

SYSTEMS AND METHODS FOR PROTECTING NUCLEIC ACID MOLECULES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/980,972 entitled "Methods of Analyzing Cell Free Nucleic Acids and Applications Thereof" filed February 24, 2020, which is incorporated by reference herein in its entirety.

STATEMENT AS TO FEDERALLY SPONSORED RESEARCH

[0002] This invention was made with government support under contracts CA186569 and CA188298 awarded by the National Institutes of Health. The government has certain rights in the invention.

SEQUENCE LISTING

[0003] The instant application contains a Sequence Listing which has been submitted electronically in ASCII format and is hereby incorporated by reference in its entirety. Said ASCII copy, created on February 22, 2021, is named 06342PCT2_SeqList_ST25.txt and is 4,818 bytes in size.

FIELD OF THE INVENTION

[0004] The present disclosure is generally directed toward methods of analyzing nucleic acids, and more specifically directed toward protecting nucleic acids from reactive oxygen species.

BACKGROUND

[0005] Nucleic acid molecules derived from a biological sample of a subject (e.g., a human subject) may encode information about the subject's condition (e.g., genetic mutation(s), presence of a disease, progress of a treatment for such disease, etc.). For example, noninvasive blood tests that can detect somatic alterations (e.g., mutated nucleic acids) based on the analysis of cell-free nucleic acids (e.g., cell-free deoxyribonucleic acid (cfDNA) or a cell-free ribonucleic acid (cfRNA)) may be attractive candidates for cancer screening and other applications due to the relative

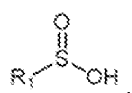
ease of obtaining biological specimens (e.g., biological fluids). There is a need for methods, systems, and compositions to promote the accurate determination of the nucleic acid sequence of nucleic acids as found in biological samples.

SUMMARY

[0006] The present disclosure provides compositions and methods for reducing or preventing alteration (e.g., via a mutation, such as a transversion) of one or more nucleic acid molecules in or derived from a biological sample of a subject. Compositions and methods of the present disclosure can reduce or prevent damage done to one or more nucleic acid molecules in an *in vitro* or *ex vivo* sample, e.g., damage done by reactive oxygen species. Compositions and methods of the present disclosure can reduce a degree and/or rate of error (e.g., sequencing error, background error) during analysis of such one or more nucleic acid molecules (e.g., cfDNA, cfRNA) for, e.g., disease diagnosis, disease monitoring, or determining treatments for the subject. Methods and systems of the present disclosure can enhance sensitivity, specificity, and/or reliability of analysis of such one or more nucleic acid molecules, e.g., detection of cancer-derived or disease-derived nucleic acids.

[0007] In one aspect, the present disclosure provides a composition comprising (i) a nucleic acid molecule and (ii) a heterologous antioxidant moiety comprising a sulfinic acid group.

[0008] In some embodiments, the heterologous antioxidant moiety has the structure:



wherein R₁ is C₁-C₆ alkylamine.

[0009] In some embodiments of any one of the compositions disclosed herein, the heterologous antioxidant moiety is hypotaurine.

[0010] In one aspect, the present disclosure provides a composition comprising (i) a nucleic acid molecule and (ii) a heterologous antioxidant moiety comprising an oligomeric protein capable of inducing a decomposition of reactive oxygen species.

[0011] In some embodiments, the oligomeric protein is a catalase. In some embodiments of any one of the compositions disclosed herein, the catalase has at least about 80%, 85%, 90%, 92%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to SEQ ID NO. 1.

[0012] In some embodiments of any one of the compositions disclosed herein, the heterologous antioxidant moiety reduces transversion of one or more nucleotides of the nucleic acid molecule. In some embodiments of any one of the compositions disclosed herein, the transversion comprises a purine to pyrimidine point mutation. In some embodiments of any one of the compositions disclosed herein, the transversion comprises a guanine to thymine point mutation, or vice versa. In some embodiments of any one of the compositions disclosed herein, when the composition is at about 47 °C for about 8 hours, the composition experiences reduced transversion of the one or more nucleotides of the nucleic molecule by at least about 20%, 30%, 40%, 50%, or more as compared to a corresponding control composition that lacks the heterologous antioxidant moiety.

[0013] In some embodiments of any one of the compositions disclosed herein, an amount of the heterologous antioxidant moiety in the composition is between about 0.1 millimolar and about 100 millimolar. In some embodiments of any one of the compositions disclosed herein, the amount of the heterologous antioxidant moiety in the composition is between about 0.5 millimolar and about 50 millimolar. In some embodiments of any one of the compositions disclosed herein, the amount of the heterologous antioxidant moiety in the composition is between about 1 millimolar and about 10 millimolar.

[0014] In some embodiments of any one of the compositions disclosed herein, an amount of a population of nucleic acid molecules comprising the nucleic acid molecule in the composition is between about 10 nanomolar and about 10 micromolar. In some embodiments of any one of the compositions disclosed herein, the amount of the population of nucleic acid molecules in the composition is between about 100 nanomolar and about 1 micromolar. In some embodiments of any one of the compositions disclosed herein, the amount of the population of nucleic acid molecules in the composition is between about 100 nanomolar and about 300 nanomolar.

[0015] In some embodiments of any one of the compositions disclosed herein, the composition further comprises a plasma sample, wherein the plasma sample comprises the nucleic acid molecule.

[0016] In some embodiments of any one of the compositions disclosed herein, the composition further comprises an isolated deoxyribonucleic acid (DNA) sample, wherein the isolated DNA sample comprises the nucleic acid molecule.

[0017] In some embodiments of any one of the compositions disclosed herein, the composition further comprises a nucleic acid analysis sample, wherein the nucleic acid analysis sample comprises the nucleic acid molecule. In some embodiments of any one of the compositions disclosed herein, the composition further comprises one or more nucleic acid probes designed to capture the nucleic acid molecule from a pool of nucleic acid molecules.

[0018] In some embodiments of any one of the compositions disclosed herein, the composition further comprises (1) a nucleic acid ligase, (2) a nucleic acid polymerase, or (3) a nucleic acid helicase for sequencing of at least a portion of the nucleic acid molecule.

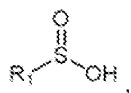
[0019] In some embodiments of any one of the compositions disclosed herein, the nucleic acid molecule is a cell-free nucleic acid molecule.

[0020] In some embodiments of any one of the compositions disclosed herein, the composition is an *ex vivo* or *in vitro* composition.

[0021] In some embodiments of any one of the compositions disclosed herein, the nucleic acid molecule is from a biological sample from a subject. In some embodiments of any one of the compositions disclosed herein, the subject is a human subject. In some embodiments of any one of the compositions disclosed herein, the subject has been or is suspected of being exposed to more oxidative stress as compared to a control.

[0022] In one aspect, the present disclosure provides a method comprising mixing (i) a nucleic acid molecule and (ii) a heterologous antioxidant moiety comprising a sulfinic acid group.

[0023] In some embodiments, the heterologous antioxidant moiety has the structure:



wherein R1 is C₁-C₆ alkylamine.

[0024] In some embodiments of any one of the methods disclosed herein, the heterologous antioxidant moiety is hypotaurine.

[0025] In one aspect, the present disclosure provides a method comprising mixing (i) a nucleic acid molecule and (ii) a heterologous antioxidant moiety comprising an oligomeric protein capable of inducing a decomposition of reactive oxygen species.

[0026] In some embodiments, the oligomeric protein is a catalase. In some embodiments of any one of the methods disclosed herein, the catalase has at least about 80%, 85%, 90%, 92%, 95%, 96%, 97%, 98%, 99%, or 100% sequence identity to SEQ ID NO. 1.

[0027] In some embodiments of any one of the methods disclosed herein, the heterologous antioxidant moiety reduces transversion of one or more nucleotides of the nucleic acid molecule. In some embodiments of any one of the methods disclosed herein, the transversion comprises a purine to pyrimidine point mutation. In some embodiments of any one of the methods disclosed herein, the transversion comprises a guanine to thymine point mutation. In some embodiments of any one of the methods disclosed herein, upon subjecting a mixture comprising the nucleic acid molecule and the heterologous antioxidant moiety to at about 47 °C for about 8 hours, the mixture experiences reduced transversion of the one or more nucleotides of the nucleic molecule by at least about 20%, 30%, 40%, 50%, or more as compared to a corresponding control composition that lacks the heterologous antioxidant moiety.

[0028] In some embodiments of any one of the methods disclosed herein, upon the mixing, an amount of the heterologous antioxidant moiety is between about 0.1 millimolar and about 100 millimolar. In some embodiments of any one of the methods disclosed herein, upon the mixing, an amount of the heterologous antioxidant moiety is between about 0.5 millimolar and about 50 millimolar. In some

embodiments of any one of the methods disclosed herein, upon the mixing, an amount of the heterologous antioxidant moiety is between about 1 millimolar and about 10 millimolar.

[0029] In some embodiments of any one of the methods disclosed herein, upon the mixing, an amount of a population of nucleic acid molecules comprising the nucleic acid molecule in the mixture is between about 10 nanomolar and about 10 micromolar. In some embodiments of any one of the methods disclosed herein, the amount of the population of nucleic acid molecules in the mixture is between about 100 nanomolar and about 1 micromolar. In some embodiments of any one of the methods disclosed herein, the amount of the population of nucleic acid molecules in the mixture is between about 100 nanomolar and about 300 nanomolar.

[0030] In some embodiments of any one of the methods disclosed herein, the method comprises mixing (i) a plasma sample that comprises the nucleic acid molecule and (ii) the heterologous antioxidant moiety comprising a sulfinic acid group.

[0031] In some embodiments of any one of the methods disclosed herein, the method comprises mixing (i) an isolated deoxyribonucleic acid (DNA) sample that comprises the nucleic acid molecule and (ii) the heterologous antioxidant moiety comprising a sulfinic acid group.

[0032] In some embodiments of any one of the methods disclosed herein, the nucleic acid molecule is a cell-free nucleic acid molecule.

[0033] In some embodiments of any one of the methods disclosed herein, the mixing is performed *ex vivo* or *in vitro* composition.

[0034] In some embodiments of any one of the methods disclosed herein, the nucleic acid molecule is from a biological sample from a subject. In some embodiments of any one of the methods disclosed herein, the subject is a human subject. In some embodiments of any one of the methods disclosed herein, the subject has been or is suspected of being exposed to more oxidative stress as compared to a control.

[0035] In some embodiments of any one of the methods disclosed herein, the method further comprises storing a mixture comprising (i) the nucleic acid molecule and (ii) the heterologous antioxidant moiety for at least 1 day, 2 days, 3 days, 4 days, 5 days, 6 days, or 7 days.

[0036] In some embodiments of any one of the methods disclosed herein, the method further comprises capturing the nucleic acid molecule via one or more nucleic acid probes prior to, concurrent with, or subsequent to the mixing.

[0037] In some embodiments of any one of the methods disclosed herein, the method further comprises, subsequent to the mixing, sequencing at least a portion of the nucleic acid molecule.

[0038] In one aspect, the present disclosure provides a device for holding a nucleic acid molecule, the device comprising a heterologous antioxidant moiety coupled to a surface of the device, wherein the heterologous antioxidant moiety comprises a sulfinic acid group.

[0039] In one aspect, the present disclosure provides a method to mitigate nucleotide transversions that arise during sequencing library preparation, comprising: performing sequence library preparation with a reactive oxygen species scavenger or enzyme in the reaction mixture.

[0040] In some embodiments, the sequence capture reaction is performed with the reactive oxygen species scavenger hypotaurine in the reaction mixture.

[0041] In some embodiments of any one of the methods disclosed herein, the reactive oxygen species scavenger is glutathione, hypotaurine, or sodium sulfite; and wherein the enzyme is uracil-DNA glycosylase (UDG), Formamidopyrimidine [fapy]-DNA glycosylase (FPG), or catalase enzyme.

BRIEF DESCRIPTION OF THE DRAWINGS

[0042] The description and claims will be more fully understood with reference to the following figures and data graphs, which are presented as exemplary embodiments of the invention and should not be construed as a complete recitation of the scope of the invention.

[0043] **Fig. 1** provides an example of a composition for protecting one or more nucleic acid molecules from oxidative damage by one or more reactive oxygen species.

[0044] **Fig. 2** provides a flow diagram of a method for protecting one or more nucleic acid molecules from oxidative damage by one or more reactive oxygen species.

[0045] Fig. 3 provides a flow diagram of a process to perform a clinical intervention on an individual based on detecting circulating tumor nucleic acid sequences in a sequencing result.

[0046] Fig. 4A provides a chart that identifies the error rates (and corresponding types of errors that arise) in samples that are treated with various chemical or enzymatic products. Fig. 4B provides a diagram illustrating the chemical mechanism by which carcinogens in cigarette smoke in vivo or reactive oxygen species (ROS) in vitro cause damage to DNA leading to the generation of 8-oxoguanine, which subsequently results in the generation of G>T transversions (Fig. 4B, top), and another diagram illustrating a proposed mechanism by which the addition of a ROS scavenger reduces oxidative-damage-derived G>T artefacts in vitro (Fig. 4B, bottom).

[0047] Fig. 5 shows that to improve sensitivity for detection of allelic levels, a few methodologies were developed and tested for maximizing the yield of unique, successfully sequenced cfDNA molecules while simultaneously minimizing their associated sequencing error profile.

[0048] Fig. 6 shows that when the error profiles of cfDNA samples from healthy adults captured with and without hypotaurine were compared, it was found that samples captured with the ROS scavenger had significantly lower background error-rates and fewer G>T errors. Shown (Fig. 6, left) is a comparison of the distribution of base substitutions in healthy control cfDNA samples (n = 12 individuals) captured with and without the ROS scavenger hypotaurine present in the hybrid capture reaction. The number of errors that are G>T transversions was compared using a paired two-sided t-test ($P < 1 \times 10^{-8}$). Also shown are aggregate selector-wide nondeduped (Fig. 6, middle) and deduped (Fig. 6, right) background error rates summarizing the results in Fig. 6, left.

[0049] Fig. 7 shows that a relative reduction of G>T errors (16% vs 57% of all errors, Wilcoxon rank-sum test, $P < 1 \times 10^{-8}$) and background error rate (about 50% reduction, Wilcoxon rank-sum test, $P < 0.0001$) was observed in healthy control cfDNA samples captured with the ROS scavenger compared to control cfDNA samples captured without hypotaurine. Shown (Fig. 7, left) is a comparison of selector-wide deduped background error rates and base substitution distributions across two cohorts of healthy controls, in which cfDNA samples were profiled with (present; bottom, n =

104) or without (absent; top, n = 69) the ROS scavenger hypotaurine in the hybrid capture reaction. Also shown (**Fig. 7, right**) are aggregate selector-wide error rates summarizing the results from **Fig. 7, left**.

