain)

MILDADYITEEGKPVIRLFFKKEGFEFKIEHDRTRFRPYIYALLKDDSKIEEVKKITAERH
 GKIVRIVDAEKVEKKFLGRPITVWRLYFEHPQDVPTIREKIREHSAVVDIFEYDIPFAK
 RYLIDKGLIPMEGDEELKLLAFDIETLYHEGEEFGKGPIMISYADEEEAKVITWKKID
 LPYVEVVSSEREMIKRFLKIIREKDPDIIITYNGDSFDLPYLAKRAEKLGIKLTIGRDGS
 5 EPKMQRIGDMTAVEVKGRIHFDFLYHVIRRTINLPTYTLEAVYEAIFGKPKKVKYADEI
 AKAWETGEGLERVAKYSMEDAKATYELGKEFFPMEAQLSRLVGGQPLWDVSRSTG
 NLVEWFLLRKA YERNELAPNKPDEREYERRLRRESYAGGFVKEPEKGLWENIVSLDFR
 ALYPSIIITHNVSPDTLNREGCRNYDVAVEVGHKFKCKDFPGFIPSLKRLLEDERQKIKT
 10 KMKASQDPIEKIMLDYRQRAIKILANSYYGYGYAKARWYCKECAESVTAWGREYI
 EFVWKELEEKFGFKVLYIDTDGLYATIPGGKSEEIKKALEFVDYINAKLPGLLELEY
 EGFYKRGFFVTKKKYALIDEEGKIITRGLEIVRRDWSEIAKETQARVLEAILKHGNVE
 EAVRIVKEVTQKLSKYEIPPEKLAIYEQITRPLHEYKAIGPHVAVAKRLAAKGVKIKP
 GMVIGYIVLRGDGPISNRAILAEYDPRKHKYDAEYYIENQVLPVAVLRILEGFGYRKE
 15 DLRWQKTKQTGLTSLWLNKKSgtggggatvkfkykgeekvdiskikkvwrvgkmisftyde
 gggktgrgav
 sekdapkellqmleqqk

SEQ ID NO:9
 YGYGYAKARWYCKECAESVTAWGR

SEQ ID NO:10
 Sso7d/Ssh7A/SsoP2
 MATVKFKYKGEEKEVDISKIKKVWRVVGKMISFTYDEGGGKTGRGAVSEKDAPKELL
 20 QMLEKQKK

SEQ ID NO:11
 Ssh7b
 >gi |3138797|dbj|BAA28275.1| [Sulfolobus shibatae]
 MVTVKFKYKGEEKEVDTSKIKKVWRVVGKMISFTYDEGGGKTGRGAVSEKDAPKEL
 25 LQMLEKQKK

SEQ ID NO:12
 Sac7d
 >gi |152933|gb|AAA80315.1| DNA-binding protein [Sulfolobus sp.]
 35 MVKVKFKYKGEEKEVDTSKIKKVWRVVGKMVSFTYDDNGKTGRGAVSEKDAPKEL
 LDMLARAEREKK

SEQ ID NO:13
 Sac7e
 >gi |70606201|ref|YP_255071.1| DNA-binding protein 7e [Sulfolobus acidocaldarius DSM
 639]
 MAKVRFKYKGEEKEVDTSKIKKVWRVVGKMVSFTYDDNGKTGRGAVSEKDAPKEL
 40 MDMLARAEEKK

SEQ ID NO:14
 Sto7e
 >gi |15920860|ref|NP_376529.1| DNA-binding protein 7e [Sulfolobus tokodaii str. 7]
 MVTVKFKYKGEEKEVDISKIKKVWRVVGKMISFTYDDNGKTGRGAVSEKDAPKELLQ
 45 MLEKSGKK

50

WHAT IS CLAIMED IS:

