

METHODS FOR SINGLE-ENDED OLIGONUCLEOTIDE ENRICHMENT AND SEQUENCING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of priority to U.S. Provisional Patent Application Serial No. 63/619,437, filed on January 10, 2024. The disclosure of the above-referenced application is herein expressly incorporated by reference in its entirety, including any drawings.

FIELD

[0002] The present disclosure relates to methods for single-ended oligonucleotide enrichment and sequencing for gene editor cleavage profiling.

BACKGROUND

[0003] *In vitro*/biochemical strategies to evaluate on- and off-target activity of site-specific nucleases generally utilize genomic DNA or a user-specified library of sequences. Genomic DNA approaches generally involve interrogation of off-target cleavage events in a set of DNA sequences in a relevant defined system (for example, the human genome). Such methods include CIRCLE-seq, SITE-seq, and Digenome-seq. However, such methods are restricted to the identification of off-target sites that are present in the particular genomic DNA sample used in the study. In contrast, user-specified library approaches, such as Nucleaseq and ONE-seq involve interrogation of a pre-enriched library of linear DNAs consisting of user-defined sequences that are generated by oligonucleotide synthesis. The library is then interrogated for those sequences that can be bound, modified or cleaved by sequence-specific enzymes or enzyme complexes. The library may include a collection of candidate off-target sequences derived from a biological source (e.g., from the sequence of a single genome or multiple genomes), or from a collection of sequences designed *in silico* with a set of base substitutions, insertions, or deletions relative to a target sequence. Random variant scanning libraries represent a systematic approach in which sequences are substituted with all possible alternative bases, insertions or deletions up to a certain number to provide insights into the biochemical function and specificity of various binding/modifying enzymes *in vitro* in a more unbiased fashion (e.g., ZFNs, TALENs, and CRISPR-Cas9). However, each of the above methods also possess significant limitations including high complexity, low accuracy, low reliability, and high costs.

[0004] The disclosure provided herein provides methods that employ a single pre-defined barcode, and analysis of only a single side of a cleavage product using UMIs to enable accurate and quantitative detection of off-target cleavage by gene editing nucleases with high sensitivity and consistency.

BRIEF SUMMARY

[0005] The present disclosure relates generally to the development of methods for quantitative detection of genomic cleavage. In particular, the present disclosure involves one-sided analysis of off-target cleavage by gene editing nucleases. The foregoing summary is illustrative only and is not intended to be in any way limiting. In addition to the illustrative embodiments and features described herein, further aspects, embodiments, objects and features of the disclosure will become fully apparent from the drawings and the detailed description and the claims.

[0006] Provided herein, among others, includes a method for quantifying off-target cleavage by a gene editing nuclease. The method includes performing *in vitro* cleavage of a library of dsDNA oligonucleotides with a gene editing nuclease such that a set of first dsDNA oligonucleotide fragments is formed, each said first dsDNA oligonucleotide fragment comprising a pre-defined barcode uniquely associated with one off-target sequence in the library, and a first fixed sequence present in every oligonucleotide in the library. The cleaved ends of the first oligonucleotide fragments are then end-repaired. The cleaved ends are ligated to an adaptor, wherein the adaptor comprises a UMI. An amplification reaction is then performed to amplify the set of first dsDNA oligonucleotide fragments using (i) a first primer comprising a nucleotide sequence complementary to a region of the adaptor and (ii) a second primer comprising a nucleotide sequence complementary to the first fixed sequence. The amplification products are then sequenced, and the reads derived from each cleaved off-target sequence in the library are quantified.

[0007] In some embodiments, the library contains sequences derived from a reference genome, sequences derived from more than one genome, or sequences corresponding to any defined set of variants with respect to an on-target sequence.

[0008] In some embodiments, the number of sequences in the library is between 10 and one billion.

[0009] In some embodiments, the first dsDNA oligonucleotide fragment comprises part of a PAM.

[0010] In some embodiments, the first dsDNA oligonucleotide fragment comprises part of a PS.

[0011] In some embodiments, the gene editing nuclease is a Cas nuclease. In some embodiments, the Cas nuclease is Cas9 or Cas12a.

[0012] In some embodiments, the library of dsDNA oligonucleotides comprises a UMI.

[0013] In some embodiments, sequencing comprises determining the sequence of the barcode and counting the reads, thereby quantifying off-target dsDNA sequences that are cleaved by a gene editing nuclease.

[0014] In some embodiments, the library is formed by synthesizing single-stranded DNA sequences, optionally on high-density oligonucleotide arrays, and performing a first amplification reaction to generate double stranded DNA (dsDNA) oligonucleotides.

[0015] In some embodiments, the first amplification reaction comprises limited cycle PCR.

[0016] In some embodiments, end-repairing comprises blunt ending and/or A tailing.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The features of the present disclosure are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present disclosure will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the disclosure are utilized, and the accompanying drawings of which:

[0018] FIG. 1 shows a diagram of Cas9 and Cas12a cleavage indicating the locations cleaved by different functional domains (dark arrows), secondary cleavage sites that may be due to trimming (lighter arrows), PAM location, target strand, and non-target strand.

[0019] FIG. 2 shows the workflow of an exemplary method of the present disclosure.

[0020] FIG. 3 shows the correlation of the disclosed method with standard ONE-seq data.

[0021] FIG. 4 shows a graphical summary of the effect of single mismatches (circles), insertions (squares) and deletions (triangles) at each position of the DNMT1 target site on the score relative to the on-target site with zero differences.

DETAILED DESCRIPTION

[0022] The present disclosure relates generally to new approaches for detecting off-target cleavage events that are associated with genome editing at specific target sites. These methods address several problems with current detection technologies, including low reliability, lack of

quantitative accuracy, and high costs. For example, the ONE-seq method requires large, expensive libraries of long synthetic oligonucleotides, and the final read counts and corresponding scores do not quantitatively reflect the true number of cleavage events. The reasons for this are as follows. The method requires a pair of identical barcodes (referred to as unique identifiers) that are associated with each candidate off-target sequence so that the data derived from each side of the starting library molecule can be identified and combined after cleavage. Surprisingly, however, the read counts from the two sides do not match. The read counts produced from non-target strand 5' ends after cleavage (PAM-side for Cas9; protospacer side for Cas12a) are nearly always significantly lower than the read counts from target strand 5' ends (protospacer-side for Cas9; PAM-side for Cas12a). For example, the Cas9 non-target strand read counts are typically only ~28% the value of the target strand read counts, on average, and the corresponding proportion of Cas12a non-target strand read counts is typically only ~10%. In addition, the ratio of target to non-target strand reads for individual off-target sequences varies significantly (from less than 2 to over 50) due to differential trimming and/or degradation. Furthermore, the statistical variance of combined Cas9 target and non-target strand scores is approximately twice that of the target strand scores alone (median coefficient of variation = 0.47 vs 0.24 across multiple sample replicates for all sites with read counts ≥ 10). Finally, the read counts do not quantitatively reflect the number of unique *in vitro* cleavage events because the cleavage products are amplified after adaptor ligation, and UMIs are not incorporated into each molecule immediately after cleavage.

[0023] The disclosure provided herein provides methods that employ a single pre-defined barcode, and analysis of only a single side of a cleavage product using UMIs to enable accurate and quantitative detection of off-target cleavage by gene editing nucleases with high sensitivity and consistency. The methods reduce the cost and error rate for oligonucleotide library synthesis by enabling shorter oligonucleotides and reduce sequencing costs by ~50%.

[0024] The section headings used herein are for organizational purposes only and are not to be construed as limiting the subject matter described.

[0025] Although various features of the disclosure can be described in the context of a single embodiment, the features can also be provided separately or in any suitable combination. Conversely, although the present disclosure can be described herein in the context of separate embodiments for clarity, the present disclosure can also be implemented in a single embodiment.

DEFINITION

[0026] The singular form “a”, “an”, and “the” include plural references unless the context clearly dictates otherwise. “A and/or B” is used herein to include all of the following alternatives: “A”, “B”, “A or B”, and “A and B.”

[0027] “Barcode,” as used herein, refers to one or more known nucleotide sequences that are used to identify a nucleic acid with which the barcode is associated, such as a unique gene editing nuclease off-target site. In some embodiments, a barcode enables multiplexing of products derived from different target sites in separate reactions or different samples.

[0028] “Index”, as used herein, refers to a class of barcode that is used to differentiate and group sets of sequencing reads derived from different samples after sequencing, by means of standard sequencing analysis software.

[0029] “UMI” (unique molecular identifier) as used herein, refers to one or more randomized or semi-randomized nucleotide sequences that are used to group and deduplicate all products derived from a single starting molecule after sequencing.

[0030] “Fixed sequence”, as used herein, refers to a nucleic acid sequence that is the same in each oligonucleotide of a library. There can be more than one fixed sequence present in an oligonucleotide and the fixed sequences may be different.

[0031] “Target”, as used herein, refers to a sequence that has homology to the intended on-target site for a gene editing nuclease. A site generically referred to as a target site may comprise an on-target sequence or an off-target sequence.

[0032] As used herein, the term “detecting” a nucleic acid molecule or fragment thereof refers to determining the presence of the nucleic acid molecule, typically when the nucleic acid molecule or fragment thereof has been fully or partially separated from other components of a sample or composition.

[0033] As used herein, the term “nuclease” refers to a polypeptide capable of cleaving the phosphodiester bonds between the nucleotide subunits of nucleic acids.

