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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 June 2025 (19.06.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/128693 A1

(51) International Patent Classification:
C12N 9/22 (2006.01) *C12N 15/63* (2006.01)
C12Q 1/6806 (2018.01) *C12Q 1/6874* (2018.01)
C12N 15/10 (2006.01) *C40B 40/06* (2006.01)
C12N 15/55 (2006.01)

(21) International Application Number:
PCT/US2024/059554

(22) International Filing Date:
11 December 2024 (11.12.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/608,550 11 December 2023 (11.12.2023) US

(71) Applicant: **THE ROCKEFELLER UNIVERSITY** [US/US]; 1230 York Avenue, New York, NY 10065 (US).

(72) Inventors: **SOROCZYNSKI, Jan**; 1230 York Avenue, New York, NY 10065 (US). **RISCA, Viviana**; 1230 York Avenue, New York, NY 10065 (US).

(74) Agent: **DIETZEL, Christine, E.** et al.; Hoffmann & Baron, LLP, 6900 Jericho Turnpike, Syosset, NY 11791 (US).

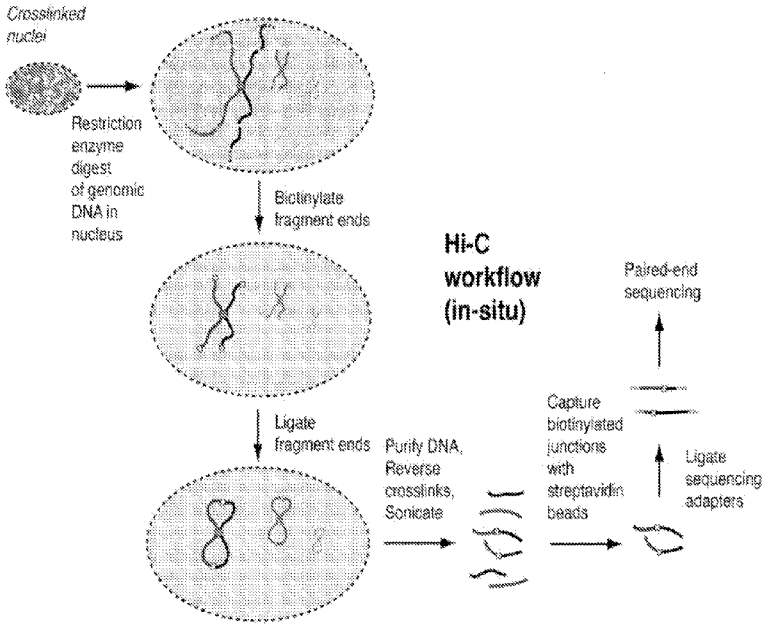
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ,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

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

WO 2025/128693 A1

(54) Title: AN ENGINEERED MAMMALIAN NUCLEASE FOR IMPROVED MAPPING OF 3-D GENOME ARCHITECTURE FROM NUCLEOSOME TO CHROMOSOME SCALE

FIGURE 1



(57) Abstract: The present invention provides engineered mammalian nucleases, particularly engineered ICAD and CAD enzymes, which are altered to be activated by a specific protease. The invention provides methods and kits for mapping of genome architecture and evaluation of chromosome conformation for determining high-resolution three-dimensional (3D) and 4 dimensional (space and time) genome features, structures and sequences, particularly utilizing the engineered nucleases, particularly ICAD and CAD.

[Continued on next page]

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

Published:

— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

AN ENGINEERED MAMMALIAN NUCLEASE FOR IMPROVED MAPPING OF 3-D GENOME ARCHITECTURE FROM NUCLEOSOME TO CHROMOSOME SCALE

STATEMENT OF GOVERNMENT RIGHTS

[0001] This invention was made with government support under numbers DP2GM150021 awarded by the National Institutes of Health. The government has certain rights in the invention.

FIELD OF THE INVENTION

[0002] The present invention and disclosure relates to engineered mammalian nucleases and methods for mapping of genome architecture and evaluation of chromosome conformation for determining high-resolution three-dimensional (3D) and 4 dimensional (space and time) genome features, structures and sequences.

SEQUENCE LISTING

[0003] A Sequence Listing conforming to the rules of WIPO Standard ST.26 is hereby incorporated by reference. The Sequence Listing has been filed as an electronic document via PatentCenter encoded as XML in UTF-8 text. The electronic document, created on December 8, 2023, is entitled "1119-81-P_ST26.xml", and is 38,826 bytes in size.

BACKGROUND OF THE INVENTION

[0004] Two decades of research have uncovered that the three-dimensional folding of chromosomes contributes significantly to gene regulation. Its dysregulation has been observed in both cancers and developmental disorders. In the last decade, technological advancements in genomics have provided appreciation and recognition of the existence of intricate 3D organization of interphase chromosomes (Mirny & Dekker, CSHL Perspect. Biol., 2022; Soroczynski & Risca, Curr Opin Cell Biol., 2023). We now know that the architecture of chromosome folding plays a major role in human gene regulation, and has been shown to be misregulated in multiple cancer types as well as in developmental disorders (Tettey et al, Nat Rev Cancer, 2023; Fujita et al, Neuron, 2022). Specifically, the development of chromosome conformation capture (3C) methods with high-throughput DNA sequencing has enabled the whole genome study of chromosome architecture. The common such method is referred to as Hi-C (Dekker et al., Science, 2002; Lieberman-Aiden et al., Science, 2009). Three dimensional (3D) conformation is interpreted via two dimensional (2D) contact maps, which provide hierarchical structural patterns across genomic scales. Pair-

wise proximity ligation measures how frequently two loci come close enough to ligate and this is used to reconstruct the dimensional organization and 'who sees who how often'.

[0005] Chromosome conformation capture (3C) methods rely on the principle of proximity ligation, where cells are chemically fixed, DNA is cut and the DNA ends in 3D proximity within the nucleus are ligated. A depiction of the steps in proximity ligation is provided in FIGURE 1. Typically, prior to ligation the ends are labeled, such as with biotin, which subsequently allows for the ligation junctions to be enriched by capture via the label, such as by streptavidin capture. Hundreds of millions of the chimeric DNA molecules corresponding to the ligation junctions are then sequenced, and used to reconstruct pairs of spatially-proximal DNA interactions in the form of a 2D contact matrix heatmap. Data from Hi-C and its variant methods capture a very wide dynamic range, from local oligo-nucleosomal interactions to multi-megabase whole-chromosome interactions, and even inter-chromosomal contacts can be measured. When obtained at high resolution, these data sets reveal tissue-specific and even individual sample-specific contacts between regulatory regions and genes that can provide insights into the function and targets of those regulatory regions or into the aberrant functions of their genetic variants in regulating gene expression.

[0006] 3C methods such as Hi-C have applications in basic biology, including de novo genome assembly, as well as emerging roles in translational and clinical research such as understanding the impact of sequence variants on gene regulation and disease. Hi-C and derivative kits as well as services, are commercially available (Arima Genomics and Dovetail Genomics (recently acquired by Cantata Bio)). In the traditional Hi-C method, restriction enzymes are utilized to cut DNA at defined sequence motifs and these are easy to ligate together. Hi-C is useful for studying long-range interactions from tens of thousands of bases and above. But these restriction sites are non-homogeneously distributed along the chromosomes, leading to irregular resolution of the contact maps. The resolution of Hi-C is limited by the restriction enzymes used (Oksuz BA et al (2021) Nature Meth 18:1046-1055; doi.org/10.1038/s41592-021-01248-7).