- 1 1. A polypeptide comprising a DNA polymerase having an amino acid
2 sequence at least 95% identical to SEQ ID NO:2.
- 1 2. The polypeptide of claim 1, wherein the DNA polymerase comprises
2 SEQ ID NO:2.
- 1 3. The polypeptide of claim 1, wherein the DNA polymerase is linked to
2 a non-specific DNA binding domain.
- 1 4. The polypeptide of claim 3, wherein the DNA polymerase linked to the
2 non-specific DNA binding domain has higher efficiency for amplification of an amplicon of
3 at least 7.5 kb as compared to a DNA polymerase consisting of SEQ ID NO:8.
- 1 5. The polypeptide of claim 3, wherein the non-specific DNA binding
2 domain comprises an amino acid sequence at least 80% identical to SEQ ID NO:4.
- 1 6. The polypeptide of claim 3, wherein the non-specific DNA binding
2 domain comprises SEQ ID NO:4.
- 1 7. The polypeptide of claim 1, wherein the polypeptide is at least 95%
2 identical to SEQ ID NO:5.
- 1 8. The polypeptide of claim 1, wherein the polypeptide comprises SEQ
2 ID NO:5.
- 1 9. An isolated nucleic acid comprising a polynucleotide encoding a
2 polypeptide, the polypeptide comprising a DNA polymerase having an amino acid sequence
3 at least 95% identical to SEQ ID NO:2.
- 1 10. The isolated nucleic acid of claim 9, further comprising a promoter
2 operably linked to the polynucleotide.
- 1 11. The isolated nucleic acid of claim 9, wherein the DNA polymerase
2 comprises SEQ ID NO:2.
- 1 12. The isolated nucleic acid of claim 9, wherein the DNA polymerase is
2 linked to a non-specific DNA binding domain.

1 13. The isolated nucleic acid of claim 9, wherein the non-specific DNA
2 binding domain comprises an amino acid sequence at least 80% identical to SEQ ID NO:4.

1 14. The isolated nucleic acid of claim 9, wherein the non-specific DNA
2 binding domain comprises SEQ ID NO:4.

1 15. The isolated nucleic acid of claim 9, wherein the polypeptide is at least
2 95% identical to SEQ ID NO:5.

1 16. The isolated nucleic acid of claim 9, wherein the polypeptide
2 comprises SEQ ID NO:5.

1 17. The isolated nucleic acid of claim 9, wherein the DNA polymerase
2 linked to the non-specific DNA binding domain has higher efficiency for amplification of an
3 amplicon of at least 7.5 kb as compared to a DNA polymerase consisting of SEQ ID NO:8.

1 18. An isolated host cell comprising a heterologous expression cassette
2 comprising a promoter operably linked to a polynucleotide encoding a polypeptide, the
3 polypeptide comprising a DNA polymerase having an amino acid sequence at least 95%
4 identical to SEQ ID NO:2.

1 19. The isolated host cell of claim 18, wherein the DNA polymerase, when
2 linked to a non-specific DNA binding domain, has higher efficiency for amplification of an
3 amplicon of at least 5 kb as compared to a DNA polymerase consisting of SEQ ID NO:8.

1 20. The isolated host cell of claim 18, wherein the DNA polymerase
2 comprises SEQ ID NO:2.

1 21. The isolated host cell of claim 18, wherein the DNA polymerase is
2 linked to a non-specific DNA binding domain.

1 22. The isolated host cell of claim 18, wherein the non-specific DNA
2 binding domain comprises an amino acid sequence at least 80% identical to SEQ ID NO:4.