[0034] As used herein, the terms “nucleic acid”, “nucleic acid molecule” or “polynucleotide” are used herein interchangeably. They refer to a polymer of deoxyribonucleotides or ribonucleotides in either single- or double-stranded form, and unless otherwise stated, encompass known analogs of natural nucleotides that can function in a similar manner as naturally occurring nucleotides. The terms encompass nucleic acid-like structures with synthetic backbones, as well as amplification products. Oligonucleotides, DNAs, RNAs and genomes are all polynucleotides.

The polymer may include natural nucleosides (*i.e.*, adenosine, thymidine, guanosine, cytidine, uridine, deoxyadenosine, deoxythymidine, deoxyguanosine, and deoxycytidine), nucleoside analogs (e.g., 2-aminoadenosine, 2-thiothymidine, inosine, pyrrolo-pyrimidine, 3-methyl adenosine, C5-propynylcytidine, C5-propynyluridine, C5-bromouridine, C5-fluorouridine, C5-iodouridine, C5-methylcytidine, 7-deazaadenosine, 7-deazaguanosine, 8-oxoadenosine, 8-oxoguanosine, O(6)-methylguanine, and 2-thiocytidine), chemically modified bases (2'-O,4'-C-methylene bridged/locked nucleic acid), biologically modified bases (e.g., methylated bases), intercalated bases, spacers, modified sugars (e.g., 2'-fluororibose, ribose, 2'-deoxyribose, arabinose, and hexose), or modified phosphate groups (e.g., phosphorothioates and 5'-N-phosphoramidite linkages).

[0035] As used herein, the terms “amplify”, “amplified”, “amplification”, or “amplifying” as used in reference to a nucleic acid or nucleic acid reactions, refers to *in vitro* methods of making copies of a particular nucleic acid, such as a target nucleic acid, or a tagged nucleic acid produced, for example, by a method described herein.

[0036] A “primer” as used herein means a nucleic acid having a sequence complementary and specific to a known sequence in a target or template nucleic acid, e.g., DNA. This means that they must be sufficiently complementary to hybridize with their respective strands to form the desired hybridized products and then be extendable by a DNA polymerase. In some instances, the primer has exact complementarity to the target or template nucleic acid. However, in many situations, exact complementarity is not possible or likely, and one or more mismatches may exist which do not prevent hybridization or the formation of primer extension products using the DNA polymerase.

[0037] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range and any other stated or intervening value in that stated range, is encompassed within the disclosure. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges, and are also encompassed within the disclosure, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the disclosure.

[0038] All ranges disclosed herein also encompass any and all possible sub-ranges and combinations of sub-ranges thereof. Any listed range can be recognized as sufficiently describing

and enabling the same range being broken down into at least equal halves, thirds, quarters, fifths, tenths, and so forth. As a non-limiting example, each range discussed herein can be readily broken down into a lower third, middle third and upper third, and so forth. As will also be understood by one skilled in the art all language such as “up to,” “at least,” “greater than,” “less than,” and the like include the number recited and refer to ranges which can be subsequently broken down into sub-ranges as discussed above. Finally, as will be understood by one skilled in the art, a range includes each individual member. Thus, for example, a group having 1-3 articles refers to groups having 1, 2, or 3 articles. Similarly, a group having 1-5 articles refers to groups having 1, 2, 3, 4, or 5 articles, and so forth.

[0039] It is appreciated that certain features of the disclosure, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the disclosure, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable sub-combination. All combinations of the embodiments pertaining to the disclosure are specifically embraced by the present disclosure and are disclosed herein just as if each and every combination was individually and explicitly disclosed. In addition, all sub-combinations of the various embodiments and elements thereof are also specifically embraced by the present disclosure and are disclosed herein just as if each and every such sub-combination was individually and explicitly disclosed herein.

[0040] Although features of the disclosures may be described in the context of a single embodiment, the features may also be provided separately or in any suitable combination. Conversely, although the disclosures may be described herein in the context of separate embodiments for clarity, the disclosures may also be implemented in a single embodiment. Any published patent applications and any other published references, documents, manuscripts, and scientific literature cited herein are incorporated herein by reference for any purpose. In the case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

METHODS OF THE DISCLOSURE

[0041] As described in greater detail below, one aspect of the present disclosure provides a method of quantifying off-target cleavage by a gene editing nuclease that includes analysis of only one side of an off-target cleavage site by a gene editing nuclease. This is in contrast to

existing approaches which rely on analysis of both sides of the off-target cleavage site. The method involves performing *in vitro* cleavage of a library of double stranded oligonucleotides with a gene editing nuclease such that a set of first dsDNA oligonucleotide fragments is formed. Each first dsDNA oligonucleotide fragment includes a pre-defined barcode uniquely associated with one off-target sequence in the library and a first fixed sequence present in every oligonucleotide in the library. The method then involves end-repairing, and optionally A-tailing the cleaved ends of the first oligonucleotide fragments. The cleaved ends are ligated to an adaptor, and an amplification reaction is performed to amplify the set of first dsDNA oligonucleotide fragments using (i) a first primer comprising a nucleotide sequence complementary to a region of the adaptor and (ii) a second primer comprising a nucleotide sequence complementary to the first fixed sequence. The amplification products are sequenced, and the reads derived from each cleaved off-target sequence in the library are quantified.

Oligonucleotide Libraries

[0042] As described above, the methods described herein involve performing *in vitro* cleavage of a library of double stranded oligonucleotides. For the purposes of the methods described herein, a library of double stranded oligonucleotides can refer to a collection of oligonucleotide sequences that by one or more criteria have an increased or decreased probability of being cleaved by a gene editing nuclease of interest.

[0043] In some embodiments, oligonucleotide libraries as disclosed herein can be synthesized utilizing various solid-phase strategies. Nucleic acid sequences can be synthesized, for example, by the sequential addition of activated monomers to an elongating polynucleotide chain on a solid support (Caruthers, M. H. et al. (1992) *Meth Enzymol* 211:3), light-directed, spatial chemical synthesis (Fodor et al. (1991) *Science*, 251:767–773) or ink-jet deposition using phosphoramidite chemistry (Cleary et al. (2004) *Nature Methods* 1:241). Nucleic acid sequences can also be synthesized enzymatically by the sequential addition of monomers to an elongating nucleic acid chain. See, for example: Hoff et al. (2020) *ACS Synth. Biol.* 9:283 and Verardo et al. (2023) *Science Adv.* 9:eadi0263. In some embodiments, in lieu of synthesizing the desired sequences in the laboratory, essentially any nucleic acid, or library of nucleic acids, can be custom ordered from a variety of commercial sources, for example: IDT Corporation, Newark NJ, Agilent Technologies, Santa Clara, CA, Twist Bioscience, South San Francisco, CA.

[0044] An oligonucleotide can refer to a nucleic acid that comprises a string of nucleotides or analogues thereof. Oligonucleotides may be obtained by a number of methods including, for example, chemical synthesis, restriction enzyme digestion or PCR. As will be appreciated by one skilled in the art, the length of an oligonucleotide (*i.e.*, the number of nucleotides) can vary widely, often depending on the intended function or use of the oligonucleotide. Generally, oligonucleotides comprise between about 5 and about 300 nucleotides, for example, between about 15 and about 200 nucleotides, between about 15 and about 100 nucleotides, or between about 15 and about 50 nucleotides. Throughout the specification, whenever an oligonucleotide is represented by a sequence of letters (for example: A, C, G, and T, which denote adenosine, cytidine, guanosine, and thymidine, respectively), the nucleotides are presented in the 5' to 3' order from the left to the right. In certain embodiments, the sequence of an oligonucleotide includes one or more degenerate residues and/or modified phosphate groups as described herein.

[0045] In some embodiments, the library of oligonucleotides is initially synthesized on high-density oligonucleotide arrays as individual single-stranded DNA sequences, each bearing a unique identifier/barcode, which is present on one side of the oligonucleotide (BC in FIG. 2). The synthesized oligonucleotides can then be released from the chip and converted into double-stranded DNA molecules by priming against a first fixed sequence present in all DNA molecules synthesized on the chip. This pooled library is then incubated with a site-specific nuclease and enriched for cleaved sequences. The DNA sequences of cleaved sites can then be determined from the barcode that originally flanked these sites.

[0046] Synthesized molecules of the library can be specified to represent on-target and off-target sequences derived from a reference genome, sequences derived from more than one genome or sequences corresponding to any defined set of variants with respect to an on-target sequence. In one embodiment, sequences are derived from a set of 4,119 genomes from the 1000 Genomes and Human Genome Diversity Project data sets, using phased haplotypes to identify variants. Examples of a defined set of variants can include any set of differences whether represented by a sequenced genome or not, or a comprehensive set representing all possible changes with up to a certain number of differences relative to the target site, wherein a difference may be a mismatch, insertion, or deletion. As used herein, an on-target sequence can be a sequence identical to a guide RNA, and an off-target sequence can be a sequence with imperfect homology to a guide RNA.

[0047] In some embodiments, the oligonucleotide library includes at least about 1000, at least about 2000, at least about 3000, at least about 4000, at least about 5000, at least about 10,000, at least about 20000, at least about 30000, at least about 40000, at least about 50000, at least about 60000, at least about 70000, at least about 80000, at least about 90000, at least about 100000, at least about 200000, at least about 300000, at least about 400000, at least about 500000 and up to 10^6 , 10^7 , 10^8 , or 10^9 different sequences.

[0048] In some embodiments, following synthesis as single-stranded DNA sequences, a first amplification reaction is performed to amplify the oligonucleotide library and generate double stranded DNA oligonucleotides. Numerous methods of amplifying nucleic acids are known in the art and are described in more detail below.