[0007] To gain high resolution contact maps, a variant of the Hi-C was developed, Micro-C, which uses micrococcal nuclease (MNase) instead of restriction enzymes (Hsieh et al., Cell, 2015; Hsieh et al., Nature Methods, 2016; Hsieh et al., Molecular Cell, 2020, Krietenstein et al., Molecular Cell, 2020). MNase can digest chromatin to its nucleosome units and obtain higher resolution. This comes at the expense of assay sensitivity and robustness because MNase digests chromosomes unevenly and over-digests some loci due to its bias towards more accessible and AT-rich DNA and exonuclease activity. It also generates poorly ligatable DNA ends. Micro-C therefore requires arduous end-user optimization and is unsuitable for precious clinical samples or rare cell types. MNase is an aggressive nuclease which attacks DNA in complex ways, first making double strand cuts, then chewing and nicking the DNA. These ends are not ligatable, so they have to be repaired first in order for proximity ligation to be performed. If the

reaction is not halted, MNase will continue to digest all of the DNA, regardless of whether proteins are bound or not. Because of this the experimenter must carefully titrate the MNase dose and reaction time and titration studies are often necessary in order to achieve accurate and reliable results. In addition, MNase only cleaves upstream of A or T (Hortz & Altenburger, NAR, 1981), so regardless of reaction condition the chromosomes are never uniformly fragmented. Taken together high resolution maps using MNase and Micro-C methodolodies suffer from biases and miss the long range contacts that Hi-C detects. The inefficiencies of Micro-C make it laborious, expensive to perform, and only applicable to highly abundant cell samples that can be used for titrations (Slobodyanyuk et al., Methods Mol Biol, 2022). Crucially, clinically important rare cell types and precious, limited samples are difficult to study using Micro-C because of this titration limitation.

[0008] In view of the deficiencies and problems with prior art methods, we sought to develop a method that would allow one to study chromosome architecture at all scales in as many sample types as possible, including rare cell types and precious samples with limited or lower cell numbers. We identified chromatin fragmentation as the primary bottleneck in the 3C methods and provide a new and useful approach to chromosome fragmentation with a controllable, activatable and effective enzyme lacking exonuclease activity and sequence specificity requirements.

SUMMARY OF THE INVENTION

[0009] The invention relates generally to the determination of the organizational structure of chromatin and chromosomes in three dimensions, including whereby information as to the structure and sequence is determined and provided. We sought to develop a method that would allow one to study chromosome architecture at all scales in as many sample types as possible, including rare cell types and precious samples with limited or lower cell numbers. We identified the mode of chromatin fragmentation, which determines the ends available for ligation, as the primary bottleneck in the efficiency of 3C methods. We chose to turn to a mammalian Caspase Activated DNase or CAD, which evolved to uniformly fragment chromosomes during the onset of apoptosis. CAD is also known as 40-kDa DNA Fragmentation Factor (DFF40). CAD has the unique property of being a strict endonuclease, and therefore retains the easy end point digestion property of restriction enzymes (as in Hi-C) while keeping the high resolution of MNase, with lesser DNA sequence bias. To overcome the limitations of Micro-C, we developed a novel recombinant nuclease based on Caspase-Activated DNase (CAD). Unlike MNase, CAD has no substantial exonuclease activity and produces blunt-ended, ligatable DNA ends with 5'-phosphates (Widlak P et al (2000) J Biol Chem 275(11):8226-32). Further, CAD does not appear to overdigest chromatin over a wide range of concentrations. These unique properties of CAD improve the prior art approaches and methods, including Hi-C and Micro-C, and their resolution without sacrificing sensitivity.

[00010] CAD is ordinarily activated by caspase cleavage of its inhibitor ICAD (inhibitor of CAD, also known as DFF45) (Widlak P and Garrard WT (2005) J Cellular Biochem 94(6):1078-87). In accordance with the invention, ICAD has been re-engineered to be activated by a an alternative enzyme/protease, common recombinant TEV protease, to allow easy CAD activation *in vitro*. Utilizing re-engineered ICAD, we have applied CAD to the chromatin fragmentation step of a proximity ligation protocol with consistently good and useful results as provided herein. This novel method has been provisionally named CAD-C. We have designed and optimized a murine CAD/I-CAD for scalable bacterial expression, which can efficiently digest human chromatin in a TEVp-controlled manner. Our data utilizing a Hi-C type methodology with CAD enzyme (CAD-C) demonstrate CAD-C's unique ability to simultaneously detect short-range (< 1 kb) and long-range (>10 Mb) chromosome contacts, which with the current state of the art, would require separate, costly Micro-C and Hi-C experiments.

[00011] In accordance with the invention, a controllably activated Caspase Activated DNase or CAD is provided, wherein the CAD is activated by an enzyme other than caspase, including other than caspase-3. In an embodiment, the controllably activated Caspase Activated DNase or CAD is provided, wherein the CAD is activated by protease digestion of the CAD inhibitor ICAD. In an embodiment, the controllably activated Caspase Activated DNase (CAD) is provided, wherein the CAD is activated by protease digestion of the CAD inhibitor ICAD (inhibitor of caspase-activated DNase (ICAD) also known as DNA fragmentation factor 45 kDa (DFF45)) with a protease with high sequence specificity, wherein the protease is not caspase and wherein the protease recognized sequence is replaced for caspase recognized amino acid sequence in ICAD. In an embodiment, the controllably activated Caspase Activated DNase or CAD is provided, wherein the CAD is activated by protease digestion of the CAD inhibitor ICAD with TEV protease.

[00012] In embodiments of the invention an engineered ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, engineered ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence whereby the engineered ICAD is cleaved by a specific alternative protease and is not cleaved by caspase. In an embodiment, engineered ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, caspase recognition sequence is replaced with TEV protease recognition sequence. In an embodiment, caspase recognition sequence selected from DEP, DET and/or DAV is deleted and is replaced with a protease recognition sequence for an alternative protease. Alternative proteases may be selected from known sequence specific proteases, particularly those having short recognition sequences of between three amino acids and ten amino acids. Alternative proteases may be selected from known sequence specific proteases, whereby replacement of the alternative protease

sequence in the ICAD generates an engineered ICAD, wherein the engineered ICAD is capable of forming a heterodimer with CAD thereby inactivating CAD enzyme and wherein the engineered ICAD is specifically cleaved by the alternative protease. In an embodiment, the engineered ICAD is specifically cleaved by the alternative protease and is not cleaved by caspase.

[00013] In an embodiment, caspase recognition sequence selected from DEP, DET and/or DAV or an equivalent corresponding caspase recognition sequence is replaced with alternative protease recognition sequence. In an embodiment, caspase recognition sequence selected from DEP, DET and/or DAV or an equivalent corresponding caspase recognition sequence is replaced with TEV protease recognition sequence. In an embodiment, caspase recognition sequence selected from DEP, DET and/or DAV is replaced with TEV protease recognition sequence ENLYFQX, wherein X is S, G, A, M, C or H (SEQ ID NO:10). In an embodiment, caspase recognition sequence selected from DEP, DET and/or DAV is replaced with TEV protease recognition sequence ENLYFQX, wherein X is S or G (SEQ ID NO:9). In an embodiment, caspase recognition sequence selected from DEP, DET and/or DAV is replaced with TEV protease recognition sequence ENLYFQS (SEQ ID N:8) or ENLYFQG (SEQ ID NO:9). In a preferred embodiment, caspase recognition sequence selected from DEP, DET and/or DAV is replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

[00014] In an embodiment, engineered or altered murine or mouse ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, engineered or altered murine ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence whereby the engineered murine ICAD is cleaved by a specific alternative protease and is not cleaved by caspase. In an embodiment, engineered or altered murine or mouse ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, engineered or altered murine or mouse ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence. In an embodiment, engineered or altered murine or mouse ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8) or ENLYFQG (SEQ ID NO:9). In an embodiment, engineered or altered murine or mouse ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

[00015] In an embodiment, engineered or altered human ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, engineered or altered human ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence whereby the engineered

human ICAD is cleaved by a specific alternative protease and is not cleaved by caspase. In an embodiment, engineered or altered human ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, engineered or altered human ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence. In an embodiment, engineered or altered human ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8) or ENLYFQG (SEQ ID NO:9). In an embodiment, engineered or altered human ICAD is provided wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

[00016] In an embodiment, ICAD sequence is engineered to insert the TEV protease recognition sequence of ENLYFQS to replace caspase recognition amino acids. In an embodiment, the TEV protease recognition sequence of ENLYFQS is inserted to replace a minimal number of caspase recognition amino acids. In an embodiment, the TEV protease recognition sequence of ENLYFQS is inserted to replace three caspase recognition amino acids. In an embodiment, the ENLYFQS sequence is inserted at both caspase recognition locations and replaces DEP in one location and replaces DAV in the other in the mouse ICAD sequence. An exemplary sequence variant is shown in FIGURE 4 and provided in SEQ ID NO:1.