DETAILED DESCRIPTION

[0050] While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

[0051] The term “antioxidant moiety” as used herein generally refers to a molecule (e.g., one or more small molecules, polypeptides, etc.) or a complex of a plurality of molecules (e.g., an oligomeric protein) that reduce or neutralize activity of reactive oxygen species (ROS) (e.g., capable of reacting with free radicals and neutralizing them). An antioxidant moiety can be a scavenger of reactive oxygen species (ROS), e.g., a small molecule capable of reacting with the ROS to transform to a different molecule. An antioxidant moiety can be a protein (e.g., an enzyme) that can decompose the ROS (or catalyze such decomposition). In some cases, a presence of an antioxidant moiety as disclosed herein can reduce or inhibit cellular damage done by the ROS. In some cases, a presence of an antioxidant moiety as disclosed herein can reduce or inhibit ROS-mediated damage to a target molecule, such as a nucleic acid molecule (e.g., a cell free nucleic acid molecule).

[0052] The term “oligomeric protein” or “oligomeric polypeptide” as used interchangeably herein generally refers to a polypeptide complex comprising two or more polypeptide molecules (or subunits), wherein the subunits complex with each other (e.g., via non-covalent interaction) to form the polypeptide complex. Such polypeptide complex can exhibit a specific activity (e.g., an enzymatic activity) to a greater degree than that of a single subunit. In some cases, the plurality of subunits can be the same (e.g., a homo-oligomeric protein). In some cases, the plurality of subunits can be different (e.g., a hetero-oligomeric protein). In some cases, an oligomeric protein can be a dimer, trimer, tetramer, etc.

[0053] The term “nucleic acid,” “polynucleotide,” or “oligonucleotide” as used interchangeably herein generally refers to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof, either in single-, double-, or multi- stranded form. A nucleic acid molecule can be exogenous to a cell. A nucleic acid molecule can exist in a cell-free environment. A nucleic acid molecule can be a gene, fragment thereof, or derivative thereof (e.g., an amplified copy). A nucleic acid molecule can be DNA. Non-limiting examples of nucleic acid molecules include coding or non-coding regions of a gene or gene fragment, loci (locus) defined from linkage analysis, exons, introns, messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), short interfering RNA (siRNA), short-hairpin RNA (shRNA), micro-RNA (miRNA), ribozymes, complimentary DNA (cDNA), recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, cell-free polynucleotides (e.g., cell-free DNA (cfDNA) (e.g., tumor cfDNA, fetal cfDNA, prenatal cfDNA, etc.), cell-free RNA (cfRNA)), nucleic acid probes, and primers.

Compositions

[0054] In an aspect, the present disclosure provides a composition comprising (i) a nucleic acid molecule and (ii) a heterologous antioxidant moiety. The heterologous antioxidant moiety can be configured to protect the nucleic acid molecule from oxidative damage by one or more reactive oxygen species (ROS). The nucleic acid molecule and the heterologous antioxidant moiety may not be derived from the same biological sample of a subject (e.g., a patient). The nucleic acid molecule and the heterologous antioxidant moiety may not be from the same source. For example, the nucleic acid molecule and the heterologous antioxidant moiety may not be naturally found together.

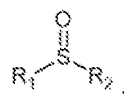
[0055] Fig. 1 shows an example composition 100 for protecting one or more nucleic acid molecules (e.g., cfDNA or cfRNA) from oxidative damage by one or more ROS. The composition 100 can comprise one or more nucleic acid molecules 110. The composition can further comprise one or more heterologous antioxidant moieties 120 configured to protect the one or more nucleic acid molecules 110 from oxidative damage.

Heterologous antioxidant moiety compounds

[0056] In some embodiments, the heterologous antioxidant moiety can comprise one or more antioxidant compounds. Such antioxidant compounds can be non-proteinaceous. Non-limiting examples of antioxidant compounds can include beta-carotene, vitamin C, vitamin E, selenium, ubiquinone, lutein, tocotrienol, isoflavone, S-adenosylmethionine, glutathione, taurine, N-acetylcysteine, lipoic acid, L-carnitine, astaxanthin, hesperidin, lutein, lycopene, polyphenol, zeaxanthin, and sodium sulfite.

[0057] In some cases, the heterologous antioxidant moiety can comprise a sulfinyl group. The heterologous antioxidant moiety can be a sulfinyl group-containing compound. In some examples, the heterologous antioxidant moiety can be a compound comprising a sulfinyl group. The compound can comprise at least or up to 1 sulfinyl group, at least or up to 2 sulfinyl groups, at least or up to 3 sulfinyl groups, at least or up to 4 sulfinyl groups, at least or up to 5 sulfinyl groups, at least or up to 6 sulfinyl groups, at least or up to 7 sulfinyl groups, at least or up to 8 sulfinyl groups, at least or up to 9 sulfinyl groups, or at least or up to 10 sulfinyl groups.

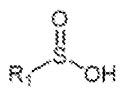
[0058] In some cases, the heterologous antioxidant moiety can have the structure:



wherein R₁ is a substituent, and R₂ is a substituent.

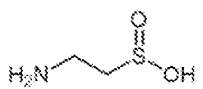
[0059] In some cases, the heterologous antioxidant moiety can comprise a sulfinic acid group. The heterologous antioxidant moiety can be a sulfinic acid group-containing compound. In some examples, the heterologous antioxidant moiety can be a compound comprising a sulfinic acid group. The compound can comprise at least or up to 1 sulfinic acid group, at least or up to 2 sulfinic acid groups, at least or up to 3 sulfinic acid groups, at least or up to 4 sulfinic acid groups, at least or up to 5 sulfinic acid groups, at least or up to 6 sulfinic acid groups, at least or up to 7 sulfinic acid groups, at least or up to 8 sulfinic acid groups, at least or up to 9 sulfinic acid groups, or at least or up to 10 sulfinic acid groups.

[0060] In some cases, the heterologous antioxidant moiety can have the structure:



wherein R₁ is a substituent. In some examples, R₁ is an alkylamine. In some examples, R₁ is a C₁-C₆ alkylamine. In some examples, R₁ is a C₁-C₄ alkylamine.

[0061] In some cases, the heterologous antioxidant moiety can be an intermediate of taurine synthesis (e.g., biosynthesis from cysteine). In some cases, the heterologous antioxidant moiety can be hypotaurine, having the structure:



[0062] Heterologous antioxidant moiety compounds as disclosed herein can include all stereoisomers, enantiomers, diastereomers, mixtures, racemates, atropisomers, and tautomers thereof.

[0063] Non-limiting examples of optional substituents include hydroxyl groups, sulfhydryl groups, halogens, amino groups, nitro groups, nitroso groups, cyano groups, azido groups, sulfoxide groups, sulfone groups, sulfonamide groups, carboxyl groups, carboxaldehyde groups, imine groups, alkyl groups, halo-alkyl groups, alkenyl groups, halo-alkenyl groups, alkynyl groups, halo-alkynyl groups, alkoxy groups, aryl groups, aryloxy groups, aralkyl groups, arylalkoxy groups, heterocyclyl groups, acyl groups, acyloxy groups, carbamate groups, amide groups, ureido groups, epoxy groups, and ester groups.

[0064] Additional examples of a heterologous antioxidant moiety can include, but are not limited to: n-octylsulfinic acid; propanesulfinic acid; 2-furansulfinic acid; cysteinesulfinic acid; F-octanesulfinic acid; 2-propanesulfinic acid; purine-6-sulfinic acid; 1-heptanesulfinic acid; 1-pentanesulfinic acid; L-cysteinesulfinic acid; 2-naphthylsulfinic acid; piperidinesulfinic acid; 3-pyridinesulfinic acid; 4-pyridinesulfinic acid; 1-naphthylsulfinic acid; ω-D-camphorsulfinic acid; chromone-3-sulfinic acid; cyclohexanesulfinic acid; but-2-ene-2-sulfinic acid; thiophene-2-sulfinic acid; 4-morpholinesulfinic acid; thiophene-2-sulfinic acid; pyrimidine-2-sulfinic acid; 2-chloroethylsulfinic acid; 3-aminopropanesulfinic acid; L-homocysteinesulfinic acid; norkhelline-6-sulfinic acid; indan-1-one-6-sulfinic acid; ethyl sulfinic acid chloride; 2-benzothiazolesulfinic acid; 4-chlorobenzenesulfinic acid; acenaphthene-3-sulfinic acid; perfluorobutanesulfinic acid; 2-amino(H)ethanesulfinic acid; 2-imidazolin-2-ylsulfinic acid; ethanesulfinic acid sodium salt; 1-methylpyrrole-2-sulfinic acid; 1-methylpyrrole-3-sulfinic acid; 3-methyl-butane-1-sulfinic acid; 2-methyl-1-propanesulfinic acid; 4-amino-toluene-2-sulfinic acid; 3-nitro-toluene-4-sulfinic acid; 2-methylpropane-2-sulfinic acid; 2-ethylhex-1-ene-1-sulfinic acid; 2-butylnon-1-ene-1-sulfinic acid; sodium, 7H-purine-6-sulfinic acid; 1-octanesulfonic-2-sulfinic acid; 1H-

benzimidazole-2-sulfinic acid; chloro sulfinic acid methyl ester; P-toluene sulfinic acid zinc salt; toluene-4-sulfinic acid-anhydride; 1-methylimidazole-2-sulfinic acid; 2-naphthylsulfinic acid sodium salt; butane-1-sulfinic acid ethyl ester; 2'-hydroxybiphenyl-2-sulfinic acid; toluene-4-sulfinic acid butyl ester; 2-pyridinesulfinic acid sodium salt; 5-methylselenophene-2-sulfinic acid; 8-nitro-naphthalene-1-sulfinic acid; 6-methylnaphthalene-2-sulfinic acid; furan-2-sulfinic acid, lithium-salt; propane-2-sulfinic acid methyl ester; toluene-4-sulfinic acid methyl ester; propane-1-sulfinic acid methyl ester; toluene-4-sulfinic acid phenyl ester; toluene-4-sulfinic acid benzyl ester; 4-chlorobenzene sulfinic acid sodium; 2-chloro-5-nitrobenzenesulfinic acid; 5-chloro-naphthalene-1-sulfinic acid; 2-propene-1-sulfinic acid, ethyl ester; 3-oxo-3-phenylpropane-1-sulfinic acid; 4,6-diaminopyrimidine-2-sulfinic acid; 6-acetylaminotoluene-3-sulfinic acid; 6-methyl-4-oxochromene-3-sulfinic acid; 2,1,3-benzothiadiazole-4-sulfinic acid; toluene-4-sulfinic acid cyclohexylamide; toluene-4-sulfinic acid, ammonium salt; benzofuran-2-sulfinic acid lithium salt; benzo-2,1,3-thiadiazole-4-sulfinic acid; toluene-4-sulfinic acid benzhydryl ester; naphthalene-2-sulfinic acid methyl ester; naphthalene-1-sulfinic acid methyl ester; 3-chlorobenzenesulfinic acid sodium salt; 3,5-dimethyl-1,2-oxazole-4-sulfinic acid; 2-chloro-5-nitro-toluene-4-sulfinic acid; 1H-purine-6-sulfinic acid, monosodium salt; 2-acetamidoanisole-4-sulfinic acid, hydrate; 5-dimethylaminonaphthalene-1-sulfinic acid; benzothiazole-2-sulfinic acid dimethylamide; 2-methyl-8-nitro-naphthalene-1-sulfinic acid; 2-methyl-5-nitro-naphthalene-1-sulfinic acid; 3-oxo-3-thiophen-2-ylpropane-1-sulfinic acid; 4-acetamido-2,6-dimethylbenzenesulfinic acid; 3-(tert-butoxy)-3-oxopropane-1-sulfinic acid; 2-methyl-propane-1-sulfinic acid methyl ester; 2-methyl-propane-2-sulfinic acid methyl ester; p-toluenesulfinic acid; (R)-(+)-2-methyl-propane-2-sulfinic acid amide; 4-bromo-2,1,3-benzothiadiazole-7-sulfinic acid; toluene-4-sulfinic acid-(1-methyl-heptyl ester); 2-butene-1-sulfinic acid, 4-phenyl-, methyl ester; 3-formyl-1H-indole-2-sulfinic acid methyl ester; 3-oxo-3-(phenethylamino)propane-1-sulfinic acid; sodium, 2-acetamido-1,3-thiazole-5-sulfinic acid; 3-(4-methoxyphenyl)-3-oxopropane-1-sulfinic acid; 2,5-dichlorothiophene-3-sulfinic acid sodium salt; 3-trifluoromethylphenyl sulfinic acid sodium salt; 2,5-dichlorothiophene-3-sulfinic acid sodium salt; sulfinic acid, 2-chloro-5-nitrobenzene-, sodium salt; 7-Octene-1-sulfinic acid, 2-oxo-2-phenylethyl ester; 9,10-dioxo-9,10-dihydro-anthracene-1-sulfinic acid; 4-amino-7H-pyrrolo[2,3-d]pyrimidine-2-sulfinic acid; 2-hydroxy-tridecane-1-sulfinic acid 4-methyl-anilide; 2-methyl-propane-2-sulfinic

acid cyclohexylideneamide; 1,1,2,2,3,3,4,4,4-nonafluoro-butane-1-sulfinic acid amide; 4-chloro-1,1,2,2,3,3,4,4,4-octafluorobutane-1-sulfinic acid; (1R)-2-methyl-propane-2-sulfinic acid 4-fluoro-benzylideneamide; 2-methyl-propane-2-sulfinic acid 1-p-tolyl-meth-(E)-ylideneamide; (R,R)-2-methylpropane-2-sulfinic acid 1-(naphthalen-1-yl)ethylamide; and toluene-sulfinic acid-(4)-(1-phenyl-ethyl ester).

[0065] Any compound herein can be purified. A compound herein can be least 1% pure, at least 2% pure, at least 3% pure, at least 4% pure, at least 5% pure, at least 6% pure, at least 7% pure, at least 8% pure, at least 9% pure, at least 10% pure, at least 11% pure, at least 12% pure, at least 13% pure, at least 14% pure, at least 15% pure, at least 16% pure, at least 17% pure, at least 18% pure, at least 19% pure, at least 20% pure, at least 21% pure, at least 22% pure, at least 23% pure, at least 24% pure, at least 25% pure, at least 26% pure, at least 27% pure, at least 28% pure, at least 29% pure, at least 30% pure, at least 31% pure, at least 32% pure, at least 33% pure, at least 34% pure, at least 35% pure, at least 36% pure, at least 37% pure, at least 38% pure, at least 39% pure, at least 40% pure, at least 41% pure, at least 42% pure, at least 43% pure, at least 44% pure, at least 45% pure, at least 46% pure, at least 47% pure, at least 48% pure, at least 49% pure, at least 50% pure, at least 51% pure, at least 52% pure, at least 53% pure, at least 54% pure, at least 55% pure, at least 56% pure, at least 57% pure, at least 58% pure, at least 59% pure, at least 60% pure, at least 61% pure, at least 62% pure, at least 63% pure, at least 64% pure, at least 65% pure, at least 66% pure, at least 67% pure, at least 68% pure, at least 69% pure, at least 70% pure, at least 71% pure, at least 72% pure, at least 73% pure, at least 74% pure, at least 75% pure, at least 76% pure, at least 77% pure, at least 78% pure, at least 79% pure, at least 80% pure, at least 81% pure, at least 82% pure, at least 83% pure, at least 84% pure, at least 85% pure, at least 86% pure, at least 87% pure, at least 88% pure, at least 89% pure, at least 90% pure, at least 91% pure, at least 92% pure, at least 93% pure, at least 94% pure, at least 95% pure, at least 96% pure, at least 97% pure, at least 98% pure, at least 99% pure, at least 99.1% pure, at least 99.2% pure, at least 99.3% pure, at least 99.4% pure, at least 99.5% pure, at least 99.6% pure, at least 99.7% pure, at least 99.8% pure, or at least 99.9% pure.