1

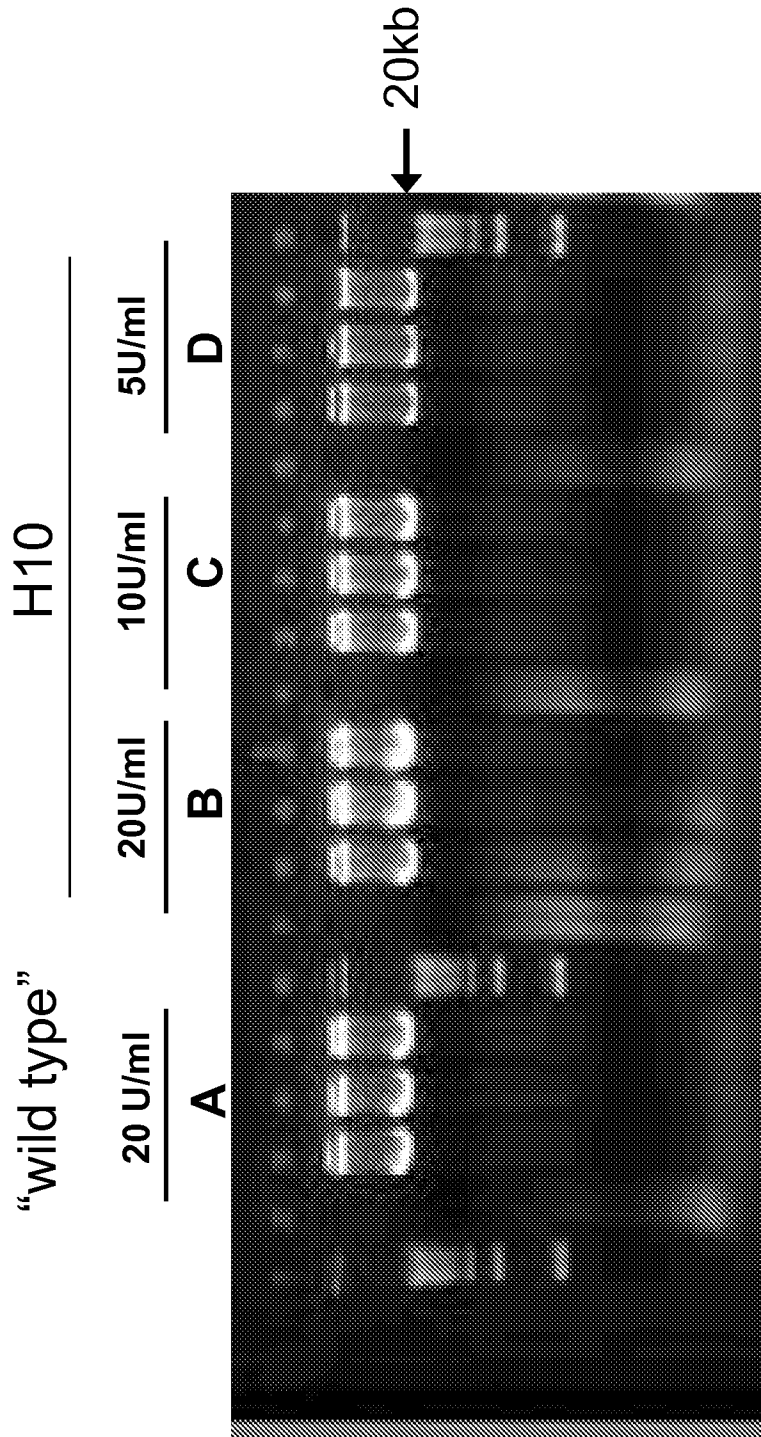


FIG. 1

SEQ NO:2 1 MILDADYITEEGKPVIRLFFKKENGFKIEHDRTRFRPYIYALLKDDSKIDEVRKITGERHG 60
MILDADYITEEGKPVIRLFFKKENG+FKIEHDRTRFRPYIYALLKDDSKI+EV+KIT ERHG
SEQ NO:7 1 MILDADYITEEGKPVIRLFFKKENGFKIEHDRTRFRPYIYALLKDDSKIEEVKKITAERHG 60

SEQ NO:2 61 KIVRIVDAEKVEKFLGRPIEVWKLYLEHPQDVPVPAIRDKVRHPAVVDIFEYDIPFAKRY 120
KIVRIVDAEKVEKFLGRPI VW+LY EHPQDVP IR+K+REH AVVDIFEYDIPFAKRY
SEQ NO:7 61 KIVRIVDAEKVEKFLGRPIVWRLYFEHPQDVPVPTIREKIREHSVVDIFEYDIPFAKRY 120

SEQ NO:2 121 LIDKGLIPMEGEEELKILAFDIEETLYHEGEEFGKGPIMISYADENEAKVITWKNIDLPI 180
LIDKGLIPMEG+EELK+LAFDIEETLYHEGEEFGKGPIMISYADE EAKVITWK IDLPI
SEQ NO:7 121 LIDKGLIPMEGDEELKLLAFDIEETLYHEGEEFGKGPIMISYADEEEAAKVITWKKIDLPI 180

SEQ NO:2 181 VEVSSEREMIKRFLRVIREKDPDVIITYNGDSFDFPYLVKRAEKLGIKLTIGRDGSEPK 240
VEVSSEREMIKRFL++IREKDPD+IITYNGDSFD PYL KRAEKLGIKLTIGRDGSEPK
SEQ NO:7 181 VEVSSEREMIKRFLKIIREKDPDIIITYNGDSFDLPLAKRAEKLGIKLTIGRDGSEPK 240

SEQ NO:2 241 MQRIGDMTAVEIKGRIHFDLYHVIRRTINLPTYTLEAVVEAIFGKPKKVVYADEIAEAW 300
MQR+GDMTAVE+KGRIFHDLYHVIRRTINLPTYTLEAVVEAIFGKPKKVVYADEIA+AW
SEQ NO:7 241 MQRIGDMTAVEIKGRIHFDLYHVIRRTINLPTYTLEAVVEAIFGKPKKVVYADEIAKAW 300