[0049] In some embodiments, the first amplification reaction uses a high-fidelity polymerase. A “high-fidelity polymerase” can include a polymerase with a low error rate and several safeguards to protect against making or propagating mistakes while replicating nucleic acids. If an incorrect nucleotide does bind in the polymerase active site, incorporation is slowed due to the sub-optimal architecture of the active site complex. This lag time increases the opportunity for the incorrect nucleotide to dissociate before polymerase progression, thereby allowing the process to start again, with a correct nucleoside triphosphate. Some high-fidelity polymerases contain a 3' - 5' exonuclease activity known as “proofreading” to excise incorrectly incorporated mononucleotides and to replace them with the correct nucleotide. In some embodiments, the high-fidelity polymerase is a high-fidelity proofreading polymerase. Exemplary the high-fidelity proofreading polymerases include, without limitation, Watchmaker Equinox, Quantabio RepliQa, or Phusion polymerase.

[0050] In some embodiments, the first amplification reaction is a limited-cycle PCR reaction.

[0051] The number of amplification cycles in a limited PCR can comprise at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12 cycles, at least 13, at least 14, at least 15 cycles.

[0052] The cycling conditions of the amplification reaction will vary depending upon the step of the reaction. In some embodiments, an initial denaturing step can be performed at about 94, 95, 96, 97, 98, or 99 °C for about 10 seconds, about 20 seconds, about 30 seconds, about 1 minute, about 2 minutes, about 3 minutes, about 4 minutes, or about 5 minutes, or about 5 minutes, and a final extension step can be performed at about 67, 68, 69, 70, 71, 72, 73, 74, 75, or 76 °C for about 1 minute, about 2 minutes, about 3 minutes, about 4 minutes, about 5 minutes, about 6

minutes, about 7 minutes, about 8 minutes, about 9 minutes, or about 10 minutes. In some embodiments, the first amplification reaction is a limited-cycle PCR reaction using a high fidelity polymerase proofreading polymerase to minimize bias for G:C and A:T rich sequences. In some embodiments, the high-fidelity polymerase proofreading polymerase is Watchmaker Equinox, Quantabio RepliQa, or Phusion polymerase.

[0053] In some embodiments, the first amplification reaction is an isothermal amplification. The term “isothermal amplification” refers to a process in which a target nucleic acid is amplified using a constant, single, amplification temperature (e.g., from about 30°C to about 95°C). Unlike standard PCR, an isothermal amplification reaction does not include multiple cycles of denaturation, hybridization, and extension, of an annealed oligonucleotide to form a population of amplified target nucleic molecules. There are various types of isothermal application known in the art, including but not limited to, loop-mediated isothermal amplification (LAMP), nucleic acid sequence-based amplification NASBA, recombinase polymerase amplification (RPA), rolling circle amplification (RCA), nicking enzyme assisted amplification (NEAA), and helicase dependent amplification (HDA). In some embodiments, the isothermal application is nicking enzyme assisted isothermal amplification.

Gene Editing Nucleases

[0054] As described above, *in vitro* cleavage of a library of double stranded oligonucleotides with a gene editing nuclease is performed such that a set of first dsDNA oligonucleotide fragments is formed. Generally, the nuclease is site-specific in that it is known or expected to cleave only at a specific sequence or set of sequences, referred to herein as the nuclease's “target site”. However, site-specific nucleases may also cleave at off-target sequences that are homologous to the target sequence, containing up to 1, or 2, or 3, or 4, or 5, or 6, or 7, or 8, or 9, or 10 differences.

[0055] “Cleavage”, as used herein, can refer to the breakage of the covalent backbone of a DNA molecule. Cleavage can be initiated by a variety of methods including, but not limited to, enzymatic or chemical hydrolysis of a phosphodiester bond. Both single-stranded cleavage and double-stranded cleavage are possible, and double-stranded cleavage can occur as a result of two distinct single-stranded cleavage events. DNA cleavage as described herein can result in the production of either blunt ends or cohesive ends.

[0056] In methods presently disclosed herein, performing *in vitro* cleavage with the nuclease is generally carried out under conditions favorable for the cleavage by the nuclease. That is, even though a given candidate target site or variant target site might not actually be cleaved by the nuclease, the incubation conditions are such that the nuclease would have cleaved at least a significant portion (e.g., at least 1%, at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, or at least 95%) of templates containing its known target site. For known and generally well-characterized nucleases, such conditions are generally known in the art and/or can easily be discovered or optimized. For newly discovered nucleases, such conditions can generally be approximated using information about related nucleases that are better characterized (e.g., homologs and orthologs).

[0057] In some embodiments, cleavage reactions can be performed by mixing a guide RNA with a Cas9 or other nuclease at an RNP to DNA ratios of about 10:1, about 9:1, about 8:1, about 7:1, about 6:1, about 5:1, about 4:1, about 3:1, about 2:1, about 1:1, about 0.5:1, or about 0.1:1.

[0058] In some embodiments, cleavage reactions can be performed by mixing a guide RNA with a Cas9 or other nuclease at 15, 20, 25, 30, 35, 40, or 45 °C for 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120 minutes.

[0059] In some embodiments, the nuclease is an endonuclease. In some embodiments, the nuclease is a site-specific endonuclease (e.g., a restriction endonuclease, a meganuclease, a transcription activator-like effector nuclease (TALEN), a zinc finger nuclease, a site-specific recombinase, etc.).

[0060] In some embodiments, the site specificity of a site-specific nuclease is conferred by an accessory molecule. For example, the CRISPR-associated (Cas) nucleases are guided to specific sites by “guide RNAs” or gRNAs as described herein. In some embodiments, the nuclease is an RNA-guided nuclease. In some embodiments, the nuclease is a CRISPR-associated nuclease.

[0061] In some embodiments, the nuclease is a homolog or an ortholog of a previously known nuclease, for example, a newly discovered homolog or ortholog.

[0062] RNA-guided nucleases according to the present disclosure include, but are not limited to, naturally-occurring Class 2 CRISPR nucleases such as Cas9, and Cas12a, Cas12f, CasΦ, CasX, as well as mutants, or engineered nucleases with altered specificities derived or obtained therefrom. In functional terms, RNA-guided nucleases are defined as those nucleases that: (a) interact with (e.g., complex with) a gRNA; and (b) together with the gRNA, associate with, and optionally cleave or modify, a target region of a DNA that includes (i) a sequence

complementary to the targeting domain of the gRNA and, optionally, (ii) an additional sequence referred to as a “protospacer adjacent motif,” or “PAM,” which is described in greater detail below. RNA-guided nucleases can be defined, in broad terms, by their PAM specificity and cleavage activity, even though variations may exist between individual RNA-guided nucleases that share the same PAM specificity or cleavage activity. Skilled artisans will appreciate that some aspects of the present disclosure relate to systems, methods and compositions that can be implemented using any suitable RNA-guided nuclease having a certain PAM specificity and/or cleavage activity. For this reason, unless otherwise specified, the term RNA-guided nuclease should be understood as a generic term, and not limited to any particular type (e.g., Cas9 vs. Cas12a), species (e.g., *S. pyogenes* vs. *S. aureus*) or variation (e.g., full-length vs. truncated or split; naturally-occurring PAM specificity vs. engineered PAM specificity, etc.) of RNA-guided nuclease.

[0063] The PAM sequence takes its name from its sequential relationship to the “protospacer” sequence that is complementary to gRNA targeting domains (or “spacers”). Together with protospacer sequences, PAM sequences define target regions or sequences for specific RNA-guided nuclease/gRNA combinations.

[0064] Various RNA-guided nucleases may require different sequential relationships between PAMs and protospacers. In general, Cas9s recognize PAM sequences that are 3' of the protospacer as visualized relative to the guide RNA targeting domain. Cas12a, on the other hand, generally recognizes PAM sequences that are 5' of the protospacer.

[0065] In addition to recognizing specific sequential orientations of PAMs and protospacers, RNA-guided nucleases can also recognize specific PAM sequences. *S. aureus* Cas9, for instance, recognizes a PAM sequence of NNGRRT (SEQ ID NO: 8) or NNGRRV (SEQ ID NO: 9), wherein the N residues are immediately 3' of the region recognized by the gRNA targeting domain. *S. pyogenes* Cas9 recognizes NGG PAM sequences. And *F. novicida* Cas12a recognizes a TTN PAM sequence. PAM sequences have been identified for a variety of RNA-guided nucleases, and a strategy for identifying novel PAM sequences has been described by Shmakov et al., 2015, Molecular Cell 60, 385-397, Nov. 5, 2015. It should also be noted that engineered RNA-guided nucleases can have PAM specificities that differ from the PAM specificities of reference molecules (for instance, in the case of an engineered RNA-guided nuclease, the reference molecule may be the naturally occurring variant from which the RNA-guided nuclease

is derived, or the naturally occurring variant having the greatest amino acid sequence homology to the engineered RNA-guided nuclease).

[0066] In some embodiments, the set of first dsDNA oligonucleotide fragments comprises part of a PAM.

[0067] In some embodiments, the set of first dsDNA oligonucleotide fragments comprises part of a protospacer.

[0068] In addition to their PAM specificity, RNA-guided nucleases can be characterized by their DNA cleavage activity: naturally-occurring RNA-guided nucleases typically form DSBs in target nucleic acids, but engineered variants have been produced that generate only SSBs (discussed above) Ran & Hsu, *et al.*, *Cell* 154(6), 1380-1389, Sep. 12, 2013), or that that do not cut at all.