Acceptable salts

[0066] Any therapeutic compound described herein can be provided in the form of a salt (e.g., a pharmaceutically-acceptable salt). The acceptable salts include, for

example, acid-addition salts and base-addition salts. The acid that is added to the compound to form an acid-addition salt can be an organic acid or an inorganic acid. A base that is added to the compound to form a base-addition salt can be an organic base or an inorganic base. In some cases, an acceptable salt is a metal salt. In some cases, an acceptable salt is an ammonium salt.

[0067] Metal salts can arise from the addition of an inorganic base to a compound of the invention. The inorganic base consists of a metal cation paired with a basic counterion, such as, for example, hydroxide, carbonate, bicarbonate, or phosphate. The metal can be an alkali metal, alkaline earth metal, transition metal, or main group metal. In some cases, the metal is lithium, sodium, potassium, cesium, cerium, magnesium, manganese, iron, calcium, strontium, cobalt, titanium, aluminum, copper, cadmium, or zinc.

[0068] In some cases, a metal salt is a lithium salt, a sodium salt, a potassium salt, a cesium salt, a cerium salt, a magnesium salt, a manganese salt, an iron salt, a calcium salt, a strontium salt, a cobalt salt, a titanium salt, an aluminum salt, a copper salt, a cadmium salt, or a zinc salt.

[0069] Ammonium salts can arise from the addition of ammonia or an organic amine to a compound of the invention. In some cases, the organic amine is triethyl amine, diisopropyl amine, ethanol amine, diethanol amine, triethanol amine, morpholine, N-methylmorpholine, piperidine, N-methylpiperidine, N-ethylpiperidine, dibenzylamine, piperazine, pyridine, pyrrazole, pipyrrazole, imidazole, pyrazine, or pipyrazine.

[0070] In some cases, an ammonium salt is a triethyl amine salt, a diisopropyl amine salt, an ethanol amine salt, a diethanol amine salt, a triethanol amine salt, a morpholine salt, an N-methylmorpholine salt, a piperidine salt, an N-methylpiperidine salt, an N-ethylpiperidine salt, a dibenzylamine salt, a piperazine salt, a pyridine salt, a pyrrazole salt, a pipyrrazole salt, an imidazole salt, a pyrazine salt, or a pipyrazine salt.

[0071] Acid addition salts can arise from the addition of an acid to a compound of the invention. In some cases, the acid is organic. In some cases, the acid is inorganic. In some cases, the acid is hydrochloric acid, hydrobromic acid, hydroiodic acid, nitric acid, nitrous acid, sulfuric acid, sulfurous acid, a phosphoric acid, isonicotinic acid, lactic acid, salicylic acid, tartaric acid, ascorbic acid, gentisinic acid, gluconic acid, glucaronic acid, saccharic acid, formic acid, benzoic acid, glutamic acid, pantothenic

acid, acetic acid, propionic acid, butyric acid, fumaric acid, succinic acid, methanesulfonic acid, ethanesulfonic acid, benzenesulfonic acid, p-toluenesulfonic acid, citric acid, oxalic acid, or maleic acid.

[0072] In some cases, the salt is a hydrochloride salt, a hydrobromide salt, a hydroiodide salt, a nitrate salt, a nitrite salt, a sulfate salt, a sulfite salt, a phosphate salt, isonicotinate salt, a lactate salt, a salicylate salt, a tartrate salt, an ascorbate salt, a gentisinate salt, a gluconate salt, a glucaronate salt, a saccharate salt, a formate salt, a benzoate salt, a glutamate salt, a pantothenate salt, an acetate salt, a propionate salt, a butyrate salt, a fumarate salt, a succinate salt, a methanesulfonate (mesylate) salt, an ethanesulfonate salt, a benzenesulfonate salt, a p-toluenesulfonate salt, a citrate salt, an oxalate salt, or a maleate salt.

Heterologous antioxidant moiety proteins

[0073] In some cases, the heterologous antioxidant moiety can comprise a polypeptide (e.g., a protein, such as an enzyme) capable of effecting (e.g., inducing) a decomposition of reactive oxygen species. For example, an enzymatic heterologous antioxidant moiety can catalyze a decomposition of hydrogen peroxide to water and oxygen. Non-limiting examples of such enzymatic heterologous antioxidant moiety can include uracil-DNA glycosylase (UDG), Formamidopyrimidine [fapy]-DNA glycosylase (FPG), catalase, superoxide dismutase, glutathione peroxidase, glutathione reductase, and glutathione S-transferase.

[0074] In some cases, the heterologous antioxidant moiety can comprise one or more iron-containing heme groups that allow the polypeptide to react with the hydrogen peroxide. The heterologous antioxidant moiety can comprise at least or up to about 1 iron-containing heme group, at least or up to about 2 iron-containing heme groups, at least or up to about 3 iron-containing heme groups, at least or up to about 4 iron-containing heme group, at least or up to about 5 iron-containing heme groups, at least or up to about 6 iron-containing heme groups, at least or up to about 7 iron-containing heme groups, at least or up to about 8 iron-containing heme groups, at least or up to about 9 iron-containing heme group, or at least or up to about 10 iron-containing heme groups.

[0075] In some cases, the polypeptide-based heterologous antioxidant moiety can comprise an oligomeric protein capable of effecting decomposition of reactive oxygen species. The oligomeric protein can comprise two or more polypeptide molecules as

subunits, in which the subunits collectively form a complex that is capable of effecting a decomposition of reactive oxygen species. The oligomeric protein can comprise at least or up to about 2, at least or up to about 3, at least or up to about 4, at least or up to about 5, or at least or up to about 6 polypeptide molecules as subunits.

[0076] In some cases, the oligomeric protein-based heterologous antioxidant moiety can be a tetramer comprising 4 subunits. Each subunit of the 4 subunits of the tetramer can be the same. Alternatively, a first subunit of the 4 subunits of the tetramer can be different than a second subunit of the 4 subunits of the tetramer. In some examples, the oligomeric protein can be a catalase or a functional variant thereof. A subunit polypeptide molecule of a catalase can have at least about 50%, at least about 55%, at least about 60%, at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90% sequence identity, at least about 95% sequence identity, at least about 99%, or about 100% sequence identity to human catalase (SEQ ID NO. 1).

SEQ ID NO. 1:

```
MADSRDPASD QMQHWKEQRA AQKADVLTG AGNPVGDKLN VITVGPRGPL
LVQDVVFTDE MAHFDREIRIP ERVHAKGAG AFGYFEVTHD ITKYSKAKVF
EHIGKKTPIA VRFSTVAGES GSADTVRDPR GFAVKFYTED GNWDLVGNNT
PIFFIRDPIIL FPSFIHSQKR NPQTHLKDPD MVWDFWSLRP ESLHQVSFLF
SDRGIPDGHR HMNGYGSHTF KLVNANGEAV YCKFHYKTDQ GIKNLSVEDA
ARLSQEDPDY GIRDLFNAIA TGKYPSTWTFY IQVMTFNQAE TFPFNPFDLT
KVWPHKDYPL IPVGKLVNLR NPVNYFAEVE QIAFDPSNMP PGIEASPDKM
LQGRLFAYPD THRHRLGPNY LHIPVNCOPYR ARVANYQRDG PMCMQDNQGG
APNYYPNSFG APEQQPSALE HSIQYSGEVR RFNTANDDNV TQVRAFYVNV
LNEEQRKRLC ENIAGHLKDA QIFIQKKAVK NFTEVHPDYG SHIQALLDKY
NAEKPKNAIH TQVQSGSHLA AREKANL
```

Additional aspects of the composition

[0077] In some embodiments of any one of the compositions disclosed herein, the heterologous antioxidant moiety can reduce alternation (e.g., mutation, such as transversion or transition) of one or more nucleotides of the nucleic acid molecule. In some cases, the heterologous antioxidant moiety can reduce transversion of one or more nucleotides of the nucleic acid molecule. A transversion can be an edit (e.g., a

point mutation) in a nucleic acid molecule in which a single purine (A or G) is changed for a pyrimidine (T or C), or vice versa. Without wishing to be bound by theory, such transversion can be (i) guanine to thymine, or vice versa, (ii) guanine to cytosine, or vice versa, (iii) adenine to cytosine, or vice versa, or (iv) adenine to thymine, or vice versa. For example, the heterologous antioxidant moiety can reduce transversion of one or more nucleotides from a guanine to thymine, or vice versa.

[0078] In some embodiments of any one of the compositions disclosed herein, when the composition is at a temperature higher than or equal to room temperature (about 25 °C) for a period of time (e.g., a predetermined period of time), the composition can experience reduced transversion of one or more nucleotides of a nucleic molecule by at least or up to about 5%, at least or up to about 10%, at least or up to about 15%, at least or up to about 20%, at least or up to about 25%, at least or up to about 30%, at least or up to about 35%, at least or up to about 40%, at least or up to about 45%, at least or up to about 50%, at least or up to about 55%, at least or up to about 60%, at least or up to about 65%, at least or up to about 70%, at least or up to about 75%, at least or up to about 80%, at least or up to about 85%, at least or up to about 90%, at least or up to about 95%, at least or up to about 99%, or about 100%, as compared to a corresponding control composition that lacks the heterologous antioxidant moiety.

[0079] In some cases, the reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature of at least or up to about 30 °C, at least or up to about 35 °C, at least or up to about 40 °C, at least or up to about 45 °C, at least or up to about 50 °C, at least or up to about 55 °C, at least or up to about 60 °C, at least or up to about 65 °C, at least or up to about 70 °C, at least or up to about 75 °C, at least or up to about 80 °C, at least or up to about 85 °C, or at least or up to about 90 °C. For example, the reduced transversion can be observed when the composition and the corresponding control composition are each at about 47 °C.

[0080] In some cases, the reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for at least or up to about 5 minutes, at least or up to about 10 minutes, at least or up to about 15 minutes, at least or up to about 20 minutes, at least or up to about 25 minutes, at least or up to about 30 minutes, at least or up to about 40 minutes, at least or up to about 50 minutes, at least or up to about 60 minutes, at

least or up to about 2 hours, at least or up to about 3 hours, at least or up to about 6 hours, at least or up to about 12 hours, at least or up to about 18 hours, at least or up to about 24 hours, at least or up to about 36 hours, at least or up to about 48 hours, at least or up to about 60 hours, or at least or up to about 72 hours.

[0081] In some cases, the reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for about 1 hour to about 96 hours. The reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for at least about 1 hour. The reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for at most about 96 hours. The reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for about 1 hour to about 2 hours, about 1 hour to about 4 hours, about 1 hour to about 8 hours, about 1 hour to about 12 hours, about 1 hour to about 16 hours, about 1 hour to about 20 hours, about 1 hour to about 24 hours, about 1 hour to about 36 hours, about 1 hour to about 48 hours, about 1 hour to about 72 hours, about 1 hour to about 96 hours, about 2 hours to about 4 hours, about 2 hours to about 8 hours, about 2 hours to about 12 hours, about 2 hours to about 16 hours, about 2 hours to about 20 hours, about 2 hours to about 24 hours, about 2 hours to about 36 hours, about 2 hours to about 48 hours, about 2 hours to about 72 hours, about 2 hours to about 96 hours, about 4 hours to about 8 hours, about 4 hours to about 12 hours, about 4 hours to about 16 hours, about 4 hours to about 20 hours, about 4 hours to about 24 hours, about 4 hours to about 36 hours, about 4 hours to about 48 hours, about 4 hours to about 72 hours, about 4 hours to about 96 hours, about 8 hours to about 12 hours, about 8 hours to about 16 hours, about 8 hours to about 20 hours, about 8 hours to about 24 hours, about 8 hours to about 36 hours, about 8 hours to about 48 hours, about 8 hours to about 72 hours, about 8 hours to about 96 hours, about 12 hours to about 16 hours, about 12 hours to about 20 hours, about 12 hours to about 24 hours, about 12 hours to about 36 hours, about 12 hours to about 48 hours, about 12 hours to about 72 hours, about 12 hours to about 96 hours, about 16 hours to about 20 hours, about 16 hours to about 24 hours, about 16 hours to about 36 hours, about 16 hours to about 48 hours, about 16 hours to about 72 hours, about 16 hours to about 96 hours, about 20 hours to about 24 hours, about 20 hours to about 36 hours, about 20 hours

to about 48 hours, about 20 hours to about 72 hours, about 20 hours to about 96 hours, about 24 hours to about 36 hours, about 24 hours to about 48 hours, about 24 hours to about 72 hours, about 24 hours to about 96 hours, about 36 hours to about 48 hours, about 36 hours to about 72 hours, about 36 hours to about 96 hours, about 48 hours to about 72 hours, about 48 hours to about 96 hours, or about 72 hours to about 96 hours. The reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for about 1 hour, about 2 hours, about 4 hours, about 8 hours, about 12 hours, about 16 hours, about 20 hours, about 24 hours, about 36 hours, about 48 hours, about 72 hours, or about 96 hours. For example, the reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for about 16 hours. In another example, the reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for about 48 hours. In a different example, the reduced transversion can be observed when the composition and the corresponding control composition are each at a temperature (or a temperature range) for about 72 hours.