SEQ NO:2 301 SGEGLERVAKYSMEDAKVTYELGKEFLPMEIQLSRLVQGQPLWDVSRSSSTGNLVEWFLLRK 360
+GEGLERVAKYSMEDAK TYELGKEF PME QLSRLVQGQPLWDVSRSSSTGNLVEWFLLRK
SEQ NO:7 301 TGEGLERVAKYSMEDAKATYELGKEFFPMEAQLSRLVQGQPLWDVSRSSSTGNLVEWFLLRK 360

SEQ NO:2 361 AYERNEVAPNKPSEEEYERRLLRESYAGGYVKEPEKGLWENIVSLDFRSLYPSIIITHNV 420
AYERNE+APNKP E EYERRLLRESYAGG+VKEPEKGLWENIVSLDFR+LYPSIIITHNV
SEQ NO:7 361 AYERNEVAPNKPDEREYERRLLRESYAGGFVKEPEKGLWENIVSLDFRALYPSIIITHNV 420

FIG. 2

SEQ NO:2 421 P D T L N R E G C R E Y D I A P E V G H K F C K D F P G F I P S L L K H L L L E E R Q K I K T K M K E S Q D P I E K K M L 480
P D T L N R E G C R Y D + A P E V G H K F C K D F P G F I P S L L K L L + E R Q K I K T K M K S Q D P I E K M L
SEQ NO:7 421 P D T L N R E G C R N Y D V A P E V G H K F C K D F P G F I P S L L K R L L D E R Q K I K T K M K A S Q D P I E K I M L 480

SEQ NO:2 481 D Y R Q R A I K I L A N S F Y G Y G A K A R W Y C K E C A E S V T A W G R K Y I E F V W K E L E E K F G F K V L Y I 540
D Y R Q R A I K I L A N S + Y G Y G Y A K A R W Y C K E C A E S V T A W G R + Y I E F V W K E L E E K F G F K V L Y I
SEQ NO:7 481 D Y R Q R A I K I L A N S Y Y G Y G A K A R W Y C K E C A E S V T A W G R E Y I E F V W K E L E E K F G F K V L Y I 540

SEQ NO:2 541 D T D G L Y A T I P G G E P E E I K K K A L E F V K Y I N S K L P G L L E L E Y E G F Y V R G F F V T K K R Y A V I D E 600
D T D G L Y A T I P G G + E E I K K K A L E F V Y I N + K L P G L L E L E Y E G F Y R G F F V T K K + Y A + I D E
SEQ NO:7 541 D T D G L Y A T I P G G K S E E I K K K A L E F V D Y I N A K L P G L L E L E Y E G F Y K R G F F V T K K K Y A L I D E 600

SEQ NO:2 601 E G K V I T R G L E I V R R D W S E I A K E T Q A R V L E T I L K H G N V E E A V K I V K E V T Q K L A N Y E I P P E K 660
E G K + I T R G L E I V R R D W S E I A K E T Q A R V L E I L K H G N V E E A V + I V K E V T Q K L + Y E I P P E K
SEQ NO:7 601 E G K I I T R G L E I V R R D W S E I A K E T Q A R V L E A I L K H G N V E E A V R I V K E V T Q K L S K Y E I P P E K 660

SEQ NO:2 661 L A I Y E Q I T R P L H E Y K A I G P H V A V A K K L A A R G V K I K P G M V I G Y I V L R G D G P I S K R A I L A E E 720
L A I Y E Q I T R P L H E Y K A I G P H V A V A K + L A A + G V K I K P G M V I G Y I V L R G D G P I S R A I L A E E
SEQ NO:7 661 L A I Y E Q I T R P L H E Y K A I G P H V A V A K R L A A K G V K I K P G M V I G Y I V L R G D G P I S N R A I L A E E 720

SEQ NO:2 721 F D P R K H K Y D A E Y Y I E N Q V L P A V L R I L E G F G Y R K E D L R W Q K T K Q V G L T S W L N I K K S 775
+ D P R K H K Y D A E Y Y I E N Q V L P A V L R I L E G F G Y R K E D L R W Q K T K Q G L T S W L N I K K S
SEQ NO:7 721 Y D P R K H K Y D A E Y Y I E N Q V L P A V L R I L E G F G Y R K E D L R W Q K T K Q T G L T S W L N I K K S 775

FIG. 2
(cont'd)