Cas9

[0069] In some embodiments of the methods described herein, the gene editing nuclease is Cas9. Crystal structures have been determined for *S. pyogenes* Cas9 (Jinek *et al.*, *Science* 343(6176), 1247997, 2014, and for *S. aureus* Cas9 in complex with a unimolecular guide RNA and a target DNA (Nishimasu, "Crystal Structure of Cas9 in Complex with Guide RNA and Target RNA", *Cell* 156(5):935-49 (2014); Anders *et al.*, *Nature*, 2014 Sep. 25; 513(7519):569-73; and Nishimasu, "Crystal Structure of Staphylococcus aureus Cas 9," *Cell* 162(5):1113-1126 (2015)).

[0070] A naturally occurring Cas9 protein comprises two lobes: a recognition (REC) lobe and a nuclease (NUC) lobe; each of which comprise particular structural and/or functional domains. The REC lobe comprises an arginine-rich bridge helix (BH) domain, and at least one REC domain (e.g., a REC1 domain and, optionally, a REC2 domain). The REC lobe does not share structural similarity with other known proteins, indicating that it is a unique functional domain. While not wishing to be bound by any theory, mutational analyses suggest specific functional roles for the BH and REC domains: the BH domain appears to play a role in gRNA:DNA recognition, while the REC domain is thought to interact with the repeat:anti-repeat duplex of the gRNA and to mediate the formation of the Cas9/gRNA complex.

[0071] The NUC lobe comprises a RuvC domain, an HNH domain, and a PAM-interacting (PI) domain (FIG. 1). The RuvC domain shares structural similarity to retroviral integrase superfamily members and cleaves the non-complementary (*i.e.*, bottom) strand of the target nucleic acid. It may be formed from two or more split RuvC motifs (such as RuvC I, RuvC II, and RuvC III in *S. pyogenes* and *S. aureus*). The HNH domain, meanwhile, is structurally similar

to HNN endonuclease motifs, and cleaves the complementary (*i.e.*, top) strand of the target nucleic acid. The PI domain, as its name suggests, contributes to PAM specificity.

[0072] While certain functions of Cas9 are linked to (but not necessarily fully determined by) the specific domains set forth above, these and other functions may be mediated or influenced by other Cas9 domains, or by multiple domains on either lobe. For instance, in *S. pyogenes* Cas9, as described in Nishimasu, “Crystal Structure of Cas9 in Complex with Guide RNA and Target RNA”, *Cell* 156(5):935-49 (2014), the repeat:antirepeat duplex of the gRNA falls into a groove between the REC and NUC lobes, and nucleotides in the duplex interact with amino acids in the BH, PI, and REC domains. Some nucleotides in the first stem loop structure also interact with amino acids in multiple domains (PI, BH and REC1), as do some nucleotides in the second and third stem loops (RuvC and PI domains).

Cas12a

[0073] In some embodiments of the methods described herein, the gene editing nuclease is Cas12a. The crystal structure of *Acidaminococcus sp.* Cas12a in complex with crRNA and a double-stranded (ds) DNA target including a TTTN PAM sequence has been solved by Yamano *et al.* (*Cell*. 2016 May 5; 165(4): 949-962). Cas12a, like Cas9, has two lobes: a REC (recognition) lobe, and a NUC (nuclease) lobe. The REC lobe includes REC1 and REC2 domains, which lack similarity to any known protein structures. The NUC lobe, meanwhile, includes three RuvC domains (RuvC-I, -II and -III) and a BH domain. However, in contrast to Cas9, the Cas12a REC lobe lacks an HNH domain, and includes other domains that also lack similarity to known protein structures: a structurally unique PI domain, three Wedge (WED) domains (WED-I, -II and -III), and a nuclease (Nuc) domain.

[0074] While Cas9 and Cas12a share similarities in structure and function, it should be appreciated that certain Cas12a activities are mediated by structural domains that are not analogous to any Cas9 domains. For instance, cleavage of the complementary strand of the target DNA appears to be mediated by the Nuc domain, which differs sequentially and spatially from the HNH domain of Cas9. Additionally, the non-targeting portion of Cas12a gRNA (the handle) adopts a pseudoknot structure, rather than a stem loop structure formed by the repeat:antirepeat duplex in Cas9 gRNAs.

[0075] In some embodiments, the nuclease is a homolog or an ortholog of a previously known nuclease, for example, a newly discovered homolog or ortholog.

Guide RNA (gRNA) Molecules

[0076] The terms “guide RNA” and “gRNA” refer to any nucleic acid that promotes the specific association (or “targeting”) of an RNA-guided nuclease such as a Cas9 or a Cas12a to a target sequence such as a genomic or episomal sequence in a cell. gRNAs can be unimolecular (comprising a single RNA molecule, and referred to alternatively as chimeric), or modular (comprising more than one, and typically two, separate RNA molecules, such as a crRNA and a tracrRNA, which are usually associated with one another, for instance by duplexing). gRNAs and their component parts are described throughout the literature, for instance in Briner *et al.* (Molecular Cell 56(2), 333-339, Oct. 23, 2014, which is incorporated by reference), and in Cotta-Ramusino.

[0077] In bacteria and archaea, type II CRISPR systems generally comprise an RNA-guided nuclease protein such as Cas9, a CRISPR RNA (crRNA) that includes a 5' region that is complementary to a foreign sequence, and a trans-activating crRNA (tracrRNA) that includes a 5' region that is complementary to, and forms a duplex with, a 3' region of the crRNA. While not intending to be bound by any theory, it is thought that this duplex facilitates the formation of—and is necessary for the activity of—the Cas9/gRNA complex. As type II CRISPR systems were adapted for use in gene editing, it was discovered that the crRNA and tracrRNA could be joined into a single unimolecular or chimeric guide RNA, in one non-limiting example, by means of a four nucleotide (e.g., GAAA) “tetraloop” or “linker” sequence bridging complementary regions of the crRNA (at its 3' end) and the tracrRNA (at its 5' end). (Mali et al. Science. 2013 Feb. 15; 339(6121): 823-826; Jiang et al. Nat Biotechnol. 2013 March; 31(3): 233-239; and Jinek et al., 2012 Science August 17; 337(6096): 816-821, all of which are incorporated by reference herein.)

[0078] Guide RNAs, whether unimolecular or modular, include a “targeting domain” that is fully or partially complementary to a target domain within a target sequence, such as a DNA sequence in the genome of a cell where editing is desired. Targeting domains are referred to by various names in the literature, including without limitation “guide sequences” (Hsu et al., Nat Biotechnol. 2013 September; 31(9): 827-832, incorporated by reference herein), “complementarity regions” (Cotta-Ramusino), “spacers” (Briner *et al.* (Molecular Cell 56(2), 333-339, Oct. 23, 2014) and generically as “crRNAs” (Jiang et al. Nat Biotechnol. 2013 March; 31(3): 233-239). Irrespective of the names they are given, targeting domains are typically 10-30 nucleotides in length, and in certain embodiments are 16-24 nucleotides in length (for instance,

16, 17, 18, 19, 20, 21, 22, 23 or 24 nucleotides in length), and are at or near the 5' terminus of in the case of a Cas9 gRNA, and at or near the 3' terminus in the case of a Cas12a gRNA.

[0079] In addition to the targeting domains, gRNAs typically (but not necessarily, as discussed below) include a plurality of domains that may influence the formation or activity of gRNA/Cas9 complexes. For instance, as mentioned above, the duplexed structure formed by first and secondary complementarity domains of a gRNA (also referred to as a repeat:anti-repeat duplex) interacts with the recognition (REC) lobe of Cas9 and can mediate the formation of Cas9/gRNA complexes. (Nishimasu et al., Cell 156, 935-949, Feb. 27, 2014, and Nishimasu et al., Cell 162, 1113-1126, Aug. 27, 2015, both incorporated by reference herein). It should be noted that the first and/or second complementarity domains may contain one or more poly-A tracts, which can be recognized by RNA polymerases as a termination signal. The sequence of the first and second complementarity domains are, therefore, optionally modified to eliminate these tracts and promote the complete *in vitro* transcription of gRNAs, for instance through the use of A-G swaps as described in Briner *et al.* (Molecular Cell 56(2), 333-339, Oct. 23, 2014, or A-U swaps. These and other similar modifications to the first and second complementarity domains are within the scope of the present disclosure.

[0080] Along with the first and second complementarity domains, Cas9 gRNAs typically include two or more additional duplexed regions that are involved in nuclease activity *in vivo* but not necessarily *in vitro*. (Nishimasu et al., Cell 162, 1113-1126, Aug. 27, 2015). A first stem-loop one near the 3' portion of the second complementarity domain is referred to variously as the "proximal domain," (Cotta-Ramusino) "stem loop 1" (Nishimasu et al., Cell 156, 935-949, Feb. 27, 2014, and Nishimasu et al., Cell 162, 1113-1126, Aug. 27, 2015) and the "nexus" (Briner *et al.* (Molecular Cell 56(2), 333-339, Oct. 23, 2014). One or more additional stem loop structures are generally present near the 3' end of the gRNA, with the number varying by species: *S. pyogenes* gRNAs typically include two 3' stem loops (for a total of four stem loop structures including the repeat:anti-repeat duplex), while *S. aureus* and other species have only one (for a total of three stem loop structures). A description of conserved stem loop structures (and gRNA structures more generally) organized by species is provided in Briner *et al.* (Molecular Cell 56(2), 333-339, Oct. 23, 2014).