[0082] In some embodiments of any one of the compositions disclosed herein, an amount of the heterologous antioxidant moiety in the composition can be between about 0.01 millimolar (mM) and about 500 mM. The amount of the heterologous antioxidant moiety in the composition can be at least about 0.001 mM. The amount of the heterologous antioxidant moiety in the composition can be at most about 500 mM. The amount of the heterologous antioxidant moiety in the composition can be about 0.001 mM to about 0.005 mM, about 0.001 mM to about 0.01 mM, about 0.001 mM to about 0.05 mM, about 0.001 mM to about 0.1 mM, about 0.001 mM to about 0.5 mM, about 0.001 mM to about 1 mM, about 0.001 mM to about 5 mM, about 0.001 mM to about 10 mM, about 0.001 mM to about 50 mM, about 0.001 mM to about 100 mM, about 0.001 mM to about 500 mM, about 0.005 mM to about 0.01 mM, about 0.005 mM to about 0.05 mM, about 0.005 mM to about 0.1 mM, about 0.005 mM to about 0.5 mM, about 0.005 mM to about 1 mM, about 0.005 mM to about 5 mM, about 0.005 mM to about 10 mM, about 0.005 mM to about 50 mM, about 0.005 mM to about 100 mM, about 0.005 mM to about 500 mM, about 0.01 mM to about 0.05 mM, about 0.01 mM to about 0.1 mM, about 0.01 mM to about 0.5 mM, about 0.01 mM to about 1 mM, about 0.01 mM to about 5 mM, about 0.01 mM to about 10 mM, about 0.01 mM to

about 50 mM, about 0.01 mM to about 100 mM, about 0.01 mM to about 500 mM, about 0.05 mM to about 0.1 mM, about 0.05 mM to about 0.5 mM, about 0.05 mM to about 1 mM, about 0.05 mM to about 5 mM, about 0.05 mM to about 10 mM, about 0.05 mM to about 50 mM, about 0.05 mM to about 100 mM, about 0.05 mM to about 500 mM, about 0.1 mM to about 0.5 mM, about 0.1 mM to about 1 mM, about 0.1 mM to about 5 mM, about 0.1 mM to about 10 mM, about 0.1 mM to about 50 mM, about 0.1 mM to about 100 mM, about 0.1 mM to about 500 mM, about 0.5 mM to about 1 mM, about 0.5 mM to about 5 mM, about 0.5 mM to about 10 mM, about 0.5 mM to about 50 mM, about 0.5 mM to about 100 mM, about 0.5 mM to about 500 mM, about 1 mM to about 5 mM, about 1 mM to about 10 mM, about 1 mM to about 50 mM, about 1 mM to about 100 mM, about 1 mM to about 500 mM, about 3 mM to about 7 mM, about 5 mM to about 10 mM, about 5 mM to about 50 mM, about 5 mM to about 100 mM, about 5 mM to about 500 mM, about 10 mM to about 50 mM, about 10 mM to about 100 mM, about 10 mM to about 500 mM, about 50 mM to about 100 mM, about 50 mM to about 500 mM, or about 100 mM to about 500 mM. The amount of the heterologous antioxidant moiety in the composition can be about 0.001 mM, about 0.005 mM, about 0.01 mM, about 0.05 mM, about 0.1 mM, about 0.5 mM, about 1 mM, about 5 mM, about 10 mM, about 50 mM, about 100 mM, or about 500 mM.

[0083] In some embodiments of any one of the compositions disclosed herein, an amount of the heterologous antioxidant moiety in the composition can be between about 10 milligrams per liter (mg/L) and about 5,000 mg/L. The amount of the heterologous antioxidant moiety in the composition can be at least about 10 mg/L. The amount of the heterologous antioxidant moiety in the composition can be at most about 5,000 mg/L. The amount of the heterologous antioxidant moiety in the composition can be about 10 mg/L to about 50 mg/L, about 10 mg/L to about 100 mg/L, about 10 mg/L to about 200 mg/L, about 10 mg/L to about 400 mg/L, about 10 mg/L to about 500 mg/L, about 10 mg/L to about 550 mg/L, about 10 mg/L to about 600 mg/L, about 10 mg/L to about 800 mg/L, about 10 mg/L to about 1,000 mg/L, about 10 mg/L to about 2,000 mg/L, about 10 mg/L to about 5,000 mg/L, about 50 mg/L to about 100 mg/L, about 50 mg/L to about 200 mg/L, about 50 mg/L to about 400 mg/L, about 50 mg/L to about 500 mg/L, about 50 mg/L to about 550 mg/L, about 50 mg/L to about 600 mg/L, about 50 mg/L to about 800 mg/L, about 50 mg/L to about 1,000 mg/L, about 50 mg/L to about 2,000 mg/L, about 50 mg/L to about 5,000 mg/L, about 100 mg/L to about 200 mg/L, about 100 mg/L to about 400 mg/L, about 100 mg/L to about

500 mg/L, about 100 mg/L to about 550 mg/L, about 100 mg/L to about 600 mg/L, about 100 mg/L to about 800 mg/L, about 100 mg/L to about 1,000 mg/L, about 100 mg/L to about 2,000 mg/L, about 100 mg/L to about 5,000 mg/L, about 200 mg/L to about 400 mg/L, about 200 mg/L to about 500 mg/L, about 200 mg/L to about 550 mg/L, about 200 mg/L to about 600 mg/L, about 200 mg/L to about 800 mg/L, about 200 mg/L to about 1,000 mg/L, about 200 mg/L to about 2,000 mg/L, about 200 mg/L to about 5,000 mg/L, about 400 mg/L to about 500 mg/L, about 400 mg/L to about 550 mg/L, about 400 mg/L to about 600 mg/L, about 400 mg/L to about 800 mg/L, about 400 mg/L to about 1,000 mg/L, about 400 mg/L to about 2,000 mg/L, about 400 mg/L to about 5,000 mg/L, about 500 mg/L to about 550 mg/L, about 500 mg/L to about 600 mg/L, about 500 mg/L to about 800 mg/L, about 500 mg/L to about 1,000 mg/L, about 500 mg/L to about 2,000 mg/L, about 500 mg/L to about 5,000 mg/L, about 550 mg/L to about 600 mg/L, about 550 mg/L to about 800 mg/L, about 550 mg/L to about 1,000 mg/L, about 550 mg/L to about 2,000 mg/L, about 550 mg/L to about 5,000 mg/L, about 600 mg/L to about 800 mg/L, about 600 mg/L to about 1,000 mg/L, about 600 mg/L to about 2,000 mg/L, about 600 mg/L to about 5,000 mg/L, about 800 mg/L to about 1,000 mg/L, about 800 mg/L to about 2,000 mg/L, about 800 mg/L to about 5,000 mg/L, about 1,000 mg/L to about 2,000 mg/L, about 1,000 mg/L to about 5,000 mg/L, or about 2,000 mg/L to about 5,000 mg/L. The amount of the heterologous antioxidant moiety in the composition can be about 10 mg/L, about 50 mg/L, about 100 mg/L, about 200 mg/L, about 400 mg/L, about 500 mg/L, about 550 mg/L, about 600 mg/L, about 800 mg/L, about 1,000 mg/L, about 2,000 mg/L, or about 5,000 mg/L.

[0084] In some embodiments of any one of the compositions disclosed herein, the composition can comprise one or more nucleic acid molecules (NA) and the heterologous antioxidant moiety (HAM) in a molar ratio (HAM:NA) of between about 1,000,000:1 and about 1,000,000:5,000, between about 1,000,000:1 and about 1,000,000:2,000, between about 1,000,000:1 and about 1,000,000:1,000, between about 1,000,000:1 and about 1,000,000:500, between about 1,000,000:1 and about 1,000,000:400, between about 1,000,000:1 and about 1,000,000:300, between about 1,000,000:1 and about 1,000,000:200, between about 1,000,000:1 and about 1,000,000:180, between about 1,000,000:1 and about 1,000,000:160, between about 1,000,000:1 and about 1,000,000:140, between about 1,000,000:1 and about 1,000,000:120, between about 1,000,000:1 and about 1,000,000:100, between about 1,000,000:1 and about 1,000,000:80, between about 1,000,000:1 and about

1,000,000:60, between about 1,000,000:1 and about 1,000,000:50, between about 1,000,000:1 and about 1,000,000:40, between about 1,000,000:1 and about 1,000,000:30, between about 1,000,000:1 and about 1,000,000:20, between about 1,000,000:1 and about 1,000,000:10, or between about 1,000,000:1 and about 1,000,000:5. In some cases, the molar ratio (HAM:NA) can be between about 1,000,000:10 and about 1,000,000:500, between about 1,000,000:20 and about 1,000,000:400, or between about 1,000,000:25 and about 1,000,000:200. In some examples, the molar ratio (HAM:NA) can be about 1,000,000:30 and about 1,000,000:40. In some examples, the molar ratio (HAM:NA) can be about 1,000,000:100 and about 1,000,000:200.

[0085] In some embodiments of any one of the compositions disclosed herein, the composition can comprise one or more nucleic acid molecules (NA) and the heterologous antioxidant moiety (HAM) in a weight ratio (HAM:NA) of between about 100:1 and about 100:100, between about 100:1 and about 100:80, between about 100:1 and about 100:60, between about 100:1 and about 100:50, between about 100:1 and about 100:40, between about 100:1 and about 100:30, between about 100:1 and about 100:20, between about 100:1 and about 100:15, between about 100:1 and about 100:12, between about 100:1 and about 100:10, or between about 100:1 and about 100:5. In some cases, the weight ratio (HAM:NA) can be between about 100:5 and about 100:50. In some examples, the weight ratio (HAM:NA) can be about 100:9. In some examples, the weight ratio (HAM:NA) can be about 100:40.

[0086] In some embodiments of any one of the compositions disclosed herein, an amount of the one or more nucleic acid molecules in the composition can be between about 10 nanomolar (nM) and about 10,000 nM. The amount of the one or more nucleic acid molecules in the composition can be at least about 10 nM. The amount of the one or more nucleic acid molecules in the composition can be at most about 10,000 nM. The amount of the one or more nucleic acid molecules in the composition can be about 10 nM to about 100 nM, about 10 nM to about 120 nM, about 10 nM to about 150 nM, about 10 nM to about 200 nM, about 10 nM to about 400 nM, about 10 nM to about 600 nM, about 10 nM to about 800 nM, about 10 nM to about 1,000 nM, about 10 nM to about 2,000 nM, about 10 nM to about 5,000 nM, about 10 nM to about 10,000 nM, about 100 nM to about 120 nM, about 100 nM to about 150 nM, about 100 nM to about 200 nM, about 100 nM to about 400 nM, about 100 nM to about 600 nM, about 100 nM to about 800 nM, about 100 nM to about 1,000 nM, about 100 nM to

about 2,000 nM, about 100 nM to about 5,000 nM, about 100 nM to about 10,000 nM, about 120 nM to about 150 nM, about 120 nM to about 200 nM, about 120 nM to about 400 nM, about 120 nM to about 600 nM, about 120 nM to about 800 nM, about 120 nM to about 1,000 nM, about 120 nM to about 2,000 nM, about 120 nM to about 5,000 nM, about 120 nM to about 10,000 nM, about 150 nM to about 200 nM, about 150 nM to about 400 nM, about 150 nM to about 600 nM, about 150 nM to about 800 nM, about 150 nM to about 1,000 nM, about 150 nM to about 2,000 nM, about 150 nM to about 5,000 nM, about 150 nM to about 10,000 nM, about 200 nM to about 400 nM, about 200 nM to about 600 nM, about 200 nM to about 800 nM, about 200 nM to about 1,000 nM, about 200 nM to about 2,000 nM, about 200 nM to about 5,000 nM, about 200 nM to about 10,000 nM, about 400 nM to about 600 nM, about 400 nM to about 800 nM, about 400 nM to about 1,000 nM, about 400 nM to about 2,000 nM, about 400 nM to about 5,000 nM, about 400 nM to about 10,000 nM, about 600 nM to about 800 nM, about 600 nM to about 1,000 nM, about 600 nM to about 2,000 nM, about 600 nM to about 5,000 nM, about 600 nM to about 10,000 nM, about 800 nM to about 1,000 nM, about 800 nM to about 2,000 nM, about 800 nM to about 5,000 nM, about 800 nM to about 10,000 nM, about 1,000 nM to about 2,000 nM, about 1,000 nM to about 5,000 nM, about 1,000 nM to about 10,000 nM, about 2,000 nM to about 5,000 nM, about 2,000 nM to about 10,000 nM, or about 5,000 nM to about 10,000 nM. The amount of the one or more nucleic acid molecules in the composition can be about 10 nM, about 100 nM, about 120 nM, about 150 nM, about 200 nM, about 400 nM, about 600 nM, about 800 nM, about 1,000 nM, about 2,000 nM, about 5,000 nM, or about 10,000 nM.

[0087] In some embodiments of any one of the compositions disclosed herein, an amount of one or more nucleic acid molecules in the composition can be between about 1 mg/L and about 5,000 mg/L. The amount of the one or more nucleic acid molecules in the composition can be at least about 1 mg/L. The amount of the one or more nucleic acid molecules in the composition can be at most about 5,000 mg/L. The amount of the one or more nucleic acid molecules in the composition can be about 1 mg/L to about 10 mg/L, about 1 mg/L to about 20 mg/L, about 1 mg/L to about 50 mg/L, about 1 mg/L to about 100 mg/L, about 1 mg/L to about 150 mg/L, about 1 mg/L to about 200 mg/L, about 1 mg/L to about 300 mg/L, about 1 mg/L to about 400 mg/L, about 1 mg/L to about 500 mg/L, about 1 mg/L to about 1,000 mg/L, about 1 mg/L to about 5,000 mg/L, about 10 mg/L to about 20 mg/L, about 10 mg/L to about 50 mg/L,

about 10 mg/L to about 100 mg/L, about 10 mg/L to about 150 mg/L, about 10 mg/L to about 200 mg/L, about 10 mg/L to about 300 mg/L, about 10 mg/L to about 400 mg/L, about 10 mg/L to about 500 mg/L, about 10 mg/L to about 1,000 mg/L, about 10 mg/L to about 5,000 mg/L, about 20 mg/L to about 50 mg/L, about 20 mg/L to about 100 mg/L, about 20 mg/L to about 150 mg/L, about 20 mg/L to about 200 mg/L, about 20 mg/L to about 300 mg/L, about 20 mg/L to about 400 mg/L, about 20 mg/L to about 500 mg/L, about 20 mg/L to about 1,000 mg/L, about 20 mg/L to about 5,000 mg/L, about 50 mg/L to about 100 mg/L, about 50 mg/L to about 150 mg/L, about 50 mg/L to about 200 mg/L, about 50 mg/L to about 300 mg/L, about 50 mg/L to about 400 mg/L, about 50 mg/L to about 500 mg/L, about 50 mg/L to about 1,000 mg/L, about 50 mg/L to about 5,000 mg/L, about 100 mg/L to about 150 mg/L, about 100 mg/L to about 200 mg/L, about 100 mg/L to about 300 mg/L, about 100 mg/L to about 400 mg/L, about 100 mg/L to about 500 mg/L, about 100 mg/L to about 1,000 mg/L, about 100 mg/L to about 5,000 mg/L, about 150 mg/L to about 200 mg/L, about 150 mg/L to about 300 mg/L, about 150 mg/L to about 400 mg/L, about 150 mg/L to about 500 mg/L, about 150 mg/L to about 1,000 mg/L, about 150 mg/L to about 5,000 mg/L, about 200 mg/L to about 300 mg/L, about 200 mg/L to about 400 mg/L, about 200 mg/L to about 500 mg/L, about 200 mg/L to about 1,000 mg/L, about 200 mg/L to about 5,000 mg/L, about 300 mg/L to about 400 mg/L, about 300 mg/L to about 500 mg/L, about 300 mg/L to about 1,000 mg/L, about 300 mg/L to about 5,000 mg/L, about 400 mg/L to about 500 mg/L, about 400 mg/L to about 1,000 mg/L, about 400 mg/L to about 5,000 mg/L, about 500 mg/L to about 1,000 mg/L, about 500 mg/L to about 5,000 mg/L, or about 1,000 mg/L to about 5,000 mg/L. The amount of the one or more nucleic acid molecules in the composition can be about 1 mg/L, about 10 mg/L, about 20 mg/L, about 50 mg/L, about 100 mg/L, about 150 mg/L, about 200 mg/L, about 300 mg/L, about 400 mg/L, about 500 mg/L, about 1,000 mg/L, or about 5,000 mg/L.