[00017] An embodiment providing an engineered murine ICAD sequence susceptible to specific and controlled inactivation and cleavage by TEV protease ($ICAD^{2xTEV}$) amino acid sequence is as follows (SEQ ID NO:1) (The amino acids added to provide TEV protease cleavage, specifically ENLYFQS are underlined):

MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPIT
LVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWIYNDSDGGTAWVSQESFEA<u>ENLYFQS</u>DSRA
GVKWKNVARQLKEDLSSIILLSEEDLQALIDIPCAELAQELCQSCATVQGLQSTLQQVLD
QREEARQSKQLLELYLQALEKEGNILSNQKESKAALSEEL<u>ENLYFQS</u>DTGVGREMASEVLLRSQ
ILTTLKEKPAPELSLSSQDLESVSKEDPKALAVALSWDIRKAETVQQACTTELALRLQQV
QSLHSLRNLSARRSPLPGEPQRPKRAKRDSS

[00018] In an embodiment, murine ICAD sequence (SEQ ID NO:3) is engineered to delete three amino acids that are caspase recognition sequences. The murine ICAD sequence with the amino acids deleted to eliminate caspase-3 proteolysis underlined and in BOLD is shown below:

MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPIT
LVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWIYNDSDGGTAWVSQESFEA<u>**DEP**</u>DSRA
GVKWKNVARQLKEDLSSIILLSEEDLQALIDIPCAELAQELCQSCATVQGLQSTLQQVLD
QREEARQSKQLLELYLQALEKEGNILSNQKESKAALSEEL<u>**DAV**</u>DTGVGREMASEVLLRSQ
ILTTLKEKPAPELSLSSQDLESVSKEDPKALAVALSWDIRKAETVQQACTTELALRLQQV
QSLHSLRNLSARRSPLPGEPQRPKRAKRDSS

[00019] An embodiment providing an engineered human ICAD sequence susceptible to specific and controlled inactivation and cleavage by TEV protease ($hICAD^{2xTEV}$) amino acid sequence is as follows (SEQ ID NO:2) (The amino acids added to provide TEV protease cleavage, specifically ENLYFQS are underlined):

MEVTGDAGVPESGEIRTLKPCLLRRNYSREQHGVAASCLEDLRSKACDILAIDKSLTPVT
LVLAEDGTIVDDDDYFLCLPSNTKFVALASNEKWAYNNSDGGTAWISQESFDVENLYFQSDSGA
GLKWKNVARQLKEDLSSIILLSEEDLQMLVDAPCSDLAQELRQSCATVQRLQHTLQQVLD
QREEVRQSKQLLQLYLQALEKEGSLLSKQEESKAAFGEEVENLYFQSDTGISRETSSDVALASH
ILTALREKQAPELSLSSQDLELVTKEDPKALAVALNWDIKKTETVQEACERELALRLQQT
QSLHSLRSISASKASPPGDLQNPKRARQDPT

[00020] In an embodiment, human ICAD sequence (SEQ ID NO:4) is engineered to delete three amino acids that are caspase recognition sequences. Comparable caspase recognition sequences in human ICAD for deletion to eliminate caspase cleavage and replacement with an alternative enzyme recognition sequence are amino acids DET and DAV. The human ICAD sequence with the amino acids to be deleted to eliminate caspase-3 proteolysis underlined and in BOLD is shown below:

MEVTGDAGVPESGEIRTLKPCLLRRNYSREQHGVAASCLEDLRSKACDILAIDKSLTPVT
LVLAEDGTIVDDDDYFLCLPSNTKFVALASNEKWAYNNSDGGTAWISQESFDV**DET**DSGA
GLKWKNVARQLKEDLSSIILLSEEDLQMLVDAPCSDLAQELRQSCATVQRLQHTLQQVLD
QREEVRQSKQLLQLYLQALEKEGSLLSKQEESKAAFGEEV**DAV**DTGISRETSSDVALASH
ILTALREKQAPELSLSSQDLELVTKEDPKALAVALNWDIKKTETVQEACERELALRLQQT
QSLHSLRSISASKASPPGDLQNPKRARQDPT

[00021] In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2, or variants thereof retaining the TEV protease recognition sequence and having up to 90% amino acid sequence identity with SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1, or variants thereof retaining the TEV protease recognition sequence and having up to 90% amino acid sequence identity with SEQ ID NO:1. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2, or variants thereof retaining the TEV protease recognition sequence and having up to 90% amino acid sequence identity with SEQ ID NO:1 or SEQ ID NO:2 and retaining the TEV protease recognition sequence SEQ ID NO: 8, SEQ ID NO:9 or SEQ ID NO:10.

[00022] In an embodiment, CAD is activated by inactivation and protease cleavage of the instant engineered ICAD with alternative protease. In an embodiment, the alternative protease is TEV protease. The TEV protease may be a wildtype/native TEV protease or may be a variant TEV protease. Native TEV protease or variant TEV protease and sequences thereof are contemplated, provided they are active and capable of cleaving at the relevant inserted TEVp amino acid sequence, such as any TEVp recognition sequences inserted in an engineered ICAD of the invention. Altered versions and variants of TEVp have been generated and are known and available to one skilled in the art, including S219V mutation, TEV-S153N, TEVp variants with multiple mutations to improve solubility and reduce self-cleavage thereby

enhancing enzymatic activity, including S219N and S219V mutants in the background of T17S, N68D and I77V mutations. In an embodiment, TEV protease is as available commercially such as through New England Biolabs (NEB).

[00023] In one or more embodiments, CAD enzyme is utilized in the methods of the invention and is combined with engineered ICAD. In an embodiment, the altered or engineered ICAD is expressed or otherwise combined with CAD. Sequences of various CAD enzymes are known and available to one skilled in the art and these could be utilized in the invention. An exemplary mouse/murine CAD amino acid sequence is as follows (SEQ ID NO:5):

[00024] Mouse CAD DFF-40 (DFFA) >sp|O54788|DFFB_MOUSE DNA fragmentation factor subunit beta OS=Mus musculus
MCAVLRQPKCVKLRALHSACKFGVAARSCQELLRKGCVRFQLPMPGSRLCLYEDGTEVTD
DCFPGLPNDAELLLLTAGETWHGYVSDITRFLSVFNEPHAGVIQAARQLLSDEQAPLRQK
LLADLLHHVSQNITAETREQDPSWFEGLESRFRNKSGYLRYSCESRIRGYLREVSAYTSM
VDEAAQEEYLRVLGSMCQKLKSVQYNGSYFDRGAEASSRLCTPEGWFSCQGPFDLESCLS
KHSINPYGNRESRILFSTWNLDHIIEKKRTVVPTLAEAIQDGREVNWEYFYSLLFTAENL
KLVHIACHKKTTHKLECDRSRIYRPQTGSRRKQPARKKRPARKR

[00025] An exemplary human CAD amino acid sequence is as follows (SEQ ID NO:6):

[00026] Human CAD (DFF-40)DFFB >sp|O76075|DFFB_HUMAN DNA fragmentation factor subunit beta OS=Homo sapiens
MLQKPKSVKLRALRSPRKFGVAGRSCQEVLRKGCLRFQLPERGSRLCLYEDGTELTEDYF
PSVPDNAELVLLTLGQAWQGYVSDIRRFLSAFHEPQVGLIQAAQQLLCDEQAPQRQRLLA
DLLHNVSQNIAAETRAEDPPWFEGLESRFQSKSGYLRYSCESRIRSYLREVSSYPSTVGA
EAQEEFLRVLGSMCQRLRSMQYNGSYFDRGAKGGSRLCTPEGWFSCQGPFDMDSCLSRHS
INPYSNRESRILFSTWNLDHIIEKKRTIIPTLVEAIKEQDGREVDWEYFYGLLFTSENLK
LVHIVCHKKTTHKLNCDPSRIYKPQTRLKRKQPVRKRQ

[00027] In an embodiment, engineered ICAD and CAD are combined at about equal amounts to form an inactive ICAD/CAD heterodimer. In an embodiment, engineered ICAD of SEQ ID NO: 1 or 2 is expressed or otherwise combined with CAD. In an embodiment, engineered ICAD of SEQ ID NO: 1 or 2 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5 or 6. In an embodiment of the invention, human and murine CAD and ICAD are utilized interchangeably. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5 or 6. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 6. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5 or 6. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD of sequence SEQ ID

NO: 5. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 6.