[0081] While the foregoing description has focused on gRNAs for use with Cas9, it should be appreciated that other RNA-guided nucleases have been (or may in the future be) discovered or invented which utilize gRNAs that differ in some ways from those described to this point. For

instance, Cas12a is a recently discovered RNA-guided nuclease that does not require a tracrRNA to function. (Zetsche et al., 2015, Cell 163, 759-771 Oct. 22, 2015, incorporated by reference herein). A gRNA for use in a Cas12a genome editing system generally includes a targeting domain and a complementarity domain (alternately referred to as a “handle”). It should also be noted that, in gRNAs for use with Cas12a, the targeting domain is usually present at or near the 3' end, rather than the 5' end as described above in connection with Cas9 gRNAs (the handle is at or near the 5' end of a Cas12a gRNA).

[0082] Those of skill in the art will appreciate, however, that although structural differences may exist between gRNAs from different prokaryotic species, or between Cas12a and Cas9 gRNAs, the principles by which gRNAs operate are generally consistent. Because of this consistency of operation, gRNAs can be defined, in broad terms, by their targeting domain sequences, and skilled artisans will appreciate that a given targeting domain sequence can be incorporated in any suitable gRNA, including a unimolecular or chimeric gRNA, or a gRNA that includes one or more chemical modifications and/or sequential modifications (substitutions, additional nucleotides, truncations, etc.). Thus, for economy of presentation in this disclosure, gRNAs may be described solely in terms of their targeting domain sequences.

[0083] More generally, skilled artisans will appreciate that some aspects of the present disclosure relate to systems, methods and compositions that can be implemented using multiple RNA-guided nucleases. For this reason, unless otherwise specified, the term gRNA should be understood to encompass any suitable gRNA that can be used with any RNA-guided nuclease, and not only those gRNAs that are compatible with a particular species of Cas9 or Cas12a. By way of illustration, the term gRNA can, in certain embodiments, include a gRNA for use with any RNA-guided nuclease occurring in a Class 2 CRISPR system, such as a type II or type V or CRISPR system, or an RNA-guided nuclease derived or adapted therefrom.

Barcode Sequences

[0084] As described above, each first dsDNA oligonucleotide fragment includes a pre-defined barcode uniquely associated with one off-target sequence in the library. A “barcode” can refer to a molecular label, or identifier, that conveys or is capable of conveying information about a sample, sequencing read, or group of samples or sequencing reads (e.g., information about an analyte in a sample, a template, a bead, a microwell, a primer, a read, and/or a capture probe). A barcode can be part of an analyte, or independent of an analyte. A barcode can be attached to an

analyte. A particular barcode will be unique relative to other barcodes, and may include error-correcting features (e.g., Hamming codes) that enable correct identification of each barcode in a mixture even in the presence of sequencing errors.

[0085] As used in herein, barcodes can comprise custom polynucleotides of pre-defined sequence. The term “pre-defined” can mean that sequence of a barcode is predetermined or known prior to identifying or without the need to identify the sequence of the nucleic acid comprising the barcode. Barcodes can allow for identification and/or quantification of groups of sequencing reads that share the same barcode.

[0086] As used herein, the term “uniquely associated with one off-target sequence” can refer to a pre-defined barcode that is associated with a specific off-target sequence in the library and no other off-target sequence present in the library.

[0087] In some embodiments, the barcode has a number of nucleotide bases (sometimes referred to as a barcode length) of at least 3 bases, at least 4 bases, at least 5 bases, at least 6 bases, at least 7 bases, at least 8 bases, at least 9 bases, at least 10 bases, at least 11 bases, at least 12 bases, at least 13 bases, at least 14 bases, at least 15 bases, at least 16 bases, at least 17 bases, at least 18 bases, at least 19 bases, at least 20 bases at least 25 bases, at least 30 bases, at least 35 bases, at least 40 bases, at least 45 bases, or at least 50 bases. In some embodiments, the barcode has a number of nucleotide bases from 12 to 18 bases.

[0088] In some embodiments, the pre-defined barcode is an error-correcting barcode. As every assay using DNA barcodes is subject to errors introduced during DNA synthesis and sequencing, error-correcting barcodes enable correct identification of the pre-defined barcode sequence even in the presence of DNA synthesis and sequencing errors. Several types of error-correcting barcodes are known in the art including, without limitation, Hamming or Reed-Solomon codes, Levenshtein codes, filled/truncated Right End Edit (FREE) barcodes. Briefly, Hamming distance describes the number of substitutions between two sequences of equal length. Levenshtein codes, also known as edit codes, can theoretically account for all three types of common error: substitutions, insertions, and deletions, but only when the corrupted length of each barcode after errors is known. In contrast, FREE barcodes can correct substitutions, insertions, and deletions even when the edited length of the barcode is unknown (Hawkins et al., “Indel-correcting DNA Barcodes for High-Throughput Sequencing,” *Proc Natl Acad Sci USA* 115(27):E6217-E6226 (2018)).

Fixed Sequence

[0089] As described above, oligonucleotides of a library also comprise at least one fixed sequence, but typically contains multiple fixed sequences. Some fixed sequences can serve as spacers between functional elements in the library oligonucleotides to facilitate identification of such functional elements after sequencing. For example, a fixed sequence can be placed between an off-target sequence and a barcode. Fixed sequences can comprise primer binding sites for amplification of oligonucleotide libraries by PCR, or site-specific nicking enzyme recognition sites for amplification using NEAA. At least one fixed sequence, referred to as a first fixed sequence, must be present near one end of the amplified library to provide a primer binding site for amplification of the first set of cleavage products comprising a barcode. Fixed sequences can be incorporated into libraries from primer sequences via library amplification by PCR. Fixed sequences can be designed with any desired length appropriate to fulfill their functional purpose, for example as primer sequences or spacers. Fixed sequences can optionally be designed to include specific sequences such as modifying or restriction enzyme recognition sites. Fixed sequences can be constrained to maintain a desired range of G+C content, or to exclude certain sequences such as PAM sequences, homopolymers, or self-complementary regions. Fixed sequences can be restricted in length and number to minimize oligonucleotide synthesis and sequencing costs.

End-Repair/A-tailing

[0090] Following *in vitro* cleavage of the library of double stranded oligonucleotides, the cleaved ends of the first oligonucleotide fragments are end-repaired, and optionally A-tailed.

[0091] In some embodiments, end-repair involves generating blunt ends on the fragments that are amenable to ligation. In some embodiments, the end-repair comprises, e.g., treating the ends with a polymerase to fill in 5' overhangs. In some embodiments, the polymerase is a native or engineered type B proofreading polymerase (e.g., KOD, Phusion, Equinox, RepliQa), or *E. coli* DNA polymerase I large fragment. These enzymes fill in 5' overhangs and leave blunt ends. In some embodiments T4 DNA polymerase can be used to fill in 5' overhangs, remove 3' overhangs and leave blunt ends. In some embodiments a single-strand specific endonuclease such as Mung Bean nuclease, or single-strand specific exonuclease such as exonuclease T, can be used to remove single-stranded tails and leave blunt ends.

[0092] In some embodiments, end-repair can be performed by mixing the cleaved ends of the first oligonucleotide fragments with a proofreading polymerase and dNTPs. Typically, this reaction is performed at the optimal reaction temperature of the enzyme being used +/- 10 degrees. For example, blunt ending using Phusion polymerase may be performed at the optimal temperature of 72 °C, or at 62, 64, 66, 68, 70, 74, 76, 78, 80, or 82 °C for 1, 5, 10, 15, 20, 25, or 30 minutes.

[0093] In some embodiments, end-repair involves the addition of a deoxyadenosine nucleotide to the 3'-end of the double-stranded DNA product, using a polymerase lacking 3'-exonuclease activity. In some embodiments, the polymerase lacking 3'-exonuclease activity is Taq polymerase or Klenow exo- polymerase.

Adaptor Ligation

[0094] Following end repair and/or A-tailing of the cleaved and repaired ends of the first oligonucleotide fragments, the methods described herein involve ligating adaptors to the first oligonucleotide fragments.

[0095] The adaptors can be any type of adaptor known in the art including, but not limited to, conventional duplex or double stranded adaptors. In some embodiments, the adaptors can be partially double stranded DNA Y-adaptors. In some embodiments, the adaptors can be T-tailed adaptors. In an embodiment, the adaptors can be oligonucleotides comprising known sequence and, thus, allow generation and/or use of sequence specific primers for amplification and/or sequencing of any polynucleotides to which the adaptor(s) is appended or attached. Preferably, the adaptors can be any adaptors that can be marked and selected for by methods known in the art.

[0096] The adaptors used in the methods disclosed herein include a unique molecular identifier, or "UMI". A UMI comprises a short stretch of random nucleotides (e.g., NNNNNNNN (SEQ ID NO: 10)) or semi-random nucleotides (e.g., NWNWNWN (SEQ ID NO: 11)) included in an oligonucleotide or adapter. UMIs are typically between 6 and 25 bases, and are generally introduced by tagmentation or adapter ligation, although they may also be introduced by a single round of primer extension. UMIs serve to enable the identification, grouping and deduplication of extension or amplification products (e.g., PCR products) that derive from a single starting genomic DNA molecule. Such grouping enables counting of unique molecules in a starting sample that are represented by multiple reads after amplification and sequencing (e.g., counting

of unique cleavage events). UMIs can also be used to distinguishing between SNVs and misincorporation or sequencing errors in amplified molecules.

[0097] In some embodiments, the adaptors include a UMI and lack any 5' phosphate such that the adaptor may only ligate to the cleaved end of the strands containing a 5' phosphate, thereby tagging each cleaved molecule with a UMI.