[0088] In some embodiments of any one of the compositions disclosed herein, the composition can comprise a biological sample, such as a blood sample (e.g., a plasma sample or a serum sample), of a subject (e.g., a mammal, such as an animal or a human), and one or more nucleic acid molecules in the composition can be from the biological sample. Non-limiting examples of a biological sample can include blood, plasma, serum, urine, perilymph fluid, feces, saliva, semen, amniotic fluid, cerebrospinal fluid, bile, sweat, tears, sputum, synovial fluid, vomit, bone, heart,

thymus, artery, blood vessel, lung, muscle, stomach, intestine, liver, pancreas, spleen, kidney, gall bladder, thyroid gland, adrenal gland, mammary gland, ovary, prostate gland, testicle, skin, adipose, eye, brain, infected tissue, diseased tissue, malignant tissue, calcified tissue, and healthy tissue.

[0089] The subject (e.g., a human subject) as disclosed herein can be exposed to or can be suspected of having been exposed to more oxidative stress as compared to a corresponding control subject. Exposure to oxidative stress can be from one or more sources including, but not limited to, smoking (e.g., tobacco products, such as cigarettes or cigars), ultraviolet (UV) exposure, alcohol consumption, obesity, diets (e.g., high fat diet, high sugar diet), exposure to radiation, pollution, exposure to pesticides, certain medications (e.g., nimustine, actinomycin D, doxorubicin, mitomycin C, mitoxantrone, carmofur, gemcitabine, mercaptopurine, camptothecin, paclitaxel, vinblastine, vinorelbine), and one or more diseases (e.g., neurodegenerative diseases, such as Lou Gehrig's disease, Parkin's disease, Alzheimer's disease, Huntington's disease, multiple sclerosis; cardiovascular disease; cancer or tumor; aging). For example, current and/or former smokers can be exposed to more oxidative stress (e.g., due to more exposure to polycyclic aromatic hydrocarbons) as compared to a non-smoking or never-smoking subject. Nucleic acid molecules (e.g., cfDNA or cfRNA) derived from a biological sample of such smokers can exhibit a unique signature that can be damaged or altered (e.g., mutagenesis from a guanine-cytosine (G-C) pair to a thymine-adenine (T-A) pair) in absence of the heterologous antioxidant moiety as disclosed herein.

[0090] In some embodiments of any one of the compositions disclosed herein, the composition can comprise an isolated nucleic acid sample (e.g., isolated DNA sample, isolated RNA sample). The isolated nucleic acid sample can comprise one or more nucleic acid (e.g., DNA, RNA) molecules that are substantially free of, or have been isolated to be substantially free of, the bulk of the total genomic and/or transcribed nucleic acids of one or more cells, e.g., prior to the addition of the heterologous antioxidant moiety to the composition.

[0091] In some embodiments of any one of the compositions disclosed herein, the composition can comprise a nucleic acid analysis sample that comprises one or more nucleic acid molecules. The one or more nucleic acid molecules of the nucleic acid analysis sample can be isolated from a biological sample of a subject as disclosed herein. Alternatively, the one or more nucleic acid molecules of the nucleic acid

analysis sample can be prepared for amplification of one or more target nucleic acid sequences (e.g., via polymerase chain reaction (PCR)) in such sample. Yet in another alternative, the one or more nucleic acid molecules of the nucleic acid analysis sample can be derived from (e.g., amplified via PCR) one or more nucleic acid templates from a biological sample.

[0092] In some cases, the nucleic acid analysis sample can be prepared for identification and/or isolation of one or more target nucleic acid sequences (e.g., via hybridization capture/pull-down assays using one or more probes) in the sample. One or more target nucleic acid sequences can be identified via using nucleic acid probes having at least partial complementarity to a nucleic acid sequence of interest. The composition as disclosed herein can comprise the nucleic acid probe(s). In some examples, a nucleic acid probe can comprise an activatable reporter agent. The activatable reporter agent can be activated by either one of (i) hybridization of the nucleic acid probe to a target nucleic acid sequence in the sample (e.g., molecular beacon, eclipse probe, amplifluor probe, scorpions PCR primer, and light upon extension fluorogenic PCR primer (LUX primer)) and (ii) dehybridization of at least a portion of the individual nucleic acid probe that has been hybridized to the target nucleic acid sequence in the sample (e.g., a hydrolysis probe (e.g., TaqMan prob), dual hybridization probes, and QZyme PCR primer). In some examples, a nucleic acid probe as disclosed herein can comprise a pull-down tag for capture/pull-down assays. The pull-down tag can be used to enrich a sample (e.g., a biological sample obtained or derived from the subject) for a specific subset (e.g., one or more target nucleic acid sequences). The pull-down tag can comprise a nucleic acid barcode (e.g., on either or both sides of the nucleic acid probe). By utilizing beads or substrates comprising nucleic acid sequences having complementarity to the nucleic acid barcode, the nucleic acid barcode can be used to pull-down and enrich for any nucleic acid probe that is hybridized to a target cell-free nucleic acid molecule. Alternatively or in addition to, the nucleic acid barcode can be used to identify the target cell-free nucleic acid molecule from any sequencing data (e.g., sequencing by amplification).

[0093] In some cases, the pull-down tag can comprise an affinity target moiety that can be specifically recognized and bound by an affinity binding moiety. The affinity binding moiety specifically can bind the affinity target moiety to form an affinity pair. In some cases, by utilizing beads or substrates comprising the affinity binding moiety, the affinity target moiety can be used to pull-down and enrich for any nucleic acid

probe that is hybridized to a target cell-free nucleic acid molecule. Alternatively, the pull-down tag can comprise the affinity binding moiety, while the beads/substrates can comprise the affinity target moiety. Non-limiting examples of the affinity pair can include biotin/avidin, antibody/antigen, biotin/streptavidin, metal/chelator, ligand/receptor, nucleic acid and binding protein, and complementary nucleic acids. In an example, the pull-down tag can comprise biotin.

[0094] In some cases, the nucleic analysis sample can be prepared for sequencing of one or more nucleic acid molecules in the sample. A sequencing method as disclosed herein can be a first-generation sequencing method (e.g., Maxam-Gilbert sequencing, Sanger sequencing). The sequencing method can be a high-throughput sequencing method, such as next-generation sequencing (NGS) (e.g., sequencing by synthesis). A high-throughput sequencing method can sequence simultaneously (or substantially simultaneously) at least about 10,000, at least about 100,000, at least about 1 million, at least about 10 million, at least about 100 million, at least about 1 billion, or more polynucleotide molecules (e.g., cell-free nucleic acid molecules or derivatives thereof). NGS can be any generation number of sequencing technologies (e.g., second-generation sequencing technologies, third-generation sequencing technologies, fourth-generation sequencing technologies, etc.). Non-limiting examples of high-throughput sequencing methods include massively parallel signature sequencing, polony sequencing, pyrosequencing, sequencing-by-synthesis, combinatorial probe anchor synthesis (cPAS), sequencing-by-ligation (e.g., sequencing by oligonucleotide ligation and detection (SOLiD) sequencing), semiconductor sequencing (e.g., Ion Torrent semiconductor sequencing), DNA nanoball sequencing, and single-molecule sequencing, sequencing-by-hybridization. The composition as disclosed herein can comprise one or more agents involved in or required for sequencing. For example, the composition can comprise one or more of the following members to perform nucleic acid sequencing: (1) a nucleic acid ligase (e.g., T4 DNA ligase for addition of single-stranded DNA or RNA oligos to target nucleic acid molecules), (2) a nucleic acid polymerase (e.g., DNA or RNA polymerase for PCR), or (3) a nucleic acid helicase (e.g., for nanopore sequencing).

[0095] In some embodiments of any one of the compositions disclosed herein, the heterologous antioxidant moiety in the composition can reduce a degree and/or rate of error (e.g., sequencing error, background error) in sequencing one or more nucleic acid molecules (e.g., cfDNA, cfRNA) in the composition by at least or up to about 5%,

at least or up to about 10%, at least or up to about 15%, at least or up to about 20%, at least or up to about 25%, at least or up to about 30%, at least or up to about 35%, at least or up to about 40%, at least or up to about 45%, at least or up to about 50%, at least or up to about 55%, at least or up to about 60%, at least or up to about 65%, at least or up to about 70%, at least or up to about 75%, at least or up to about 80%, at least or up to about 85%, at least or up to about 90%, at least or up to about 95%, at least or up to about 99%, or about 100%, as compared to a corresponding control composition that lacks the heterologous antioxidant moiety. The degree and/or rate of error in sequencing can be reduced by between about 1% and about 100%, between about 1% and about 80%, between about 1% and about 60%, between about 1% and about 50%, between about 5% and about 50%, between about 10% and about 50%, between about 10% and about 40%, between about 10% and about 30%, or between about 10% and about 20%. The degree and/or rate of error in sequencing can be an overall degree and/or rate of error in sequencing due to a plurality of different nucleotide mutations (e.g., a plurality of transversion mutations and/or transition mutations). Alternatively, the degree and/or rate of error in sequencing can be a degree and/or rate of error in sequencing due to a specific type of nucleotide mutation (e.g., a G>T transversion).

[0096] In some embodiments of any one of the compositions disclosed herein, the heterologous antioxidant moiety in the composition can enhance sensitivity or specificity of analyzing (e.g., probe-mediated identification or pull-down, sequencing) one or more nucleic acid molecules (e.g., cfDNA, cfRNA) in the composition by at least or up to about 5%, at least or up to about 10%, at least or up to about 15%, at least or up to about 20%, at least or up to about 25%, at least or up to about 30%, at least or up to about 35%, at least or up to about 40%, at least or up to about 45%, at least or up to about 50%, at least or up to about 55%, at least or up to about 60%, at least or up to about 65%, at least or up to about 70%, at least or up to about 75%, at least or up to about 80%, at least or up to about 85%, at least or up to about 90%, at least or up to about 100%, at least or up to about 150%, or at least or up to about 200%, as compared to a corresponding control composition that lacks the heterologous antioxidant moiety.

Methods

[0097] In another aspect, the present disclosure provides a method for generating any one of the compositions disclosed herein. As illustrated by a flowchart shown in Fig. 2, the method can comprise providing (i) any one of the nucleic acid molecule(s) disclosed herein and (ii) any one of the heterologous antioxidant moiety disclosed herein (process 210). The method can further comprise mixing the nucleic acid molecule(s) and the heterologous antioxidant moiety (process 220), to generate a composition for, e.g., (1) identification and/or isolation of one or more target nucleic acid sequences, (2) sequencing of one or more nucleic acid molecule(s) in the composition, or (3) transfer or storage of the nucleic acid molecule(s). In some examples, the heterologous antioxidant moiety can comprise a sulfinic acid group (e.g., a hypotaaurine). In some examples, heterologous antioxidant moiety can comprise a protein (e.g., catalase).

Devices

[0098] In another aspect, the present disclosure provides a method for holding any one of the nucleic acid molecule(s) disclosed herein. In some embodiments, the device can comprise any one of the heterologous antioxidant moieties disclosed herein. In some cases, the heterologous antioxidant moiety can be coupled (e.g., directly or indirectly) to a surface (e.g., an inner surface) of the device. For exemplifying some cases, the surface of the device can be coated with the heterologous antioxidant moiety. In some cases, the heterologous antioxidant moiety can be covalently attached to the surface of the device. The device can be usable for holding the nucleic acid molecule(s) for, e.g., (1) identification and/or isolation of one or more target nucleic acid sequences, (2) sequencing of one or more nucleic acid molecule(s) in the composition, or (3) transfer or storage of the nucleic acid molecule(s). In some examples, heterologous antioxidant moiety can comprise a protein (e.g., catalase).

[0099] Non-limiting examples of the device as disclosed herein can include syringes, syringe tips, tubes (e.g., conical tubes for collecting a biological sample or in vitro sample), vials (e.g., cryotube vials), pipette tips, plates (e.g., tissue culture places, PCR plates), etc.

Additional embodiments***Cell-free nucleic acid sequencing and detection***

[0100] Turning now to the drawings and data, embodiments related to cell-free nucleic acid sequencing and detection of cancer are provided. In some embodiments, cell-free nucleic acids (cfDNA or cfRNA) are extracted from a liquid biopsy and prepared for sequencing. In many embodiments, sequencing results of cell-free nucleic acids are analyzed by computational models to detect circulating tumor nucleic acid (ctDNA or ctRNA) sequences (e.g., sequences of nucleic acids that derive from a neoplasm). Accordingly, in various embodiments, neoplasms (including cancer) can be detected in an individual by extracting a liquid biopsy from the individual and sequencing the cell-free nucleic acids derived from that liquid biopsy to detect circulating tumor nucleic acid sequences, and the presence of circulating tumor nucleic acid sequences indicates that the individual has a neoplasm. In some embodiments, a clinical intervention is performed on the individual based on the detection of a neoplasm.

[0101] Provided in Fig. 3 is a process to perform a clinical intervention based on detecting circulating tumor nucleic acids in an individual's biological sample. In some embodiments, detection of circulating tumor nucleic acids indicates a neoplasm (e.g., cancer) is present, and thus appropriate clinical intervention can be performed.

[0102] Process 300 may comprise obtaining, preparing, and sequencing (301) cell-free nucleic acids obtained from a non-invasive biopsy (e.g., liquid or waste biopsy). In some embodiments, cfDNA and/or cfRNA is extracted from plasma, blood, lymph, saliva, urine, stool, and/or other appropriate bodily fluid. In some embodiments, a biopsy is extracted prior to any indication of cancer. In some embodiments, a biopsy is extracted to provide an early screen in order to detect a neoplasm (e.g., cancer). In some embodiments, a biopsy is extracted to detect if residual neoplasm (e.g., cancer) exists after a treatment. Screening of any particular cancer can be performed. For more on examples of cancers that can be detected for intervention, see the section entitled "Clinical Interventions."