[00028] In another embodiment of the invention mutant or variant CAD enzyme may be utilized, provided it is active enzymatically. In an embodiment of the invention mutant or variant CAD enzyme may be utilized, provided it is active enzymatically and capable of forming a heterodimer with ICAD, particularly with the engineered ICAD hereof. Therefore, the CAD enzyme sequence, such as that of SEQ ID NO: 5 or SEQ ID NO:6 may be a variant sequence, particularly a variant sequence having 80%, 85%, 89%, 90%, 95%, 97% or 99% amino acid sequence identity to SEQ ID NO:5 or SEQ ID NO:6. If an alternative animal or mammalian CAD sequence as known, for example a native rat or other mammalian ICAD sequence, is utilized, variants thereof are also contemplated, including variant sequence having 80%, 85%, 89%, 90%, 95%, 97% or 99% amino acid sequence identity to the alternative animal or mammalian CAD sequence.

[00029] In embodiments of the invention, nucleic acids encoding the engineered ICAD and CAD are provided. In an embodiment, the encoding nucleic acid is codon optimized for expression, for example in a particular cell type, for example in bacteria such as E.coli.

[00030] Vectors or plasmids comprising the encoding nucleic acid are also provided as embodiments. In an embodiment, the altered or engineered ICAD is co-expressed with CAD in the same vector or plasmid. In an embodiment, a vector is provided in Figure 5. In an embodiments, vector comprising sequence set out in SEQ ID NO:14 is provided. In an embodiment, the altered or engineered ICAD is co-expressed with CAD via the same promoter. Suitable promoters will provide ample expression and approximately equal expression of ICAD and of CAD. An exemplary promoter includes an inducible promoter. In an embodiment, an IPTG-inducible promoter is utilized.

[00031] In particular aspects, the ICAD and/or the CAD of the invention and of use herein is tagged or otherwise includes a label or other means for purification, such as by a protein or molecule having affinity for or otherwise recognizing the tag or label. In an embodiment, a label or tag may include strepdavidin, biotin, GFP etc. In an embodiment, one or more linker, such as between the tag or label, may be added and included. In an embodiment, a recognition sequence for cleavage to remove the one or more tag or labels may also be included. This provides for cleavage and specific separation of the ICAD or CAD, such as of the one or more tag or labels, after initial expression so that a purified ICAD and/or CAD is provided free of alternative or added N terminal or C terminal sequence.

[00032] A purified and isolated engineered ICAD and CAD are provided herein. These may be provided as heterodimers for further activation of CAD by the alternative protease. Compositions of the isolated engineered ICAD, of CAD, and of engineered ICAD/CAD heterodimers are provided herein

[00033] In accordance with the invention, CAD-C methods are provided and utilized for determination of organizational structure of chromatin and chromosomes in three dimensions, including whereby information as to the structure and sequence is determined. In accordance with the invention, kits including components for CAD-C methods are provided for determination of organizational structure of chromatin and chromosomes in three dimensions, including whereby information as to the structure and sequence is determined.

[00034] A method is provided for capturing the sequence and structure (three dimensional conformation) of one or more target regions of one or more genome in a sample comprising:

(a) crosslinking chromatin from a sample to preserve the genome sequence and structure;

(b) introducing a CAD enzyme and digesting the crosslinked chromatin with CAD to generate digested ends of DNA, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase;

(c) labeling the digested ends of DNA;

(d) ligating the digested and labeled ends of DNA to generate ligated DNA; and

(e) purifying the ligated DNA to generate proximally-ligated DNA.

[00035] In an embodiment, the step (c) labeling step is omitted. In accordance with this embodiment, a method is provided for capturing the sequence and structure (three dimensional conformation) of one or more target regions of one or more genome in a sample comprising:

(a) crosslinking chromatin from a sample to preserve the genome sequence and structure;

(b) introducing a CAD enzyme and digesting the crosslinked chromatin with CAD to generate digested ends of DNA, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase;

(d) ligating the digested and labeled ends of DNA to generate ligated DNA; and

(e) purifying the ligated DNA to generate proximally-ligated DNA.

[00036] In an embodiment, the method further comprises:

(f) fragmenting the proximally-ligated DNA; and

(g) capturing the proximally-ligated DNA.

[00037] In accordance with the method provided herein, the sequence and structure (three dimensional conformation) of one or more target regions of one or more genome in a sample can be evaluated or determined via CAD enzyme without any requirement for end repair or labeling of the CAD-digested end. In embodiments of the method, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase

enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8) or ENLYFQG (SEQ ID NO:9). In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

[00038] In some embodiments, the ICAD is an engineered murine ICAD or is an engineered human ICAD. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2, or variants thereof retaining the TEV protease recognition sequence and having up to 90% amino acid sequence identity with SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1.

[00039] In some embodiments, the CAD is a murine or human CAD. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6, or variants thereof retaining CAD endonuclease activity and having up to 90% amino acid sequence identity with SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5.

[00040] In an embodiment of the method, in step (g) the capturing is achieved by enriching for proximally-ligated DNA via the label of (c).

[00041] In an embodiment of the method, in step (g) the capturing is achieved by size selection of the fragments.

[00042] In other embodiments, the proximally-ligated DNA is sequenced. In other embodiments, the proximally-ligated DNA is size selected.

[00043] In an embodiment, the crosslinking of step (a) utilizes UV. In an embodiment, the crosslinking of step (a) utilizes formaldehyde.

[00044] In an embodiment, the purified or captured proximally-ligated DNA is used to prepare a library of the proximally-ligated DNA. In an embodiment, the library of the proximally-ligated DNA is sequenced in its entirety or in part. In an embodiment, the library of the proximally-ligated DNA is sequenced to assess the library's quality in terms of library complexity and long-range interaction information prior to proceeding with the capture.

[00045] In some embodiments, the purified or captured proximally-ligated DNA is enriched using one or more probe directed against or specific for one or more target or sequence of interest.

[00046] In some embodiments of the method, the fragmentation of (f) uses mechanical methods. In one embodiment, the mechanical means is sonication. In other embodiments, fragmentation is carried out via enzyme fragmentation. In one embodiment, fragmentase is used for fragmentation. In an embodiment, the fragmentase NEBNext UltraExpress FS kit (E3340S) is utilized.

[00047] In some embodiments of the method, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of <1 kb. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of 500-800bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of 500-600bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of about 500bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of 300-600 bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of a fragment size of greater than 300bp and less than 1000 bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size >400bp.

[00048] In embodiments, the method is utilized for preparation of one or more libraries of ligated fragments. In an embodiment, di-nucleosomal fragment libraries are generated by size selection and the pairwise contacts between nucleosomes are assembled in contact matrices. Classes of loops (tens of kb) and topologically associating domains (hundreds of kb) within a sample genome can be identified and characterized.

[00049] In embodiments, fragment libraries with a range of fragment sizes are sequenced and mapped to generate pairwise contact matrices. Classes of loops (tens of kb) and topologically associating domains (hundreds of kb) within a sample genome can be identified and characterized.

[00050] In an embodiment, multiple (more than two) consecutively ligated fragments are kept intact and the library is size-selected for concatamers of > 1 kb total length or more, and sequenced with long-read methods. The fragments within each read are aligned to the genome to determine the order of the multi-step ligation "walk" among chromosome loci.

[00051] In an embodiment, the library is fragmented using an enzymatic method to yield fragments of 400-500 bp and sequenced with paired-end sequencing. The fragments within each pair of reads are aligned to the genome to determine the order of the three-step ligation "walk" among chromosome loci.

[00052] In an embodiment, genome-aligned ligated fragments are produced and can be identified. Consecutive steps between such genome-aligned ligated fragments can be identified and analyzed to represent consecutive or adjacent three-dimensional chromatin contacts, termed herein as CADwalks.

[00053] In some embodiments, the purifying of (e) and/or the enriching of (g) is via beads. In some embodiments, the purifying of (e) and/or the enriching of (g) is via magnetic beads.

[00054] In an aspect of the invention, the engineered ICAD/CAD heterodimers are activated by one or more alternative protease prior to combining with a sample for evaluation. In an aspect of the invention, the engineered ICAD/CAD heterodimers are activated by one or more alternative protease upon or shortly after combining with a sample for evaluation.

[00055] In an aspect, a method is provided for detecting special proximity relationships between nucleic acid sequences comprising contacting a sample containing nucleic acid with a CAD enzyme and digesting the nucleic acid with CAD to generate digested ends of nucleic acid, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase.