Amplification

[0098] In the method described herein, after adaptor ligation, an amplification reaction is performed to amplify the set of first dsDNA oligonucleotide fragments using (i) a first primer comprising a nucleotide sequence complementary to a region of the adaptor and (ii) a second primer comprising a nucleotide sequence complementary to the first fixed sequence.

[0099] Numerous methods of amplifying nucleic acids are known in the art, and amplification reactions include polymerase chain reactions, ligation reactions, strand displacement amplification reactions, nicking enzyme assisted amplification, rolling circle amplification reactions, transcription-mediated amplification methods such as NASBA (e.g., U.S. Pat. No. 5,409,818), loop mediated amplification methods (e.g., "LAMP" amplification using loop-forming sequences, e.g., as described in U.S. Pat. No. 6,410,278). A nucleic acid that is amplified can be DNA comprising, consisting of, or derived from DNA or RNA or a mixture of DNA and RNA, including modified DNA and/or RNA. The products resulting from amplification of a nucleic acid molecule or molecules (i.e., "amplification products"), whether the starting nucleic acid is DNA, RNA or both, can be either DNA or RNA, or a mixture of both DNA and RNA nucleosides or nucleotides, or they can comprise modified DNA or RNA nucleosides or nucleotides. A "copy" does not necessarily mean perfect sequence complementarity or identity to the target sequence. For example, copies can include nucleotide analogs such as deoxyinosine or deoxyuridine, intentional sequence alterations (such as sequence alterations introduced through a primer comprising a sequence that is hybridizable, but not complementary, to the target sequence), and/or sequence errors that occur during amplification.

[0100] Protocols for amplification can include, e.g., one round of amplification or multiple rounds of amplification. For example, a first amplification round can be followed by a second amplification round with or without one or more processing (e.g., cleanup, concentrating, etc.) steps in between the two rounds of amplification. Additional rounds of amplification may be used in some embodiments, with or without one or more processing steps in between.

[0101] For PCR, the number of amplification cycles can be varied depending on the embodiment. For example, each round of amplification can comprise at least 4, at least 8, at least 10, at least 12, at least 14, at least 16, at least 18, at least 20, at least 22, at least 24, at least 26, at least 28, at least 30, at least 32, at least 34, at least 38, or at least 40 cycles. In embodiments involving more than one round of amplification, the number of amplification cycles used in each round can be the same or it can be different. In cases in which two rounds of amplification are used and the number of amplification cycles in the two rounds differ, the first round can comprise more cycles or it may comprise fewer cycles than the second round. As an example, a first round of amplification can comprise 12 cycles and a second round of amplification can comprise 15 cycles. As another example, a first round of amplification can comprise 10 cycles and a second round of amplification can comprise 12 cycles.

[0102] The temperature of the PCR amplification reaction will vary depending upon the step of the reaction. In some embodiments, an initial denaturing can be performed at about 93, 94, 95, 96, 97, 98, or 99 °C for about 1, 2, 3, 4, or 5 minutes, and a final extension can be performed at about 67, 68, 69, 70, 71, 72, 73, 74, or 75 °C for about 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 minutes.

Sequencing and Analysis

[0103] Following the amplification step, the amplification products are sequenced and reads derived from each cleaved off-target sequence in the library are then analyzed and quantified.

[0104] As used herein, “sequencing” includes any method of determining the sequence of a nucleic acid. Any method of sequencing can be used in the present methods. In preferred embodiments, Next Generation Sequencing (NGS), a high-throughput sequencing technology that performs millions or billions of sequencing reactions in parallel, is used. Although different NGS platforms vary in their assay chemistries and detection technologies, they all generate sequence data from a large number of reactions in parallel. Exemplary NGS sequencing systems include but are not limited to Illumina MiSeq, NextSeq and NovaSeq, MGI DNBSEQ, and PacBio Revio.

[0105] The sequencing data with associated quality scores provided by the NGS system are analyzed using a bioinformatics pipeline. The pipeline processes the data, removes low quality sequences, trims adaptor sequences, and identifies functional elements in the cleavage products such as different fixed sequences, barcodes, UMIs, and the target site. Quality control criteria are applied to eliminate reads with unreliable data in which one or more of the functional elements

are missing or compromised in sequence length or quality. Passing reads that share a single barcode and corresponding target site are deduplicated using the UMIs and counted. The read counts corresponding to each target site in the library are then compared to those of the on-target site to generate a score, for example the ratio of off-target reads / on-target reads. The resulting information can be annotated to provide information on genes and other functional elements that overlap each candidate off-target site.

[0106] Together, these new methods address several problems with current detection technologies including low reliability, lack of quantitative accuracy, and high costs. For example, in contrast to ONE-seq, the new methods reduce the required length of synthetic oligonucleotides from approximately 200 to ~100 nucleotides and eliminate the need for non-quantitative sequencing reads from non-target strand 5' ends after cleavage – thereby reducing oligonucleotide synthesis and NGS sequencing costs significantly. The new methods also enable, for the first time, quantitative counting of the number of unique *in vitro* cleavage events using UMIs that are incorporated into each molecule immediately after cleavage and before amplification using index primers.

Exemplary Embodiment

[0107] An exemplary embodiment of the method described herein is shown in FIG. 2. In step 1, a library of single-stranded synthetic oligonucleotides that contain a Barcode (BC) on one side of a target sequence (with optional flanking sequences fl and fr) is amplified using left and right end primers complementary to fixed sequences (F1 and F2) at the ends of the synthetic oligonucleotides. One of the primers (left end) contains an Illumina sequencing primer (S7). Quality control of the amplified library can be performed by sequencing in step 2. In step 3, a gene editing nuclease, such as Cas9, is added to initiate an *in vitro* cleavage reaction. The *in vitro* cleavage reaction results in a set of first cleaved dsDNA oligonucleotide fragments. In step 4, these fragments are end-repaired to produce blunt ends using a proofreading DNA polymerase (step 4) and are subsequently ligated to adaptors containing a UMI and a second Illumina sequencing primer (S5) (step 5). In step 6, the bottom strand of one side (left) is amplified using index primers that contain sequences complementary to the sequencing primers (S5 and S7), index sequences (i5 and i7), and sequences to enable amplification on an Illumina flow cell for sequencing (P5 and P7). The amplified products are then sequenced on a NextSeq 1000/2000 instrument.

SYSTEMS AND KITS

[0108] Also provided herein are systems and kits including the reagents needed for performing the methods described herein as well as written instructions for making and using the same.

[0109] Any of the above-described systems and kits can further include one or more additional reagents.

[0110] In some embodiments, a system or kit can further include instructions for using the components of the kit to practice the methods. The instructions for practicing the methods are generally recorded on a suitable recording medium. For example, the instructions can be printed on a substrate, such as paper or plastic, and the like. The instructions can be present in the kits as a package insert, in the labeling of the container of the kit or components thereof (i.e., associated with the packaging or sub-packaging), and the like. The instructions can be present as an electronic storage data file present on a suitable computer readable storage medium, e.g., CD-ROM, diskette, flash drive, and the like. In some instances, the actual instructions are not present in the kit, but means for obtaining the instructions from a remote source (e.g., via the internet), can be provided. An example of this embodiment is a kit that includes a web address where the instructions can be viewed and/or from which the instructions can be downloaded. As with the instructions, this means for obtaining the instructions can be recorded on a suitable substrate.

[0111] No admission is made that any reference cited herein constitutes prior art. The discussion of the references states what their authors assert, and the Applicant reserves the right to challenge the accuracy and pertinence of the cited documents. It will be clearly understood that, although a number of information sources, including scientific journal articles, patent documents, and textbooks, are referred to herein; this reference does not constitute an admission that any of these documents forms part of the common general knowledge in the art.

[0112] The discussion of the general methods given herein is intended for illustrative purposes only. Other alternative methods and alternatives will be apparent to those of skill in the art upon review of this disclosure and are to be included within the spirit and purview of this application.

[0113] Throughout this specification, various patents, patent applications and other types of publications (e.g., journal articles, electronic database entries, *etc.*) are referenced. The disclosure of all patents, patent applications, and other publications cited herein are hereby incorporated by reference in their entirety for all purpose.

EXAMPLES

[0114] While particular alternatives of the present disclosure have been disclosed, it is to be understood that various modifications and combinations are possible and are contemplated within the true spirit and scope of the appended claims. There is no intention, therefore, of limitations to the exact abstract and disclosure herein presented.

EXAMPLE 1: VARIANT AWARE *IN VITRO* OFF-TARGET PROFILING FOR CAS9 TARGET

PCSK9-1: CCCGCACCTTGGCGCAGCGG/TGG

[0115] Exemplary oligonucleotides for use in the methods are listed in Table 1.

Table 1.