[0103] In some embodiments, a biopsy is extracted from an individual with a known risk of developing cancer, such as those with a familial history of the disorder or have known risk factors (e.g., cigarette smoker). In many embodiments, a biopsy is

extracted from any individual within the general population. In some embodiments, a biopsy is extracted from individuals within a particular age group with higher risk of cancer, such as aging individuals above the age of 50.

[0104] In many embodiments, extracted cell-free nucleic acids are prepared for sequencing. Accordingly, cell-free nucleic acids are converted into a molecular library for sequencing. In some embodiments, adapters and primers are attached onto cell-free nucleic acids to facilitate sequencing. In some embodiments, targeted sequencing of particular genomic loci is to be performed, and thus particular sequences corresponding to the particular loci are captured via hybridization prior to sequencing. In some embodiments, various reagents are included during the library and/or capture operations to mitigate confounding factors. In some embodiments, an antioxidant is included during one or more sequencing preparation operations to prevent oxidation of various nucleotides that result in nucleotide transversions. In some embodiments, the antioxidant hypotaurine is utilized in various sequencing preparation operations.

[0105] In some embodiments, any appropriate sequencing technique can be utilized that can detect sequence variations indicative of circulating tumor nucleic acids. Sequencing techniques include (but are not limited to) 454 sequencing, Illumina sequencing, SOLiD sequencing, Ion Torrent sequencing, single-read sequencing, paired-end sequencing, etc.

[0106] Process 300 analyzes (303) the cell-free nucleic acid sequencing result to detect circulating tumor nucleic acid sequences. Because neoplasms (especially metastatic tumors) are actively growing and expanding, neoplastic cells are often releasing biomolecules (especially nucleic acids) into the vasculature, lymph, and/or waste systems. In addition, due to biophysical constraints in their local environment, neoplastic cells are often rupturing, releasing their inner cell contents into the vasculature, lymph, and/or waste systems. Accordingly, it is possible to detect distal primary tumors and/or metastases from a liquid or waste biopsy.

[0107] In a number of embodiments, a cell-free nucleic acid sequencing result is analyzed to detect whether somatic single nucleotide variants (SNVs), copy number variations (CNVs), genomic position features, and/or germline SNVs exist within the cell-free nucleic acid sample. In some embodiments, presence of particular somatic SNVs, CNVs, genomic position features, and/or germline SNVs is indicative of circulating tumor nucleic acid sequences (and thus indicative of a tumor present). In various embodiments, a computational model is utilized to analyze detected somatic

SNVs, CNVs, genomic position features, and/or germline SNVs to determine whether these detected molecular elements are indicative of circulating tumor nucleic acids. In some embodiments, a computational model provides a relative indication (e.g., numerical confidence score) on whether a particular sample contains circulating tumor nucleic acids. In some embodiments, a computational model is trained on somatic SNVs, CNVs, genomic position features, and/or germline SNVs detected in patients and matched controls.

[0108] In some embodiments, confounding factors are removed from a cell-free nucleic acid sequencing result. It is now understood that clonal hematopoiesis (CH) is a confounding source of somatic SNVs and CNVs within a cell-free nucleic acid sample. Accordingly, in various embodiments, somatic SNVs and CNVs associated with CH are removed from further analysis. In some embodiments, somatic SNVs and CNVs derived from CH are determined for each particular individual analyzed. To detect an individual's particular somatic SNVs and CNVs derived from CH, leukocytes or white blood cells (WBCs) or hematopoietic cells of the individual are collected and their nucleic acids extracted and sequenced to detect somatic SNVs and CNVs derived from those cells. In some embodiments, somatic SNVs and CNVs detected in WBCs are removed during analysis of cell-free nucleic acid sequencing result.

[0109] Detection of circulating tumor nucleic acid sequences indicates that a neoplasm is present in the individual being examined. Accordingly, based on detection of circulating tumor nucleic acids, a clinical intervention may be performed (305). In some embodiments, a clinical procedure is performed, such as (for example) a blood test, medical imaging, physical exam, a tumor biopsy, or any combination thereof. In some embodiments, diagnostics are performed to determine the particular stage of cancer. In some embodiments, a treatment is performed, such as (for example) chemotherapy, radiotherapy, immunotherapy, hormone therapy, targeted drug therapy, medical surveillance, or any combination thereof. In some embodiments, an individual is assessed and/or treated by medical professional, such as a doctor, nurse, dietician, or similar.

[0110] While specific examples of processes for molecularly analyzing cell-free nucleic acids and performing a clinical intervention are described above, some operations of the process can be performed in different orders and certain operations may be optional. As such, some operations of the process may be used as appropriate

to the requirements of specific applications. Furthermore, any of a variety of processes for molecularly analyzing cell-free nucleic acids appropriate to the requirements of a given application can be utilized.

Sequence Library Preparation

[0111] Some embodiments are directed toward preparing a cell-free sample of nucleic acids, including cell-free DNA (cfDNA) and/or cell-free RNA (cfRNA), for sequencing. Accordingly, embodiments involve extracting nucleic acids from a biological sample having extracellular nucleic acids. Biological samples include (but not limited to) blood, plasma, lymphatic fluid, cerebral spinal fluid, saliva, urine, stool, etc. Cell-free nucleic acids can be isolated and purified by any appropriate means, as known in the art. In some embodiments, column purification is utilized (e.g., QIAamp Circulating Nucleic Acid Kit from Qiagen, Hilden, Germany). In some embodiments, isolated RNA fragments can be converted into complementary DNA for further downstream analysis.

[0112] Some embodiments are directed toward preparing cell-derived nucleic acid samples for sequencing. Accordingly, some embodiments isolate cells and or tissue to be analyzed (e.g., tumor cells, neoplastic cells, blood cells). Cells and tissue can be extracted and isolated as understood in the art. In some embodiments, blood cells (e.g., leukocytes) are isolated from plasma via centrifugation. Furthermore, nucleic acids from the cells and tissues can be isolated and purified by any appropriate means, as known in the art. In some embodiments, column purification is utilized (e.g., DNeasy Blood and Tissue Kit from Qiagen, Hilden, Germany). Nucleic acids can be broken down into smaller fragments (e.g., 50-450 bp) for library preparation by any appropriate means (e.g., sonication).

[0113] In some embodiments, isolated nucleic acid fragments can be prepared into a sequencing library. In many embodiments, adapters having unique identifiers (UIDs) and dual index sample barcodes, each with optimized GC content and sequence diversity, are utilized to build a library. In many of these embodiments, the UIDs and dual index barcodes are decoupled (e.g., each are distinct barcodes). In some embodiments, the UIDs are predefined (e.g., not random) sequences to provide an error-correcting benefit. Errors in UIDs or sample barcodes are often introduced during library preparation, which can lead to inaccurate enumeration of unique molecules observed by sequencing. To correct these errors, some embodiments utilize pre-

defined sequences with pair-wise Hamming edit distances, which can be utilized for error correction. For example, when 6 bp UID sequences are utilized, the sequences can be designed with pair-wise Hamming edit distances ≥ 3 , enabling correction of 1 bp errors and detection of 2 bp errors. Likewise, when 8 bp sample barcodes sequences are utilized, the sequences can be designed with pair-wise Hamming edit distances ≥ 5 , which enables correction of 1 or 2 bp errors and detection of 3 bp errors.

[0114] Some embodiments are directed toward library molecules to be used in a sequencing reaction. In some embodiments, nucleic acids are DNA, and thus can be used directly for library preparation. In some embodiments, nucleic acids are RNA, and thus conversion into cDNA is necessary before library preparation. In many embodiments, a pair of error-correcting UID is attached to the DNA (or cDNA) fragment such that DNA (or cDNA) is flanked by on each side by the UID. A pair of flanking UIDs provides an indication of a particular nucleic acid molecule derived from a biological source, which may enable more accurate enumeration of original unique molecules (e.g., each pair of UIDs indicates a ligation event of that nucleic acid molecule which occurs prior to amplification operations, enabling identification of duplicate molecules that arise due to amplification operations). In some embodiments, a pair of index sample barcodes is attached to the DNA (or cDNA) fragment such that DNA (or cDNA) is flanked by on each side by the index sample barcodes, which indicate the sample source (e.g., all molecules derived from a sample are flanked with the pair of index sample barcodes). In some embodiments, the use of dual index sample barcodes better ensures that a sequencing product is in fact a *bona fide* product from the sample source, as determined by having both index barcodes properly flanked. In some embodiments, an isolated sample DNA (or cDNA) fragment incorporating flanking UIDs and flanking sample barcodes further incorporates an annealing site for a universal primer for PCR and/or sequencing.

[0115] In some embodiments, libraries are prepared for a number of samples that may be combined to perform sequencing. Accordingly, in many of these embodiments, each sample has its own sample-specific error-correcting barcode, which may be derived from a grafting PCR. Further, in some embodiments, each sample library share the same universal PCR primer annealing sequence(s), which allows for the combined samples to be amplified in the same reaction prior to sequencing. And in some embodiments, the combined samples are sequenced in the same reaction.

[0116] In some embodiments, libraries are enhanced to help detect certain molecular elements, such as (for example), single nucleotide variants (SNVs) in particular loci of the genome. Enhancement may be necessary in order to be able to detect molecular elements above the limit of detection, especially when the molecular elements are rare and/or somatic SNVs. Accordingly, in some embodiments, targeted sequencing is performed on prepared libraries. In many embodiments, capture hybridization is utilized to selectively pull down library molecules having a particular sequence (e.g., sequence of genomic loci of interest). In some embodiments, captured hybridization is performed on a library to pull down DNA molecules with specific genomic loci in order to detect molecular features in those loci via sequencing. In some embodiments, captured hybridization is performed on a library in order to detect rare and/or somatic SNVs in genomic loci known to harbor SNVs involved in cancer and/or oncogenic pathology. In some embodiments, captured hybridization is performed on a library in order to detect rare and/or somatic SNVs in genomic loci known to harbor SNVs, as detected in a prior sequencing result of a tumor sample.

Capture Hybridization

[0117] Some embodiments utilize capture hybridization techniques to perform targeted sequencing. When performing sequencing on cell-free nucleic acids, in order to enhance resolution on particular genomic loci, library products can be captured by hybridization prior to sequencing. Capture hybridization can be particularly useful when trying to detect somatic variants and/or germline variants from a sample at particular genomic loci. In some situations, detection of somatic variants is indicative that the source of nucleic acids, including nucleic acids derived from a tumor or other neoplastic source. In some situations, identification of particular germline variants that are associated with neoplasm pathogenesis can provide support that a neoplasm is present. Accordingly, capture hybridization is a tool that can enhance detection of circulating tumor nucleic acids within cell-free nucleic acids.

[0118] One of the most common sequencing artifact observed in capture-based sequencing methods can be oxidation of guanine (G) that occurs during the hybrid capture step, which resulted in transformation of guanine into 8-oxoguanine (e.g., as observed through *in silico* analysis as provided herein). This unintended *in vitro* oxidation result can result in a G>T transversion, which can confound sequencing results, especially when searching for polymorphic variants in a sample. The G>T

transversions can be common mutagenesis events that occurs *in vivo*, especially in a neoplasm or cancer. Some environmental agents (e.g., UV radiation, cigarette smoke, free radicals) oxidize guanine (G) causing G>T transversions and thus a G>T transversion may have already occurred within the biological source prior to extraction (Figs. 4A and 4B). Accordingly, to mitigate the confounding *in vitro* mutagenesis, any one of the heterologous antioxidant moieties as disclosed herein (e.g., an enzyme and/or antioxidant) may prevent the oxidation occurring during hybrid capture. To confirm, enzymes and/or reactive oxygen species (ROS) scavengers were utilized (e.g., at 5 mM in a final mixture) to see which scavengers may prevent *in vitro* formation of 8-oxoguanine during capture hybridization. Enzymes tested included uracil-DNA glycosylase (UDG), Formamidopyrimidine [fapy]-DNA glycosylase (FPG), and catalase enzyme. Antioxidants tested included glutathione (GTT), hypotaurine, and sodium sulfite (Na₂SO₃). It was found that some these enzymes and compounds, e.g., hypotaurine and catalase, individually mitigated formation of 8-oxoguanine during capture hybridization (Figs. 4A and 4B).

[0119] In some embodiments, an antioxidant and/or enzyme is included during a hybrid capture assay. In some of these embodiments, the antioxidant is hypotaurine. Various embodiments are directed to capture hybridization methods in which hypotaurine is added to the hybridization reaction mixture. In many of these embodiments, hypotaurine is utilized within a sequencing protocol to mitigate the detection of *in vitro* G>T transversions in the sequencing result that occur during sequencing preparation. Accordingly, in some embodiments, hypotaurine is utilized to capture particular DNA molecules that are then used for a sequencing reaction.

Detection of Circulating tumor Nucleic Acids from Cell-Free Nucleic Acids

[0120] Some embodiments are directed to utilization of computational models to determine whether a cell-free nucleic acid sample includes circulating tumor nucleic acids. In some embodiments, SNVs and/or CNVs within a sequencing result of a cell-free nucleic acid sample are analyzed via computational models to determine whether the SNVs and/or CNVs are derived from circulating tumor nucleic acids. In some embodiments, computational models are trained on nucleic acid samples derived from cancer patients and unaffected individuals.

[0121] In some embodiments, the computational model can utilize a somatic single nucleotide variant module to determine whether a variant within a cell-free nucleic acid sequencing result is derived from circulating tumor nucleic acids. Somatic SNVs are highly common in nucleic acids derived from neoplastic cells, and thus are common in circulating tumor nucleic acids. Accordingly, detection of somatic SNVs in a cell-free nucleic acid sequencing result provides an indication that the source of the SNV is from neoplastic tissue.

[0122] Although somatic SNVs are often derived from neoplastic tissue, detected somatic SNVs can often arise due to reasons other than neoplastic growth, including (but not limited to) natural aging, clonal hematopoiesis, and other innocuous sources. It is therefore beneficial to utilize a system capable of accurately predicting whether a detected SNV is derived from a neoplastic source. In some embodiments, a computational model is utilized to provide an indication of whether a detected SNV in a cell-free nucleic acid sequencing result is truly derived from circulating tumor nucleic acid molecules.