[00056] The invention provides a system and kit for determining sequence and structure (three dimensional conformation) of one or more target regions or one or more genome in a sample comprising nucleic acid, comprising an engineered ICAD and CAD, wherein the engineered ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence, and further comprising an alternative protease to inactivate the engineered ICAD and activate the CAD such that the nucleic acid is specifically digested by the CAD.

[00057] In an embodiment, the system or kit further comprises:

(a) a ligation enzyme for ligating the labeled digested nucleic acid to generate ligated nucleic acid; and

(b) a means for purifying the ligated nucleic acid to generate proximally-ligated nucleic acid;

(c) a means for capturing the proximally-ligated nucleic acid.

[00058] In an embodiment the system or kit further comprises a means for labeling the nucleic acid digested by the CAD.

[00059] In an embodiment, the system or kit further comprises:

(a) a means for labeling the nucleic acid digested by the CAD;

(b) a ligation enzyme for ligating the labeled digested nucleic acid to generate ligated nucleic acid; and

(c) a means for purifying the ligated nucleic acid to generate proximally-ligated nucleic acid;

(d) a means for capturing the proximally-ligated nucleic acid.

[00060] In an embodiment, the system or kit further comprises reagents for sequencing the proximally-ligated nucleic acid.

[00061] In an embodiment, the system or kit further comprises reagents for generating a library of the proximally-ligated nucleic acid. In some embodiments, reagents for sequencing the proximally-ligated nucleic acid in the library are also included.

[00062] In embodiments of the methods, systems or kits, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8) or ENLYFQG (SEQ ID NO:9). In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

[00063] In some embodiments, the ICAD is an engineered murine ICAD or is an engineered human ICAD. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2, or variants thereof retaining the TEV protease recognition sequence and having up to 90% amino acid sequence identity with SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1.

[00064] In some embodiments, the CAD is a murine or human CAD. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6, or variants thereof retaining CAD endonuclease activity and having up to 90% amino acid sequence identity with SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5.

[00065] In accordance with the methods, kits and systems provided herein, a sample for evaluation or assessment may include eukaryotic cells derived from cell cultures, tissues, patient samples, or whole organisms. The eukaryotic cells may be cancer or tumor cells,, including cells from a biopsy. Prokaryotic cells or archae may also be suitable. Samples of cell free DNA are also contemplated.

[00066] Other objects and advantages will become apparent to those skilled in the art from a review of the ensuing detailed description, which proceeds with reference to the following illustrative drawings, and the attendant claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[00067] FIGURE 1 depicts the steps in proximity ligation via Hi-C. Genomic DNA is crosslinked and near regions are brought together. The crosslinked DNA is cut with an enzyme. Cleavage with restriction enzyme is depicted here. Then the ends are filled in as needed and marked with a marker or tag – biotin is depicted here as an example. The DNA ends are ligated together to permanently join the near regions. The DNA is purified and sheared or otherwise manipulated to generate smaller pieces or sections/fragments of DNA. Having the ends labeled (e.g. with biotin) subsequently allows for the ligation junctions to be enriched for by capture (e.g. with streptavidin). Paired-end sequencing can then proceed, for example after ligation of sequence adapters.

[00068] FIGURE 2A and B depict CAD enzyme activation via ICAD cleavage and homo-oligomerization to form active CAD. (A) Ordinary activation of CAD by caspase-3 cleavage. The inactive form of CAD (DFF) is a heterodimer composed of a 45-kDa chaperone inhibitor subunit ICAD (DFF45 or DFFA) and a latent endonuclease subunit CAD (DFF40 or DFFB). Upon caspase-3 cleavage of ICAD, CAD forms active endonuclease homo-oligomers. (B) Orthogonal activation of CAD by TEVp cleavage of $ICAD^{2xTEV}$. With the caspase-3 cleavae sites in ICAD deleted and replaced with TEV protease

recognition sequences, upon TEV cleavage of ICAD, CAD forms active endonuclease homo-oligomers. TST-muGFP-SUMO tag not shown for clarity.

[00069] FIGURE 3 provides a comparison of MNase and CAD mediated process. MNase's DNA exonuclease, endonuclease and nickase activity produces frayed, 3'-phosphorylated, unligatable DNA ends/DNA termini, in contrast to CAD's strict endonuclease activity, which generatesblunt, 5' phosphorylated DNA ends/DNA termini.

[00070] FIGURE 4 shows engineering of ICAD (murine ICAD) for orthogonal *in vitro* activation of CAD (murine) by the enzyme TEV protease (TEVp). The caspase recognition sites in ICAD are removed and TEV cleavage sites inserted in their place. This generates $ICAD^{2XTEV}$, which has TEV cleavage sites at each of the original caspase cleavage sites 1 and 2. Caspase no longer cleaves $ICAD^{2XTEV}$. Cleavage of $ICAD^{2XTEV}$ with TEV protease now specifically and controllably activates CAD enzyme.

[00071] FIGURE 5 depicts the modified vector pRSFDuet-1 for optimized co-expression of CAD and $ICAD^{2XTEV}$ at similar ratios. The vector as designed expresses an affinity tagged CAD as pRSFDuet-V2:: $ICAD^{2XTEV}$_TwinStrepTag-muGFP-$SUMO^{Eu1}$-CAD. This vector permits recombinant production and purification of $ICAD^{2XTEV}$-CAD proenzyme heterodimer from *E. coli*. TwinStrepTag affinity purification followed by size -exclusion chromatography is used for purification.

[00072] FIGURE 6 shows efficient *in vitro* cleavage of recombinant $ICAD^{2XTEV}$ by TEVp.

[00073] FIGURE 7 shows *in vitro* TEVp-dependent DNase activity of CAD-ICAD 2xTEV on λ phage genomic DNA substrate, providing TEVp activatable specific activity of CAD enzyme.

[00074] FIGURE 8 shows *in vitro* TEVp-dependent activity on human 1% formaldehyde-fixed chromatin.

[00075] FIGURE 9 depicts the evaluation of the effects of CAD concentration on digestion properties and digestion behavior across a 100 fold range of CAD concentration.

[00076] FIGURE 10 provides sequenced fragment length distribution of chromatin fragments produced from human K562 cells digested using varying MNase concentrations (left panel) versus FLDs of libraries generated from digestion of quiescent human hTERT-RPE1 cells using varying CAD concentrations (CAD-seq) (right panel). Fragment frequency plotted as sequencing depth-normalized counts per million read-pairs.

[00077] FIGURE 11 depicts normalized frequency of A/T, G/C nucleotide composition for first ten bases (5'-3') of aligned sequenced DNA fragments in CAD-seq libraries, relative enrichment calculated relative to bases 11-20. K562 MNase-seq data obtained from Mieczkowski et al. 2016 and hTERT-RPE1 CAD digest as in (B), data plotted from lowest nuclease concentration assayed, 5.4 U MNase, 2.5 U CAD.

[00078] FIGURE 12 depicts genome coverage as a function of GC content. Human hg38 coverage of MNase and CAD sequenced digest fragment (CAD-seq) is plotted as function of bin G/C%.

[00079] FIGURE 13 depicts Reads Per Genome Coverage (RPGC) of CAD (left panel) versus MNase (right panel). RPGC was normalized with 1.0 being equal to sequence coverage expected from overall sequencing depth if coverage is perfectly uniform genome wide. Mean RPGC is graphed versus position relative to CTCF sites. MNase and CAD digest titration (CAD-seq) aggregated fragment coverage plotted over top 5% occupied CTCF JASPAR MA0139.1 motifs calculated using publicly-available CTCF ChIP-seq from K562 and hTERT-RPE1 cells (Methods). All fragment lengths plotted for 3 kilobase window around occupied CTCF motifs, reads per genome coverage (RPGC) sequencing depth-normalized signal. Aggregated fragment coverage of MNase and CAD-seq digest titration at sub-nucelosomal length (<100bp) was also plotted (data not shown). CAD was evaluated at each of 2.5U, 25U and 250U, as well as 25U - RNase. MNase was evaluated and compared at 5.4U, 20.6U, 79.2U and 304 U.

[00080] FIGURE 14 shows ICE-corrected contact probability curve (upper) and slope of the contact probability curve (lower) for merged four technical replicates of CAD-C in quiescent human hTERT-RPE1 cells in comparison with Micro-C.