Name	Sequence and modifications
Library 1	CAGTCGACGTCGATTCGTGT[Barcode]TTGACATTCTGCAATTA[10nt flank - target - 10nt flank]AGAGCTGCGAGTCTTACAGC (SEQ ID NO: 1)
Library 2	CAGTCGACGTCGATTCGTGT[barcode][AAGCTT][10nt flank - target - 10nt flank]AGAGCTGCGAGTCTTACAGC (SEQ ID NO: 2)
Oligo-1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGACGTTCTCACAGCAATTCGTACAGTCGACGTCGATTCG*T*G*T (SEQ ID NO: 3)
Oligo-2	GCGTAATCACTGATGCTTCGTATATGAGACATGCAATGCTGTAAGACTCGCAGC*T*C*T (SEQ ID NO: 12)
Oligo-3	T*G*A*GAGAGTGTGATAGC*T*C*A (SEQ ID NO : 4)
Oligo-4	A*C*A*CTCTTCCCTACACGACGCTCTCCGATCT[NNNNNNNNNN]ACTGCTATCACACTCTC*T*C*A (SEQ ID NO: 5)
Oligo-5	AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTCCCTACACGACGCTCTCCGATC*T (SEQ ID NO: 6)
Oligo-6	CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T (SEQ ID NO: 7)

[0116] A set of candidate off-target sites with up to six differences relative to the target site (including mismatches and up to 2 indels) is generated using custom software tools to integrate sites identified using Cas-Designer and a modified version of Calitas. The sites comprise reference sites from the HG38 genome and phased variant sites from the 1000 Genomes (1kG) and Human Genome Diversity Project (HGDP) data sets (4119 genomes). A library of 98,331 sequences is generated, each sequence (Library 1 or Library 2) comprising the target site or a candidate off-target site (with 10 nt genomic flanking regions), a single barcode, and three fixed sequences (F1 and F2 as shown in **FIG. 2**, and a third fixed sequence between the target site and barcode). An oligonucleotide library comprising the designed sequences is synthesized (Agilent

Technologies) and amplified by 12 cycles of PCR using left and right primers (Oligo-1 and Oligo-2, respectively) with Equinox polymerase (Watchmaker Bio). The amplified library is sequenced for quality control and subjected to *in vitro* cleavage in triplicate using complexed Cas9 and PCSK9-1 gRNA at three RNP to DNA ratios (10:1, 1:1 and 0.1:1). The cleaved ends are repaired using Equinox polymerase, and a partially double-stranded universal adaptor (Oligo-3 and Oligo-4) comprising a UMI and fixed sequence (S5) is ligated to the repaired ends. The ligation products from each sample are amplified using a pair of Illumina-compatible dual index primers (for example, Oligo-5 and Oligo-6) and sequenced on a NextSeq 2000 sequencing instrument. The resulting reads are processed using a custom bioinformatics analysis pipeline. The software verifies the presence of intact elements in the cleaved products, including the barcode, fixed sequences, UMI, and candidate off-target sequences, and generates deduplicated read counts for valid cleavage events within a defined window of the expected cleavage site. A cleavage score is calculated for each candidate off-target site (ratio of read counts for each off-target site to the read counts of the on-target site). The data for each candidate off-target site is linked with the corresponding library information, including genomic location(s), phased variant information and population allele frequencies. The resulting table is annotated to include information on overlapping gene sequences, gnomAD gene constraint information, ENCODE candidate cis regulatory elements, cancer gene hits, and the predicted effects of indels at each candidate off-target cut site.

[0117] The above example is exemplary and can be adapted to other CRISPR associated nucleases such as Cas12a, CasX, or to TALEN or Zinc Finger nucleases by substituting the appropriate nuclease and guide RNA combination and performing the *in vitro* cleavage reaction under the appropriate conditions for each enzyme.

EXAMPLE 2: CAS12A SPECIFICITY PROFILING USING A RANDOM VARIANT SCANNING LIBRARY FOR TARGET DNMT1, SITE 3: TTTC/CTGATGGTCCATGTCTGTACTC

[0118] Exemplary oligonucleotides for use in the methods are listed in Table 1 above.

[0119] A set of candidate off-target sites with up to two differences relative to the target site (including mismatches, insertions or deletions in any combination) at all possible positions is generated using a custom software tool. A library of approximately 19,000 sequences is generated, each comprising the target site or a candidate off-target site with fixed 5 nt flanking regions, a single barcode, and a pair of fixed sequences (F1 and F2), as shown in FIG. 2. An

oligonucleotide library comprising the designed sequences is synthesized (Agilent Technologies) and amplified by 12 cycles of PCR using left and right primers (Oligo-1 and Oligo-2, respectively) with Equinox polymerase (Watchmaker Bio). The amplified library is sequenced for quality control and subjected to *in vitro* cleavage in triplicate using complexed Cas9 and DNMT1-site3 gRNA at multiple RNP to DNA ratios (e.g., 10:1, 1:1 and 0.1:1) and/or multiple reaction times (e.g., 10, 30, 60, 90 minutes). The cleaved ends are repaired using Equinox polymerase, and a partially double-stranded universal adaptor (Oligo-3 and Oligo-4) comprising a UMI and fixed sequence (S7) is ligated to the repaired ends. The ligation products from each sample are amplified using a pair of Illumina-compatible dual index primers (for example, Oligo 5 and Oligo 6) and sequenced on a NextSeq 2000 sequencing instrument. The resulting reads are processed using a custom bioinformatics analysis pipeline. The software verifies the presence of intact elements in the cleaved products, including the barcode, fixed sequences, UMI, and candidate off-target sequences, and generates deduplicated read counts for valid cleavage events within a defined window of the expected cleavage site. A cleavage score is calculated for each candidate off-target site (ratio of read counts for each off-target site to the read counts of the on-target site).

EXAMPLE 3: VARIANT AWARE *IN VITRO* OFF-TARGET PROFILING FOR CAS9 TARGET

PCSK9-1: CCCGCACCTTGGCGCAGCGG/TGG

[0120] A set of candidate off-target sites, with up to six differences (mismatches and up to two indels) relative to the target site (PCSK9-1), was generated using custom software tools. The sites were identified by integrating data from Cas-Designer and a modified version of CALITAS, leveraging reference sites from the hg38 reference genome and 3502 genomes with phased variant data from the 1000 Genomes (1kG) and Human Genome Diversity Project (HGDP) datasets. This resulted in 98,331 designed sequences, each comprising the target site, or a candidate off-target site flanked by 10 nt genomic regions, a single barcode, and three fixed sequences.

[0121] The oligonucleotide library (Library 2:

CAGTCGACGTCGATTCGTGT[barcode][AAGCTT][10nt flank - target - 10nt

flank]AGAGCTGCGAGTCTTACAGC (SEQ ID NO: 2)) was synthesized by Agilent

Technologies. Upon receipt, the library was resuspended and subsequently amplified by PCR (11 cycles) using Oligo-1

(GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGACGTTCTCACAGCAATTCGTACAGTCGACGTCGATTCTG*T*G*T (SEQ ID NO: 3), left primer) and Oligo-2 (GCGTAATCACTGATGCTTCGTATATGAGACATGCAATGCTGTAAGACTCGCAGC*T*C*T (SEQ ID NO: 12), right primer) with Equinox polymerase (Watchmaker Genomics). A portion of the amplified library was indexed via PCR amplification (9 cycles) using i7 and i5 primers (Oligo-5, AATGATACGGCGACCAACCGAGATCTACACTATAGCCTACACTCTTCCCTACACGACGCTCTTCCGATC*T (SEQ ID NO: 6) and Oligo-6, CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T (SEQ ID NO: 7)), to enable sequencing alongside the cleaved samples in a single pool.

[0122] The amplified library was then subjected to *in vitro* cleavage reactions in triplicate. Sp Cas9 (New England Biolabs) was complexed with PCSK9 gRNA (Integrated DNA Technologies), and cleavage reactions were conducted at an RNP-to-DNA ratio of 1:1. Following the cleavage reaction, the products were subjected to end repair using Klenow polymerase (New England Biolabs). A pre-annealed, partially double-stranded universal adaptor (comprised of Oligo-3, T*G*A*GAGAGTGTGATAGC*T*C*A (SEQ ID NO: 4), and Oligo-4, A*C*A*CTCTTCCCTACACGACGCTCTTCCGATCT[NNNNNNNNNN]ACTGCTATCACACTCTC*T*C*A (SEQ ID NO: 5)) with a unique molecular identifier (UMI) and fixed sequence (S5) was ligated to the repaired ends using T4 DNA ligase (New England Biolabs).

[0123] The ligation products were amplified by PCR (13 cycles) using Illumina-compatible dual-index primers and sequenced with the indexed pre-cleaved sample on an Illumina NextSeq 2000 sequencing instrument. The sequencing data were processed using a custom bioinformatics pipeline. The pipeline verified the presence of intact barcodes, fixed sequences, UMIs, and partial candidate off-target sequences within the cleaved products. Deduplicated read counts were generated for valid cleavage events occurring within the expected cleavage site window.

[0124] For each candidate off-target site, a cleavage score was calculated as the ratio of read counts for the off-target site to those of the on-target site. The resulting data table included annotations for genomic locations, phased variant information, population allele frequencies, overlapping gene sequences, gnomAD gene constraint data, ENCODE candidate cis-regulatory elements, cancer gene hits, and predicted mismatches and indels at each candidate off-target cut site.

[0125] A summary of results for 20 highest scoring sites is shown below in Table 2. These include reference sites from the hg38 reference genome and variant sites from the 1kG+HGDP genomes in equal numbers. The PCSK9 on-target site received a score of 1, and the various off-target sites received lower scores corresponding to their relative cleavage efficiencies in the assay. Off-target sites cleaved at high frequencies include those located in genes *FGF18*, *ZBTB4*, *RPLP1*, *ASB15*, *PIK3C2G*, and *LncRNA ENSG00000232325*. The *LncRNA ENSG00000232325* site includes a variant with approximately 26% global allele frequency.