[0123] In some embodiments, a model to identify SNVs derived from circulating tumor nucleic acid molecules integrates biological and technical features that are specific to each individual variant, including (but not limited to) background frequency of variant, fragment size of the cell-free nucleic acid molecule, variant signatures common to a particular source, presence in genomic loci (e.g., oncogenic genes) frequently mutated in cancer (or in particular cancer type), the likelihood that the variant is derived from CH, and whether or not the presence of the mutation may be confidently assessed in host hematopoietic cells relative to the VAF of the variant in the cfDNA and positional depth in the hematopoietic cells. For example, a set of model features can be used to determine whether a particular SNV is derived from circulating tumor nucleic acid molecules and their contribution to the model. This exemplary set of features includes WBC Bayesian background, cfDNA Bayesian background, variant allele frequency (VAF %), germline depth, mean barcode family size, short fragment score 1, short fragment score 2, transition/transversion, duplex support, pass outlier, mapping quality, cancer hotspot, UMI error corrected, Phred quality, and variant position in read. For details on these features, see the Exemplary Embodiments section. Although this exemplary set of features were developed specifically to identify ctDNA in non-small cell lung cancer (NSCLC), the same and/or similar set of features can be used in models for pan-cancer or other specific cancers as well. Accordingly,

various embodiments utilize a model to detect circulating tumor nucleic acids based on identification of SNVs that integrate one or more of the following features: cell-derived DNA Bayesian background, cfDNA Bayesian background, variant allele frequency (VAF %), germline depth, mean barcode family size, short fragment score 1, short fragment score 2, transition/transversion, duplex support, pass outlier, mapping quality, cancer hotspot, UMI error corrected, Phred quality, and variant position in read. In some embodiments, a model incorporates two or more of these features. In some embodiments, a model incorporates three or more of these features. In some embodiments, a model incorporates four or more of these features. In some embodiments, a model incorporates five or more of these features. In some embodiments, a model incorporates six or more of these features. In some embodiments, a model incorporates seven or more of these features. In some embodiments, a model incorporates eight or more of these features. In some embodiments, a model incorporates nine or more of these features. In some embodiments, a model incorporates ten or more of these features. In some embodiments, a model incorporates eleven or more of these features. In some embodiments, a model incorporates twelve or more of these features. In some embodiments, a model incorporates thirteen or more of these features. In some embodiments, a model incorporates fourteen or more of these features. In some embodiments, a model incorporates all fifteen of these features.

Clinical Interventions

[0124] Various embodiments are directed toward utilizing detection of cancer to perform clinical interventions. In some embodiments, an individual has a liquid or waste biopsy screened and processed by methods described herein to indicate that the individual has cancer and thus an intervention is to be performed. Clinical interventions include clinical procedures and treatments. Clinical procedures include (but are not limited to) blood tests, medical imaging, physical exams, and tumor biopsies. Treatments include (but are not limited to) chemotherapy, radiotherapy, immunotherapy, hormone therapy, targeted drug therapy, and medical surveillance. In some embodiments, diagnostics are performed to determine the particular stage of cancer. In some embodiments, an individual is assessed and/or treated by medical professional, such as a doctor, nurse, dietician, or similar.

A. Detection of Cancer for Clinical Intervention

[0125] In some embodiments as described herein a cancer can be detected utilizing a sequencing result of cell-free nucleic acids derived from blood, serum, cerebrospinal fluid, lymph fluid, urine or stool. In some embodiments, another host source is sequenced (e.g., hematopoietic cells) to provide a more robust determination of whether the sequencing result of cell-free nucleic acids includes sequences of circulating tumor nucleic acids. Use of hematopoietic cells for sequencing can help identify and remove confounding signals, such as somatic SNVs and CNVs derived from natural aging, clonal hematopoiesis, and other innocuous sources. Various embodiments utilize an antioxidant (e.g., hypotaurine) during hybrid capture in embodiments that perform targeted sequencing. In addition, some embodiments utilize computational models, including those described herein, to determine whether a sequencing result of cell-free nucleic acids includes sequences of circulating tumor nucleic acids based on a confidence score provided by the computational model. Accordingly, in some embodiments, cell-free nucleic acids are extracted, processed, and sequenced, and the sequencing result is analyzed to detect cancer. This process is especially useful in a clinical setting to provide a diagnostic scan.

[0126] An exemplary procedure for a diagnostic scan of an individual is as follows:

- (a) extract liquid or waste biopsy from individual
- (b) prepare and sequence cell-free nucleic acids and a host source (e.g., WBCs)
- (c) utilize sequencing results in one or more computational models to detect circulating tumor nucleic acid sequences within the cell-free nucleic acid sequencing result
- (d) perform clinical intervention based on detection of circulating tumor nucleic acid sequences

[0127] In various embodiments, diagnostic scans can be performed for any neoplasm type, including (but not limited to) acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), anal cancer, astrocytomas, basal cell carcinoma, bile duct cancer, bladder cancer, breast cancer, cervical cancer, chronic lymphocytic leukemia (CLL) chronic myelogenous leukemia (CML), chronic myeloproliferative neoplasms, colorectal cancer, endometrial cancer, ependymoma, esophageal cancer, esthesioneuroblastoma, Ewing sarcoma, fallopian tube cancer, gallbladder cancer, gastric cancer, gastrointestinal carcinoid tumor, hairy cell leukemia, hepatocellular

cancer, Hodgkin lymphoma, hypopharyngeal cancer, Kaposi sarcoma, Kidney cancer, Langerhans cell histiocytosis, laryngeal cancer, leukemia, liver cancer, lung cancer, lymphoma, melanoma, Merkel cell cancer, mesothelioma, mouth cancer, neuroblastoma, non-Hodgkin lymphoma, non-small cell lung cancer, osteosarcoma, ovarian cancer, pancreatic cancer, pancreatic neuroendocrine tumors, pharyngeal cancer, pituitary tumor, prostate cancer, rectal cancer, renal cell cancer, retinoblastoma, skin cancer, small cell lung cancer, small intestine cancer, squamous neck cancer, T-cell lymphoma, testicular cancer, thymoma, thyroid cancer, uterine cancer, vaginal cancer, and vascular tumors.

[0128] In some embodiments, diagnostic scans are utilized to provide an early detection of cancer. In some embodiments, diagnostic scans can detect cancer in individuals having stage I, II, or III cancer. In some embodiments, diagnostic scans are utilized to detect residual cancer in individuals after treatment of the cancer.

B. Cancer Diagnostics and Treatments

[0129] Some embodiments are directed toward performing a diagnostic scan on cell-free nucleic acids of an individual and then based on results of the scan indicating cancer, performing further clinical procedures and/or treating the individual.

[0130] In some embodiments, numerous types of neoplasms can be detected, including (but not limited to) acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), anal cancer, astrocytomas, basal cell carcinoma, bile duct cancer, bladder cancer, breast cancer, cervical cancer, chronic lymphocytic leukemia (CLL) chronic myelogenous leukemia (CML), chronic myeloproliferative neoplasms, colorectal cancer, endometrial cancer, ependymoma, esophageal cancer, esthesioneuroblastoma, Ewing sarcoma, fallopian tube cancer, gallbladder cancer, gastric cancer, gastrointestinal carcinoid tumor, hairy cell leukemia, hepatocellular cancer, Hodgkin lymphoma, hypopharyngeal cancer, Kaposi sarcoma, Kidney cancer, Langerhans cell histiocytosis, laryngeal cancer, leukemia, liver cancer, lung cancer, lymphoma, melanoma, Merkel cell cancer, mesothelioma, mouth cancer, neuroblastoma, non-Hodgkin lymphoma, non-small cell lung cancer, osteosarcoma, ovarian cancer, pancreatic cancer, pancreatic neuroendocrine tumors, pharyngeal cancer, pituitary tumor, prostate cancer, rectal cancer, renal cell cancer,

retinoblastoma, skin cancer, small cell lung cancer, small intestine cancer, squamous neck cancer, T-cell lymphoma, testicular cancer, thymoma, thyroid cancer, uterine cancer, vaginal cancer, and vascular tumors.

[0131] In some embodiments, once a diagnosis of neoplastic growth is indicated, some follow-up diagnostic procedures can be performed, including (but not limited to) physical exam, medical imaging, mammography, endoscopy, stool sampling, pap test, alpha-fetoprotein blood test, CA-125 test, prostate-specific antigen (PSA) test, biopsy extraction, bone marrow aspiration, and tumor marker detection tests. Medical imaging includes (but is not limited to) X-ray, magnetic resonance imaging (MRI), computed tomography (CT), ultrasound, and positron emission tomography (PET). Endoscopy includes (but is not limited to) bronchoscopy, colonoscopy, colposcopy, cystoscopy, esophagoscopy, gastroscopy, laparoscopy, neuroendoscopy, proctoscopy, and sigmoidoscopy.

[0132] In some embodiments, once a diagnosis of neoplastic growth is indicated, some treatments can be performed, including (but not limited to) surgery, chemotherapy, radiation therapy, immunotherapy, targeted therapy, hormone therapy, stem cell transplant, and blood transfusion. In some embodiments, an anti-cancer and/or chemotherapeutic agent is administered, including (but not limited to) alkylating agents, platinum agents, taxanes, vinca agents, anti-estrogen drugs, aromatase inhibitors, ovarian suppression agents, endocrine/hormonal agents, bisphosphonate therapy agents and targeted biological therapy agents. Medications include (but are not limited to) cyclophosphamide, fluorouracil (or 5-fluorouracil or 5-FU), methotrexate, thiotepa, carboplatin, cisplatin, taxanes, paclitaxel, protein-bound paclitaxel, docetaxel, vinorelbine, tamoxifen, raloxifene, toremifene, fulvestrant, gemcitabine, irinotecan, ixabepilone, temozolomide, topotecan, vincristine, vinblastine, eribulin, mutamycin, capecitabine, capecitabine, anastrozole, exemestane, letrozole, leuprolide, abarelix, buserlin, goserelin, megestrol acetate, risedronate, pamidronate, ibandronate, alendronate, zoledronate, tykerb, daunorubicin, doxorubicin, epirubicin, idarubicin, valrubicin mitoxantrone, bevacizumab, cetuximab, ipilimumab, ado-trastuzumab emtansine, afatinib, aldesleukin, alectinib, alemtuzumab, atezolizumab, avelumab, axitinib, belimumab, belinostat, bevacizumab, blinatumomab, bortezomib, bosutinib, brentuximab vedotin, briatinib, cabozantinib, canakinumab, carfilzomib, certinib, cetuximab, cobimetnib, crizotinib, dabrafenib, daratumumab, dasatinib, denosumab, dinutuximab, durvalumab, elotuzumab, enasidenib, erlotinib, everolimus,

gefitinib, ibritumomab tiuxetan, ibrutinib, idelalisib, imatinib, ipilimumab, ixazomib, lapatinib, lenvatinib, midostaurin, nectinmab, neratinib, nilotinib, niraparib, nivolumab, obinutuzumab, ofatumumab, olaparib, loaratumab, osimertinib, palbociclib, panitumumab, panobinostat, pembrolizumab, pertuzumab, ponatinib, ramucirumab, reorafenib, ribociclib, rituximab, romidepsin, rucaparib, ruxolitinib, siltuximab, sipuleucel-T, sonidebib, sorafenib, temsirolimus, tocilizumab, tofacitinib, tositumomab, trametinib, trastuzumab, vandetanib, vemurafenib, venetoclax, vismodegib, vorinostat, and ziv-aflibercept. In some embodiments, an individual may be treated, by a single medication or a combination of medications described herein. A common treatment combination is cyclophosphamide, methotrexate, and 5-fluorouracil (CMF).

[0133] Many embodiments are directed to diagnostic or companion diagnostic scans performed during cancer treatment of an individual. When performing diagnostic scans during treatment, the ability of agent to treat the neoplastic growth can be monitored. Most anti-cancer therapeutic agents result in death and necrosis of neoplastic cells, which may release higher amounts nucleic acids from these cells into the samples being tested. Accordingly, the level of circulating tumor nucleic acids can be monitored over time, as the level may increase during treatments and begin to decrease as the number of neoplastic cells are decreased. In some embodiments, treatments are adjusted based on the treatment effect on neoplastic cells. For instance, if the treatment isn't cytotoxic to neoplastic cells, a dosage amount may be increased or an agent with higher cytotoxicity can be administered. In the alternative, if cytotoxicity of neoplastic cells is good but unwanted side effects are high, a dosage amount can be decreased or an agent with less side effects can be administered.

[0134] Various embodiments are also directed to diagnostic scans performed after treatment of an individual to detect residual disease and/or recurrence of neoplastic growth. If a diagnostic scan indicates residual and/or recurrence of neoplastic growth, further diagnostic tests and/or treatments may be performed as described herein. If the neoplastic growth and/or individual is susceptible to recurrence, diagnostic scans can be performed frequently to monitor any potential relapse.

EXAMPLES

[0135] The embodiments of the present disclosure may be better understood with the several examples provided within. Many exemplary results of cell free nucleic acid sequencing tools and methods are described. Also provided are description of diagnostics, especially for non-small-cell lung cancer (NLCLC).

Example 1: Integrating genomic features for noninvasive early lung cancer detection

[0136] Lung cancer is the leading cause of cancer deaths and the majority of patients are diagnosed with metastatic disease that is generally incurable. Nevertheless, a significant fraction of patients with localized disease (stage I-III) can be cured, illustrating the utility of early detection. Indeed, screening of high-risk adults via low-dose computed tomography (LDCT) scans reduces lung cancer-related mortality, and as a result, annual radiologic screening may be recommended for high-risk populations. Despite its efficacy, the clinical utility of LDCT screening is complicated by a high false discovery rate (>90%) and low compliance, with <5% of eligible individuals in the US currently undergoing screening. Multiple factors contribute to this low adoption rate, including limited access to qualified radiology centers and patient inconvenience. Therefore, there is an unmet need for new approaches to improve early detection of early stage resectable lung cancers in high risk individuals.

[0137] Noninvasive blood tests that can detect tumor-derived somatic alterations based on the analysis of cfDNA are attractive candidates for cancer screening applications due to the relative ease of obtaining blood specimens. However, cfDNA assays currently in clinical use are intended for noninvasive genotyping of patients with advanced disease where ctDNA levels are significantly higher than in patients with early stage tumors. Separately, some studies examining ctDNA in patients with localized non-small-cell lung cancers (NSCLC) may use tumor-informed approaches where tumor tissue must be genotyped first. While this approach maximizes sensitivity, it may not be useful for screening. Lastly, clonal hematopoiesis (CH), which involves acquisition of somatic alterations in non-malignant hematopoietic progenitors and produces mutant cell-free DNA fragments, complicates use of ctDNA for early cancer detection.

[0138] Described within this example are methodological enhancements to Cancer Personalized Profiling by deep Sequencing (CAPP-Seq) that facilitate detection of ctDNA in early stage cancers or detection of residual cancer after treatment (for more on CAPP-Seq, see A. M. Newman *Nat. Biotechnol.* **34**, 547–555 (2016), which is incorporated herein by reference). The improved method was applied to plasma and tumor samples from patients with early stage NSCLC, initially employing a tumor-informed strategy to determine the fraction of patients whose tumors shed detectable ctDNA. The method was extended to early detection using a tumor-naïve approach to screen plasma samples from lung cancer patients and controls at high risk for lung cancer. It was found that cfDNA from both cases and controls harbor circulating somatic variants, the majority of which can be attributed to CH. Importantly, key molecular features were identified, including mutational signatures and fragment length profiles that distinguish CH variants from tumor-derived mutations. Finally, these findings were leveraged to develop and independently validate a Lung Cancer Likelihood in Plasma (Lung-CLiP) assay for noninvasive early lung cancer detection.