[00081] FIGURE 15 depicts Reads Per Genome Coverage (RPGC) of CAD (left panels) versus MNase (right panels). The sequenced DNA fragment coverage of sub-100 basepair length fragments around the top 5% CTCF-occupied CTCF motifs in the human hg38 genome was evaluated. The 'Reads Per Genome Coverage" (RPGC) was normalized with 1.0 being equal to sequence coverage expected from overall sequencing depth if coverage is perfectly uniform genome wide. CAD was evaluated at each of 2.5U, 25U and 250U, as well as 25U - RNase. MNase was evaluated and compared at 5.4U, 20.6U, 79.2U and 304 U.

[00082] FIGURE 16 depicts that CAD cleavage is moderately enhanced in open chromatin over closed chromatin. Aggregated coverage of CAD digest fragments (all fragment length RPGC-normalized signal) over H3K9me3 marked regions in hTERT-RPE1 cells, over H3K27ac marked regions in hTERT-RPE1 cells, over H3K4me1 marked regions in hTERT-RPE1 cells and over H3K27me3 marked regions in hTERT-RPE1 cells. Region width-scaled coverage with 5000 bp flanking regions. Merged two technical replicates per CAD concentration condition.

[00083] FIGURE 17 shows CAD-C contact probability curves in different lines and comparison of long-range contacts between CAD-C, Omni-C and Micro-C. A) Contact probability curve of ICE-balanced matrices from the three methods in GM12878 cells. Micro-C and Omni-C data from Dovetail Genomics (cantatabio.com; Cantata Biotech). Matrix base resolution was 500 bp. B) Contact probability curves of CAD-C libraries from hTERT-RPE-1, GM12879 and K562 cells. All were generated form ICE-balanced matrices with a base resolution of 500 bp.

[00084] FIGURE 18A and 18B depicts that end-repair and biotinylation are not necessary for efficient proximity ligation of nucleosomal fragment libraries generated by CAD cleavage. (A) Libraries were

prepared in duplicate from GM12878 cells as indicated in the gel labels. The upward shift of the library fragment lengths is due to proximity ligation of nucleosomal fragments into multi-nucleosome "walks." (B) The insert length distribution of the PacBio HiFi sequenced library, size-selected to enrich for multimers over 1 kb.

[00085] FIGURE 19 shows CADwalks, which are defined herein to consist of a consecutive series of genomic fragments within the long read. The long concatenated library molecules sequenced with PacBio HiFi reads consist of genomic fragments (nucleosomal or sub-nucleosomal) that are ligated together *in situ* before being removed from the fixed cell and prepared for sequencing.

[00086] Figure 20 shows CAD-C contact distance statistics in GM12878 cells depending on whether ligation junctions are required to be directly sequenced in one of the reads (UU direct ligations) or can be inferred from mapping the pair of reads (all UU).

[00087] Figure 21 depicts CADwalk chromosome contact and walk length statistics from PacBio HiFi reads

[00088] Figure 22 depicts size distribution of CAD-C library after UltraExpress FS enzymatic fragmentation as an alternative to sonication. This system uses fragmentase. The average size is ~400-500 bp. In this, the fragmentation of fragments is carried out with NEBNext UltraExpress FS kit (E3340S) following the manufacturer's instructions with the following modifications: 5 minutes FS fragmentation reaction time, modified library bead cleanup ratios of 0.5x Ampure XP beads and 0.6x NEB bead reconstitution buffer. 6 PCR cycles were used for material amplification. Libraries were eluted in 35 µL 0.1x TE, library concentration exceeded 100 nM.

[00089] Figure 23 provides a scheme of 3-fragment CADwalks readable by Illumina short-read paired-end sequencing (150 x 150 bp paired end).

[00090] Figure 24 depicts that middle ("B") fragment inferred length distribution in 3-fragment CADwalks shows a predominantly mononucleosomal distribution, indicating that 3-fragment CADwalks are formed by mononucleosome fragments ligating to other chromatin.

[00091] Figure 25 shows fragment pileup coverage (top) and fragment end coverage (bottom) of 3-fragment CADwalks around CTCF-bound sites show that CADwalk fragments reproduce nucleosome positioning and chromatin footprinting patterns seen with MNase, ATAC-seq and other methods. A CTCF footprint is observed at the CTCF motif (graph center), flanked by arrays of positioned nucleosomes.

DETAILED DESCRIPTION

[00092] In accordance with the present disclosure there may be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature. See, e.g., Sambrook et al, "Molecular Cloning: A Laboratory Manual"

(1989); "Current Protocols in Molecular Biology" Volumes I-III [Ausubel, R. M., ed. (1994)]; "Cell Biology: A Laboratory Handbook" Volumes I-III [J. E. Celis, ed. (1994))]; "Current Protocols in Immunology" Volumes I-III [Coligan, J. E., ed. (1994)]; "Oligonucleotide Synthesis" (M.J. Gait ed. 1984); "Nucleic Acid Hybridization" [B.D. Hames & S.J. Higgins eds. (1985)]; "Transcription And Translation" [B.D. Hames & S.J. Higgins, eds. (1984)]; "Animal Cell Culture" [R.I. Freshney, ed. (1986)]; "Immobilized Cells And Enzymes" [IRL Press, (1986)]; B. Perbal, "A Practical Guide To Molecular Cloning" (1984).

[00093] Therefore, if appearing herein, the following terms shall have the definitions set out below.

A. TERMINOLOGY

[00094] Amplification: To increase the number of copies of a nucleic acid molecule, such as one or more end joined nucleic acid fragments that includes a junction, such as a ligation junction. The resulting amplification products are called "amplicons." Amplification of a nucleic acid molecule (such as a DNA or RNA molecule) refers to use of a technique that increases the number of copies of a nucleic acid molecule (including fragments).

[00095] An example of amplification is the polymerase chain reaction (PCR), in which a sample is contacted with a pair of oligonucleotide primers under conditions that allow for the hybridization of the primers to a nucleic acid template in the sample. The primers are extended under suitable conditions, dissociated from the template, re-annealed, extended, and dissociated to amplify the number of copies of the nucleic acid. This cycle can be repeated. The product of amplification can be characterized by such techniques as electrophoresis, restriction endonuclease cleavage patterns, oligonucleotide hybridization or ligation, and/or nucleic acid sequencing.

[00096] Other examples of in vitro amplification techniques include quantitative real-time PCR; reverse transcriptase PCR (RT-PCR); real-time PCR (rt PCR); real-time reverse transcriptase PCR (rt RT-PCR); nested PCR; strand displacement amplification (see U.S. Pat. No. 5,744,311); transcription-free isothermal amplification (see U.S. Pat. No. 6,033,881, repair chain reaction amplification (see WO 90/01069); ligase chain reaction amplification (see European patent publication EP-A-320 308); gap filling ligase chain reaction amplification (see U.S. Pat. No. 5,427,930); coupled ligase detection and PCR (see U.S. Pat. No. 6,027,889); and NASBA™ RNA transcription-free amplification (see U.S. Pat. No. 6,025,134) amongst others.

[00097] Binding or stable binding (of an oligonucleotide): An oligonucleotide, such as a nucleic acid probe that specifically binds to a target junction in an end joined nucleic acid fragment, binds or stably binds to a target nucleic acid if a sufficient amount of the oligonucleotide forms base pairs or is hybridized to its target nucleic acid. For example depending in the hybridization conditions, there need not be

complete matching between the probe and the nucleic acid target, for example there can be mismatch, or a nucleic acid bubble. Binding can be detected by either physical or functional properties.

[00098] Binding site: A region on a protein, DNA, or RNA to which other molecules stably bind. In one example, a binding site is the site on an end joined nucleic acid fragment.

[00099] Capture moieties: Molecules or other substances that when attached to a nucleic acid molecule, such as an end joined nucleic acid, allow for the capture of the nucleic acid molecule through interactions of the capture moiety and something that the capture moiety binds to, such as a particular surface and/or molecule, such as a specific binding molecule that is capable of specifically binding to the capture moiety.