Target Sequence	Score	Chr	Start	End	Strand	Gene Name	Variant AF
CCCGCACCTTGGCGCAGCGGTG G (SEQ ID NO: 13)	1.000 0	chr1	55040028	5504005 1	-	PCSK9	
GGCGCACCTTGGCGCAGCGGA GG (SEQ ID NO: 14)	0.692 3	chr5	17145118 7	1.71E+0 8	+	FGF18	
GGTGCACCTTGGCGCAGCGGA GG (SEQ ID NO: 15)	0.690 5	chr5	17145118 7	1.71E+0 8	+	FGF18	0.00024
TGCACACCTTGGCGCAGTGGGG G (SEQ ID NO: 16)	0.476 8	chr1 7	7462658	7462681	+	ZBTB4	
CCCACACCAGGCACAGCGGGG G (SEQ ID NO: 17)	0.346 2	chr7	93101	93123	+	ENSG0000023 2325	0.26324
CGAGCACCTTGGCGCAGCGG AGG (SEQ ID NO: 18)	0.156 6	chr1 5	69452824	6945284 8	-	RPLP1	0.00085
ACCACACCTTGGAGCAACGGAG G (SEQ ID NO: 19)	0.108 4	chr7	12362022 0	1.24E+0 8	+	ASB15-AS1	0.00049
ACCACACCTTGGAGCAACGGAG G (SEQ ID NO: 20)	0.108 4	chr7	12362022 0	1.24E+0 8	+	ASB15	0.00049
TTAGCACCTTGGCCAGCGGCGG (SEQ ID NO: 21)	0.101 5	chr1 2	18500277	1850029 9	+	PIK3C2G	
GCTGCACCTTGGCACAGTGGAG G (SEQ ID NO: 22)	0.093 1	chr1 9	50490617	5049064 0	-	EMC10	
AGGGCACCTTGGAGCAGCGGG GG (SEQ ID NO: 23)	0.061 6	chr1	15348693	1534871 6	-	FHAD1	0.00671
CAGGCACCATGGCAGCAGCGG AGG (SEQ ID NO: 24)	0.058 9	chr2	11063393 2	1.11E+0 8	-	RPL22P12	0.00073
CCCACACCTTGGTGTGAGCGGA GG (SEQ ID NO: 25)	0.045 4	chr3	14242672	1424269 6	-	intergenic	
CAAGCCCCTTGGCACAGCGGAG G (SEQ ID NO: 26)	0.032 0	chr1 4	33749638	3374966 1	+	NPAS3	
GGAGCACCATGACGACGCGAC GG (SEQ ID NO: 27)	0.030 1	chr1 3	25772154	2577217 7	-	ATP8A2	0.00171
CTAACACCTTGGTGCAGAGGTG G (SEQ ID NO: 28)	0.029 7	chr1 6	67748648	6774867 1	-	RANBP10	
CCTGCAGCCTGGCGCAGCGGTG G (SEQ ID NO: 29)	0.028 3	chr1 8	75340345	7534036 8	-	intergenic	0.00024
CGGGCACCATGGCAGCAGCGG AGG (SEQ ID NO: 30)	0.025 0	chr2	10791489 5	1.08E+0 8	-	intergenic	
CGGGCACCATGGCAGCAGCGG AGG (SEQ ID NO: 31)	0.025 0	chr2	10973751 1	1.1E+08	+	RPL22P11	0.01000
CGGGCACCATGGCAGCAGCGG AGG (SEQ ID NO: 32)	0.025 0	chr2	11063393 2	1.11E+0 8	-	RPL22P12	

[0126] The method successfully profiled off-target cleavage events in a variant-aware and quantitative manner, providing detailed annotations for each candidate site, and enabling further insights into Cas9 specificity and potential off-target effects at the selected locus. The results also showed a high correlation ($R^2 = 0.8$) with standard ONE-seq data as shown in **FIG. 3**.

EXAMPLE 4: CAS12A SPECIFICITY PROFILING USING A RANDOM VARIANT SCANNING LIBRARY
DNMT1-SITE 3: TTTC/CTGATGGTCCATGTCTGTACTC

[0127] A set of candidate off-target sites with up to two differences relative to the target site (DNMT1-3, including mismatches, insertions or deletions in any combination) at all possible positions was generated using a custom software tool. A library of 13,546 sequences was generated, comprising the on-target site and a large number of variant off-target sites with fixed flanking regions, a single barcode, and three fixed sequences.

[0128] The oligonucleotide library (Library 2: CAGTCGACGTCGATTCGTGT[barcode][AAGCTT][10nt flank - target - 10nt flank]AGAGCTGCGAGTCTTACAGC (SEQ ID NO: 2) was designed and synthesized by Agilent Technologies. Upon receipt, the library was resuspended and subsequently amplified by PCR (11 cycles) using Oligo-1 (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGACGTTCTCACAGCAATTCGTA CAGTCGACGTCGATTCG*T*G*T (SEQ ID NO: 3), left primer) and Oligo-2 (GCGTAATCACTGATGCTTCGTATATGAGACATGCAATGCTGTAAGACTCGCAGC*T* C*T (SEQ ID NO: 12), right primer) with Equinox polymerase (Watchmaker Genomics). A portion of the amplified library was indexed via PCR amplification (9 cycles) using i7 and i5 primers (Oligo-5,

AATGATACGGCGACACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGAC GCTCTTCCGATC*T (SEQ ID NO: 6) and Oligo-6,

CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCT CTTCCGATC*T (SEQ ID NO: 7)), to enable sequencing with the cleaved samples.

[0129] The amplified library was then subjected to *in vitro* cleavage reactions in triplicate. Lb Cas12a (New England Biolabs) was complexed with DNMT1-3 gRNA (Integrated DNA Technologies), and cleavage reactions were conducted at an RNP-to-DNA ratio of 1:1. Following cleavage, the reaction products underwent end repair using Klenow polymerase (New England Biolabs). A pre-annealed, partially double-stranded universal adaptor (comprised of Oligo-3, T*G*A*GAGAGTGTGATAGC*T*C*A (SEQ ID NO: 4), and Oligo-4,

A*C*A*CTCTTCCCTACACGACGCTCTCCGATCT[NNNNNNNNNN]ACTGCTATCAC
ACTCTC*T*C*A (SEQ ID NO: 5)) with a unique molecular identifier (UMI) and fixed

sequence (S5) was ligated to the repaired ends using T4 DNA ligase (New England Biolabs).

[0130] The ligation products were amplified by PCR (13 cycles) using Illumina-compatible dual-index primers and sequenced with the indexed pre-cleaved sample on an Illumina NextSeq 2000 sequencing instrument. The sequencing data were processed using a custom bioinformatics pipeline. The pipeline verified the presence of intact barcodes, fixed sequences, UMIs, and partial candidate off-target sequences within the cleaved products. Deduplicated read counts were generated for valid cleavage events occurring within the expected cleavage site window.

[0131] For each candidate off-target site, a cleavage score was calculated as the ratio of read counts for the off-target site to those of the on-target site. The resulting data table included annotations for genomic locations, and the positions of mismatches, insertions and deletions at each candidate off-target cut site.

[0132] A graphical summary of the effect of single mismatches (circles), insertions (squares) and deletions (triangles) at each position of the target site on the score relative to the on-target site with zero differences is shown in **FIG. 4**. The method successfully profiled off-target cleavage frequency in a systematic manner based on the variant position within the target site, enabling further insights into Lb Cas12a specificity and potential off-target effects.

WHAT IS CLAIMED IS:

1. A method of quantifying off-target cleavage by a gene editing nuclease comprising:
 - a) performing *in vitro* cleavage of a library of dsDNA oligonucleotides with a gene editing nuclease such that a set of first dsDNA oligonucleotide fragments is formed, each said first dsDNA oligonucleotide fragment comprising a pre-defined barcode uniquely associated with one off-target sequence in the library, and a first fixed sequence present in every oligonucleotide in the library;
 - b) end-repairing the cleaved ends of the first oligonucleotide fragments;
 - c) ligating the cleaved ends to an adaptor, wherein the adaptor comprises a UMI;
 - d) performing an amplification reaction to amplify the set of first dsDNA oligonucleotide fragments using (i) a first primer comprising a nucleotide sequence complementary to a region of the adaptor and (ii) a second primer comprising a nucleotide sequence complementary to the first fixed sequence;
 - e) sequencing the amplification products, and quantifying the reads derived from each cleaved off-target sequence in the library.
2. The method of claim 1, wherein the library contains sequences derived from a reference genome, sequences derived from more than one genome, or sequences corresponding to any defined set of variants with respect to an on-target sequence.
3. The method of claim 1, wherein the number of sequences in the library is between 10 and one billion.
4. The method of any preceding claim, wherein the first dsDNA oligonucleotide fragment comprises part of a PAM.
5. The method of any one of claims 1 to 3, wherein the first dsDNA oligonucleotide fragment comprises part of a PS.

6. The method of any preceding claim, wherein the gene editing nuclease is a Cas nuclease.
7. The method of claim 6, wherein the Cas nuclease is Cas9 or Cas12a.
8. The method of any preceding claim, wherein the library of dsDNA oligonucleotides comprises a UMI.
9. The method of any preceding claim, wherein sequencing comprises determining the sequence of the barcode and counting the reads, thereby quantifying off-target dsDNA sequences that are cleaved by a gene editing nuclease.
10. The method of any preceding claim, wherein the library is formed by synthesizing single-stranded DNA sequences, optionally on high-density oligonucleotide arrays, and performing a first amplification reaction to generate double stranded DNA (dsDNA) oligonucleotides.
11. The method of claim 11, wherein the first amplification reaction comprises limited cycle PCR.
12. The method of any preceding claim, wherein end-repairing comprises blunt ending and/or A tailing.