Improving detection of ultra-rare circulating variants

[0139] It has been demonstrated that ctDNA levels in localized lung cancers are low, with the majority of patients with stage I disease having circulating variant allele frequency (VAF) levels below about 0.1%. To improve sensitivity for detection of such low allelic levels, a few methodologies were developed and tested for maximizing the yield of unique, successfully sequenced cfDNA molecules while simultaneously minimizing their associated sequencing error profile (Fig. 5).

[0140] A new adapter schema was developed for library preparation by combining dual-indexed error-correcting sample barcodes, which guard against sample cross-contamination, with error-correcting duplex molecular barcodes (e.g., unique identifiers or ‘UIDs’) that enable more accurate enumeration of unique cfDNA molecules. Furthermore, de-coupling of the UIDs and sample barcodes allows for independent tailoring of UID diversity and multiplexing capacity based on the application.

[0141] Using these custom adapters, we then sought to identify key operations associated with the largest loss of unique cfDNA molecules. To do so, individual strands of cfDNA fragments were tracked from the start of library preparation to their ultimate sequencing within an *in silico* simulation of the CAPP-Seq molecular biology

workflow. The simulation predicted that the largest losses occurred at the hybrid capture operation and were due to the typical input of only a small fraction of each amplified sequencing library into the hybridization reaction for target enrichment. This effect arises due to uneven representation of original molecules following PCR. Many hybrid capture sequencing methods multiplex samples in the capture operation (e.g. capture many samples together in a single reaction), and this can result in a small fraction of the total amount of each library being captured. For example, if one has 2,000 ng of each sequencing library and were to multiplex 20 samples into a single 1,000 ng capture reaction, only 2.5% (50 ng) of each individual sequencing library is input into the capture reaction. Increasing the fraction of library input into the reaction improves molecular recovery. For example, increasing the fraction of library input from 8.3% to 100% significantly improved recovery of both total unique molecules and the fraction of sourced cfDNA duplexes for which both strands were sequenced. Notably, increasing the input percentage of sequencing library from 8.3% to 25% achieved most of the possible gains in unique molecule recovery and inputting 50% or more improved the fraction of original cfDNA duplexes for which both strands were sequenced. In addition, the ratio of sequencing library input to capture baits (e.g. biotinylated oligonucleotides used to enrich for genomic regions of interest) also influences molecular recovery following the capture reaction.

[0142] It was additionally sought to further improve the technical error profile of CAPP-Seq. The most common sequencing artifact observed in CAPP-Seq and other hybrid capture-based sequencing methods are G>T transversions arising due to oxidative damage occurring during the hybrid capture reaction and leading to the generation of 8-oxoguanine (See A. M. Newman, *et al.*, *Nat. Biotechnol.* (2016), cited *supra*; and M. Costelloe, *et al.*, *Nucleic Acids Res.* **41**, 1–12 (2013), which is incorporated herein by reference). Interestingly, G>T transversions are also the most common base substitution in lung cancers, arising *in vivo* as a result of exposure to the carcinogens in cigarette smoke (Figs. 4A and 4B). Therefore, G>T transversions from *in vitro* oxidation during hybrid capture can mimic and confound detection of genuine lung cancer-derived mutations. It was hypothesized that the addition of a scavenger of reactive oxygen species (ROS) would reduce oxidative damage-derived G>T artifacts (Figs. 4A and 4B).

[0143] DNA was extracted from a biological sample (e.g., tumor biopsy samples), quantified and fragmented, and sheared DNA (e.g., less than about 100 nanograms) was used for library preparation (e.g., coupling to an adapter comprising a sample barcode). After library preparation, hybrid capture (e.g., SeqCap EZ Choice, NimbleGen) was performed. This study utilized a custom 355 kb NSCLC focused panel targeting 255 genes that are recurrently mutated in lung cancer and 11 genes that are canonically associated with clonal hematopoiesis. Hybrid capture was performed according to the manufacturer's protocol, with the exception that heterologous antioxidant moieties (e.g., hypotaurine) was added to the hybrid capture reaction at a final working concentration (e.g., 5 mM). All capture steps were conducted on a thermal cycler at 47 °C. After enrichment, libraries were sequenced (e.g., on an Illumina HiSeq4000) with 2 × 150-bp paired-end reads. Sequencing lane share was determined based on cfDNA input and the desired barcode family size. Median sequencing depths were 23,570×/5,012× (nominal/unique) for cases and 19,534×/4,075× for controls.

[0144] After testing several antioxidants and free-radical scavengers, hypotaurine, a sulfinic acid, was identified as a favorable candidate. Hypotaurine is a naturally occurring intermediate of the cysteine-to-taurine pathway and has a non-enzymatic protective effect against ROS. When we compared the error profiles of cfDNA samples from 12 healthy adults captured with and without hypotaurine, it was found that samples captured with the ROS scavenger had significantly lower background error-rates and fewer G>T errors (Wilcoxon rank-sum test $P < 0.001$, Fig. 6). A similar relative reduction of G>T errors (16% vs 57% of all errors, Wilcoxon rank-sum test, $P < 1 \times 10^{-8}$) and background error rate (about 50% reduction, Wilcoxon rank-sum test, $P < 0.0001$) was observed in 104 healthy control cfDNA samples captured with the ROS scavenger compared to 69 control cfDNA samples captured without hypotaurine (Fig. 7).

A method for estimating lung cancer likelihood in plasma

[0145] A Lung Cancer Likelihood in Plasma (Lung-CLiP) assay was developed. A probabilistic approach was utilized to estimate the likelihood that a plasma sample contains tumor-derived cfDNA without using prior knowledge of tumor variants. This approach involves deep sequencing of plasma cfDNA and matched leukocytes and integrates both SNVs and genome-wide copy number analysis. The Lung-CLiP assay

was trained using samples from a discovery cohort of 104 lung cancer patients and 56 high-risk controls undergoing annual radiologic screening for lung cancer at 4 cancer centers. To develop the assay, a multi-tiered machine learning approach was employed in which a model was first trained to estimate the probability that a given cfDNA SNV is tumor-derived. The SNV model leverages key biological and technical features specific to each individual variant including background frequencies, cfDNA fragment size, smoking signature contribution, presence in a gene frequently mutated in NSCLC, and CH likelihood. Additionally, to identify copy number variants (CNVs), the genome was binned into 5 MB regions and both the on- and off-target sequencing reads from CAPP-Seq were used to identify genome-wide copy number alterations. The results of the SNV model were integrated with genome-wide copy number alterations (generated via analysis of both on- and off-target sequencing reads) within a final patient-level probabilistic classifier that estimates the likelihood a given blood sample contains lung cancer derived cfDNA (*i.e.*, “CLiP score”).

[0146] Lung-CLiP scores were compared to tumor-informed ctDNA levels and clinicopathological features. Importantly, sensitivities at 98% specificity were not significantly different than those observed using tumor-informed ctDNA analysis, indicating that Lung-CLiP achieves sensitivities similar to tumor-informed ctDNA detection.

Study design and patients

[0147] All biospecimens analyzed in this study were collected with informed consent from subjects enrolled on Institutional Review Board-approved protocols at their respective centers, including Stanford University, MD Anderson Cancer Center, Mayo Clinic, Vanderbilt University Medical Center, and Massachusetts General Hospital. All patients were de-identified and had AJCC v7 stage I-III NSCLC and received curative-intent treatment with surgery or radiotherapy.

[0148] This study consisted of two cohorts, a discovery cohort and a validation cohort. The discovery cohort consisted of two groups of patients: (1) tumor-informed NSCLC patients (and (2) Lung-CLiP training NSCLC cases. These two groups consisted of lung cancer patients enrolled at Stanford University (n=80), Vanderbilt University (n=21), Mayo Clinic (n=14) and MD Anderson Cancer Center (n=7) between November of 2009 and July of 2018. The tumor-informed NSCLC cases consisted of 85 patients with matched tumor tissue available, the majority of which (67/85) were

analyzed with all aspects of the improved CAPP-Seq workflow described in Fig. 5. The Lung-CLiP training group was restricted only to patients analyzed with the improved workflow (n=104) and studied for the tumor-naïve analyses, serving as the training group for the Lung-CLiP classifier. Among the 104 Lung-CLiP training NSCLC cases, 67 overlap with the 85 patients in the tumor-informed group. After initial training of a noninvasive classifier, NSCLC patients in the independent validation cohort (46 lung cancer cases) were prospectively enrolled at Massachusetts General Hospital (MGH) between January and December of 2018.

[0149] The discovery cohort consisted of two separate control groups. The first group consisted of 42 adult blood donors who were un-matched for risk (“low-risk controls”). The second group consisted of 56 age-, sex- and smoking status-matched adults (“risk-matched controls”) who had negative low-dose computed tomography (LDCT) screening scans for lung cancer at Stanford University and served as the training group for the Lung-CLiP classifier. The validation cohort contained a third control group, comprised of 48 risk-matched adults undergoing LDCT screening at Massachusetts General Hospital that were prospectively enrolled between January and December of 2018. This control group was only considered for the validation of the Lung-CLiP model.

Blood collection and processing

[0150] Whole blood collected in K₂EDTA tubes was processed immediately or within 4 hours following storage at 4 °C. Whole blood collected in Cell-Free DNA BCT (STRECK) tubes was processed within 72 hours. K₂EDTA tubes were centrifuged once at 1,800 x g for 10 min and STRECK tubes were centrifuged twice at 1,600 x g for 10 min at room temperature. Following centrifugation, plasma was stored at -80°C in 1.8 ml aliquots until cfDNA isolation. Plasma-depleted whole blood was stored at -80°C for DNA isolation from leukocytes.

[0151] Cell-free DNA was extracted from 2 to 16 mL of plasma (median of 3.6 mL) using the QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer’s instructions. After isolation, cfDNA was quantified using the Qubit dsDNA High Sensitivity Kit (Thermo Fisher Scientific) and High Sensitivity NGS Fragment Analyzer (Agilent). Genomic DNA (gDNA) from matched plasma-depleted whole blood (i.e. “WBCs” or “leukocytes”) was extracted using the Qiagen DNeasy Blood and Tissue kit, quantified using Qubit dsDNA High Sensitivity Kit, and

fragmented to a target size of 170 bp using Covaris S2 sonicator. Post-sonication, fragmented gDNA was purified using the QIAquick PCR Purification Kit (Qiagen). For cfDNA, a median of 38 ng (8-85 ng) was input into library preparation. DNA input was scaled to control for high molecular weight DNA contamination, targeting input of 40 ng of cfDNA in the 50-450 bp size range based on Fragment Analyzer data when available. For gDNA from leukocytes, ≤ 100 ng of fragmented gDNA was input into library preparation.

[0152] Logistical considerations related to the prospective collection of the validation cohort required the use of STRECK blood collection tubes, while K₂EDTA collection tubes were used for the training cohort. The study design guards against such pre-analytical variables driving classification of cases versus controls because all samples within the validation cohort (*i.e.*, cases and controls) were collected in STRECK tubes. Nevertheless, to confirm that the type of collection tube does not confound the Lung-CLiP model blood was collected from three healthy donors in K₂EDTA and STRECK tubes and compared key metrics including Lung-CLiP classification, cfDNA mutation concordance, fragment size, cfDNA concentration, molecular recovery and error profiles and found that none of these were significantly affected by the type of collection tube used.

Tumor tissue collection and processing

[0153] Tumor DNA was extracted from frozen biopsy samples using the Qiagen DNeasy Blood and Tissue kit or from FFPE biopsy samples using the Qiagen AllPrep DNA/RNA FFPE kit according to the manufacturer's instructions. Following extraction, DNA was quantified and fragmented in the same manner as gDNA from plasma depleted whole blood and ≤ 100 ng of sheared DNA was input into library preparation.

Library preparation and sequencing

[0154] A new adapter schema, Flexible Error-correcting duplex adapters ("FLEX adapters"), was developed that de-couples the portion of the adapter containing the duplex molecular barcode (*i.e.* unique identifier or "UID") from the portion containing the sample barcode. FLEX adapters utilize dual-index 8 bp sample barcodes (pairwise edit distances ≥ 5) and 6 bp error correcting UIDs (pairwise edit distances ≥ 3) with optimized GC content and sequence diversity. End repair, A-tailing, and adapter ligation are performed following the KAPA Hyper Prep Kit manufacturer's instructions

with ligation performed overnight at 4°C. Adapter ligation was performed using a partial Y adapter containing a 6 bp UID and the T overhang required for ligation. Following ligation, a bead cleanup was performed using SPRIselect magnetic beads (Beckman Coulter). Next, “grafting PCR” was performed to add dual-index 8 bp sample barcodes and the remaining adapter sequence necessary to make a functional Illumina sequencing library. Following another SPRI bead cleanup, universal PCR was performed.

[0155] Following library preparation, hybrid capture (SeqCap EZ Choice, NimbleGen) was performed. In this study a custom 355 kb NSCLC-focused panel targeting 255 genes recurrently mutated in lung cancer and 11 genes canonically associated with clonal hematopoiesis were utilized. Hybrid capture was performed, e.g., in the presence of the heterologous antioxidant moiety (e.g., hypotaurine). Following enrichment, libraries were sequenced on an Illumina HiSeq4000 with 2x150 bp paired-end reads.

Sequencing data analysis and variant calling

[0156] Fastq files were demultiplexed using a custom pipeline in which read pairs were only considered if both 8 bp sample barcodes and 6 bp UIDs matched expected sequences following error-correction. Following demultiplexing, UIDs were removed and adapter read-through was trimmed from the 3' end of the reads using AfterQC to preserve short fragments. Reads were aligned to the human reference genome (hg19) using BWA ALN.

[0157] *Error suppression and variant calling:* Molecular barcode-mediated error suppression and background polishing were performed as previously described (See A. M. Newman, *Nat. Biotechnol.* (2016), cited *supra*). To leverage the improved error profile afforded by capturing samples with the heterologous antioxidant moiety as disclosed herein (e.g., hypotaurine), a background database built from 12 withheld healthy control plasma samples captured with hypotaurine was used for background polishing. Following error suppression, selector-wide single nucleotide variant (SNV) calling was performed as previously described using a custom variant calling algorithm optimized for the detection of low allele frequency variants from deep sequencing data (See A. M. Newman, *Nat. Biotechnol.* (2016), cited *supra*). This approach, termed

“adaptive variant calling,” considers local and global variation in background error rates in order to determine position-specific variant calling thresholds within each sample.

DOCTRINE OF EQUIVALENTS

[0158] While the above description contains many specific embodiments, these should not be construed as limitations on the scope of the invention, but rather as an example of one embodiment thereof. Accordingly, the scope of the invention should be determined not by the embodiments illustrated, but by the appended claims and their equivalents.