[000100] Complementary: A double-stranded DNA or RNA strand consists of two complementary strands of base pairs. Complementary binding occurs when the base of one nucleic acid molecule forms a hydrogen bond to the base of another nucleic acid molecule. Normally, the base adenine (A) is complementary to thymidine (T) and uracil (U), while cytosine (C) is complementary to guanine (G). For example, the sequence 5'-ATCG-3' of one ssDNA molecule can bond to 3'-TAGC-5' of another ssDNA to form a dsDNA. In this example, the sequence 5'-ATCG-3' is the reverse complement of 3'-TAGC-5'. Nucleic acid molecules can be complementary to each other even without complete hydrogen-bonding of all bases of each molecule. For example, hybridization with a complementary nucleic acid sequence can occur under conditions of differing stringency in which a complement will bind at some but not all nucleotide positions.

[000101] Contacting: Placement in direct physical association, including both in solid or liquid form, for example contacting a sample with a crosslinking agent or a probe.

[000102] Control: A reference standard. A control can be a known value or range of values indicative of basal levels or amounts or present in a tissue or a cell or populations thereof. A control can also be a cellular or tissue control, for example a tissue from a non-diseased state and/or exposed to different environmental conditions. A difference between a test sample and a control can be an increase or conversely a decrease. The difference can be a qualitative difference or a quantitative difference, for example a statistically significant difference.

[000103] Covalently linked: Refers to a covalent linkage between atoms by the formation of a covalent bond characterized by the sharing of pairs of electrons between atoms. In one example, a covalent link is a bond between an oxygen and a phosphorous, such as phosphodiester bonds in the backbone of a nucleic acid strand. In another example, a covalent link is one between a nucleic acid protein, another protein and/or nucleic acid that has been crosslinked by chemical means. In another example, a covalent link is one between fragmented nucleic acids.

[000104] Crosslinking agent: A chemical agent or even light, which facilitates the attachment of one molecule to another molecule. Crosslinking agents can be protein-nucleic acid crosslinking agents, nucleic

acid-nucleic acid crosslinking agents, and protein-protein crosslinking agents. Examples of such agents are known in the art. In some embodiments, a crosslinking agent is a reversible crosslinking agent. In some embodiments, a crosslinking agent is a non-reversible crosslinking agent.

[000105] Detect: To determine if an agent (such as a signal or particular nucleic acid or protein) is present or absent. In some examples, this can further include quantification in a sample, or a fraction of a sample, such as a particular cell or cells within a tissue.

[000106] Detectable label: A compound or composition that is conjugated directly or indirectly to another molecule to facilitate detection of that molecule. Specific, non-limiting examples of labels include fluorescent tags, enzymatic linkages, and radioactive isotopes and other physical tags, such as biotin. In some examples, a label is attached to a nucleic acid, such as an end-joined nucleic acid, to facilitate detection and/or isolation of the nucleic acid.

[000107] DNA sequencing: The process of determining the nucleotide order of a given DNA molecule. Generally, the sequencing can be performed using automated Sanger sequencing (AB13730xl genome analyzer), pyrosequencing on a solid support (454 sequencing, Roche), sequencing-by-synthesis with reversible terminations (ILLUMINA® Genome Analyzer), sequencing-by-ligation (ABI SOLiD®) or sequencing-by-synthesis with virtual terminators (HELISCOPE®). In some embodiments, DNA sequencing is performed using a chain termination method developed by Frederick Sanger, and thus termed "Sanger based sequencing" or "SBS." This technique uses sequence-specific termination of a DNA synthesis reaction using modified nucleotide substrates. Extension is initiated at a specific site on the template DNA by using a short oligonucleotide primer complementary to the template at that region. The oligonucleotide primer is extended using DNA polymerase in the presence of the four deoxynucleotide bases (DNA building blocks), along with a low concentration of a chain terminating nucleotide (most commonly a di-deoxynucleotide). Limited incorporation of the chain terminating nucleotide by the DNA polymerase results in a series of related DNA fragments that are terminated only at positions where that particular nucleotide is present. The fragments are then size-separated by electrophoresis a polyacrylamide gel, or in a narrow glass tube (capillary) filled with a viscous polymer. An alternative to using a labeled primer is to use labeled terminators instead; this method is commonly called "dye terminator sequencing."

[000108] "Pyrosequencing" is an array based method, which has been commercialized by 454 Life Sciences. In some embodiments of the array-based methods, single-stranded DNA is annealed to beads and amplified via EmPCR®. These DNA-bound beads are then placed into wells on a fiber-optic chip along with enzymes that produce light in the presence of ATP. When free nucleotides are washed over this chip, light is produced as the PCR amplification occurs and ATP is generated when nucleotides join with their complementary base pairs. Addition of one (or more) nucleotide(s) results in a reaction that generates a light signal that is recorded, such as by the charge coupled device (CCD) camera, within the instrument.

The signal strength is proportional to the number of nucleotides, for example, homopolymer stretches, incorporated in a single nucleotide flow.

[000109] Fluorophore: A chemical compound, which when excited by exposure to a particular stimulus such as a defined wavelength of light, emits light (fluoresces), for example at a different wavelength (such as a longer wavelength of light). Fluorophores are part of the larger class of luminescent compounds. Luminescent compounds include chemiluminescent molecules, which do not require a particular wavelength of light to luminesce, but rather use a chemical source of energy. Therefore, the use of chemiluminescent molecules (such as aequorin) eliminates the need for an external source of electromagnetic radiation, such as a laser. Examples of particular fluorophores that can be used in the probes disclosed herein are provided in U.S. Pat. No. 5,866,336 issued to Nazarenko et al., such as 4-acetamido-4'-isothiocyanatostilbene-2,2'disulfonic acid, acridine and derivatives such as acridine and acridine isothiocyanate, 5-(2'-aminoethyl)aminonaphthalene-l-sulfonic acid (EDANS), 4-amino-N-[3 vinylsulfonyl)phenyl]naphthalimide-3,5 disulfonate (Lucifer Yellow VS), N-(4-anilino-lnaphthyl) maleimide, anthranilamide, Brilliant Yellow, coumarin and derivatives such as coumarin, 7-amino-4-methylcoumarin (AMC, Coumarin 120), 7-amino-4- trifluoromethylcouluarin (Coumaran 151); cyanosine; 41,6-20 diaminidino-2-phenylindole (DAPI); 5', 5"-dibromopyrogallol-sulfonephthalein (Bromopyrogallol Red); 7 -diethy lamino-3-(4 '-isothiocyanatophenyl)-4-methylcoumarin; diethylenetriamine pentaacetate; 4,4'-diisothiocyanatodihydrostilbene-2,2'-di sulfonic acid; 4,4'-diisothio-cyanatostilbene-2,2'-disulfonic acid; 5-[dimethylamino]naphthalene-1-sulfonyl chloride (DNS, dansyl chloride); 4-dimethy laminopheny lazopheny 1-4 '-isothiocyanate (DABITC); eosin and derivatives such as eosin and eosin isothiocyanate; erythrosin and derivatives such as erythrosine and erythrosin isothiocyanate; ethidium; fluorescein and derivatives such as 5-carboxyfluorescein (FAM), 5-(4,6-dichlorotriazin-2-yl)aminofluorescein (DTAF), 2'7'-dimethoxy- 4'5'-dichloro-6-carboxyfluorescein (JOE), fluorescein, fluorescein isothiocyanate (FITC), and QFITC (XRITC); fluorescamine; IR144; IR1446; Malachite Green isothiocyanate; 4-methylumbelliferone; ortho cresolphthalein; nitrotyrosine; pararosaniline; Phenol Red; B-phycoerythrin; o-phthaldialdehyde; pyrene and derivatives such as pyrene, pyrene butyrate and succinimidyl 1-pyrene butyrate; Reactive Red 4 (Cibacron™. Brilliant Red 3B-A); rhodamine and derivatives such as 6-carboxy-X-rhodamine (ROX), 6-carboxyrhodamine (R6G), lissamine rhodamine B sulfonyl chloride, rhodamine (Rhod), rhodamine B, rhodamine 123, rhodamine X isothiocyanate, sulforhodamine B, sulforhodamine 101 and sulfonyl chloride derivative ofsulforhodamine 101 (Texas Red); N,N,N',N'-tetramethyl-6-carboxyrhodamine (TAMRA); tetramethyl rhodamine;tetramethyl rhodamine isothiocyanate (TRITC); riboflavin; rosolic acid and terbium chelate derivatives; LightCycler Red 640; Cy5.5; and Cy56-carboxyfluorescein; 5-carboxyfluorescein (5-FAM); boron dipyrromethene difluoride (BOD IPY); N,N,N' ,N'-tetramethy 1-6-carboxyrhodamine (TAMRA);

acridine, stilbene, -6-carboxy-fluorescein (HEX), TET (Tetramethyl fluorescein), 6-carboxy-X-rhodamine (ROX), Texas Red, 2',7'-dimethoxy-4',5'-dichloro-6-carboxyfluorescein (JOE), Cy3, Cy5, VIC® (Applied Biosystems), LC Red 640, LC Red 705, Yakima yellow amongst others.

[000110] "Specifically hybridizable" and "specifically complementary" are terms that indicate a sufficient degree of complementarity such that stable and specific binding occurs between the oligonucleotide (or it's analog) and the DNA, RNA, and or DNA-RNA hybrid target. The oligonucleotide or oligonucleotide analog need not be 100% complementary to its target sequence to be specifically hybridizable. An oligonucleotide or analog is specifically hybridizable when there is a sufficient degree of complementarity to avoid non-specific binding of the oligonucleotide or analog to non-target sequences under conditions where specific binding is desired. Such binding is referred to as specific hybridization.

[000111] Isolated: An "isolated" biological component (such as the end joined fragmented nucleic acids described herein) has been substantially separated or purified away from other biological components in the cell of the organism, in which the component naturally occurs, for example, extra-chromatin DNA and RNA, proteins and organelles. Nucleic acids and proteins that have been "isolated" include nucleic acids and proteins purified by standard purification methods, for example from a sample. The term also embraces nucleic acids and proteins prepared by recombinant expression in a host cell as well as chemically synthesized nucleic acids. It is understood that the term "isolated" does not imply that the biological component is free of trace contamination, and can include nucleic acid molecules that are at least 50% isolated, such as at least 75%, 80%, 90%, 95%, 98%, 99%, or even 100% isolated.

[000112] Junction: A site where two nucleic acid fragments or joined, for example using the methods described herein. A junction encodes information about the proximity of the nucleic acid fragments that participate in formation of the junction. For example, junction formation between to nucleic acid fragments indicates that these two nucleic acid sequences where in close proximity when the junction was formed, although they may not be in proximity in linear nucleic acid sequence space. Thus, a junction can define long range interactions. In some embodiments, a junction is labeled, for example with a labeled nucleotide, for example to facilitate isolation of the nucleic acid molecule that includes the junction.

[000113] Nucleic acid (molecule or sequence): A deoxyribonucleotide or ribonucleotide polymer including without limitation, cDNA, mRNA, genomic DNA, and synthetic (such as chemically synthesized) DNA or RNA or hybrids thereof. The nucleic acid can be double-stranded (ds) or singlestranded (ss). Where single-stranded, the nucleic acid can be the sense strand or the antisense strand. Nucleic acids can include natural nucleotides (such as A, T/U, C, and G), and can also include analogs of natural nucleotides, such as labeled nucleotides. Examples of modified base moieties which can be used to modify nucleotides at any position on its structure include, but are not limited to: 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, acetylcytosine, 5-

(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethy laminomethy luracil, dihydrouracil, beta-Dgalactosylqueosine, inosine, N-6-sopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methyl cytosine, N6-adenine, 7-methylguanine, 5-methylaminomethyluracil, methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-S-oxyacetic acid, 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, 2,6-diaminopurine and biotinylated analogs, amongst others.

[000114] Examples of modified sugar moieties which may be used to modify nucleotides at any position on its structure include, but are not limited to arabinose, 2-fluoroarabinose, xylose, and hexose, or a modified component of the phosphate backbone, such as phosphorothioate, a phosphorodithioate, a phosphoramidothioate, a phosphoramidate, a phosphordiamidate, a methylphosphonate, an alkyl phosphotriester, or a formacetal or analog thereof.

[000115] Primers: Short nucleic acid molecules, such as a DNA oligonucleotide, which can be annealed to a complementary target nucleic acid molecule by nucleic acid hybridization to form a hybrid between the primer and the target nucleic acid strand. A primer can be extended along the target nucleic acid molecule by a polymerase enzyme. Therefore, primers can be used to amplify a target nucleic acid molecule, wherein the sequence of the primer is specific for the target nucleic acid molecule, for example so that the primer will hybridize to the target nucleic acid molecule under very high stringency hybridization conditions. The specificity of a primer increases with its length. Thus, for example, a primer that includes 30 consecutive nucleotides will anneal to a target sequence with a higher specificity than a corresponding primer of only 15 nucleotides. Thus, to obtain greater specificity, probes and primers can be selected that include at least 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 or more consecutive nucleotides. In particular examples, a primer is at least 15 nucleotides in length, such as at least 5 contiguous nucleotides complementary to a target nucleic acid molecule. Particular lengths of primers that can be used to practice the methods of the present disclosure include primers having at least 5, at least 10, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 45, at least 50, or more contiguous nucleotides complementary to the target nucleic acid molecule to be amplified, such as a primer of 5-60 nucleotides, 15-50 nucleotides, 15-30 nucleotides or greater. Primer pairs can be used for amplification of a nucleic acid sequence, for example, by PCR, or other nucleic-acid amplification methods known in the art.

[000116] Probe: A probe comprises an isolated nucleic acid capable of hybridizing to a target nucleic acid (such as end joined nucleic acid fragment). A detectable label or reporter molecule can be attached to a probe. Typical labels include radioactive isotopes, enzyme substrates, co-factors, ligands, chemiluminescent or fluorescent agents, haptens, and enzymes. Methods for labeling and guidance in the choice of labels appropriate for various purposes are known in the art. Probes are generally at least 5 nucleotides in length, such as at least 10, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 41, at least 42, at least 43, at least 44, at least 45, at least 46, at least 47, at least 48, at least 49, at least 50 at least 51, at least 52, at least 53, at least 54, at least 55, at least 56, at least 57, at least 58, at least 59, at least 60, or more contiguous nucleotides complementary to the target nucleic acid molecule, such as 50-60 nucleotides, 20-50 nucleotides, 20-40 nucleotides, 20-30 nucleotides or greater.

[000117] Sample: A sample, such as a biological sample, that includes biological materials (such as nucleic acid and proteins, for example double-stranded nucleic acid binding proteins) obtained from an organism or a part thereof, such as a plant, animal, bacteria, and the like. In particular embodiments, the biological sample is obtained from an animal subject, such as a human subject. A biological sample is any solid or fluid sample obtained from, excreted by or secreted by any living organism, including without limitation, single celled organisms, such as bacteria, yeast, protozoans, and amebas among others, multicellular organisms (such as plants or animals, including samples from a healthy or apparently healthy human subject or a human patient affected by a condition or disease to be diagnosed or investigated, such as cancer). For example, a biological sample can be a biological fluid obtained from, for example, blood, plasma, serum, urine, bile, ascites, saliva, cerebrospinal fluid, aqueous or vitreous humor, or any bodily secretion, a transudate, an exudate (for example, fluid obtained from an abscess or any other site of infection or inflammation), or fluid obtained from a joint (for example, a normal joint or a joint affected by disease, such as a rheumatoid arthritis, osteoarthritis, gout or septic arthritis). A sample can also be a sample obtained from any organ or tissue (including a biopsy or autopsy specimen, such as a tumor biopsy) or can include a cell (whether a primary cell or cultured cell) or medium conditioned by any cell, tissue or organ.

[000118] Specific Binding Agent: An agent that binds substantially or preferentially only to a defined target such as a protein, enzyme, polysaccharide, oligonucleotide, DNA, RNA, recombinant vector or a small molecule. In an example, a "specific binding agent that specifically binds to the label" is capable of binding to a label that is covalently linked to a targeting probe. A nucleic acid-specific binding agent binds substantially only to the defined nucleic acid, such as DNA, or to a specific region within the nucleic acid, for example a nucleic acid probe. A protein-specific binding agent binds substantially only the defined

protein, or to a specific region within the protein. For example, a "specific binding agent" includes antibodies and other agents that bind substantially to a specified polypeptide.

[000119] Antibodies can be monoclonal or polyclonal antibodies that are specific for the polypeptide, as well as immunologically effective portions ("fragments") thereof. The term "antibody" describes an immunoglobulin whether natural or partly or wholly synthetically produced. The term also covers any polypeptide or protein having a binding domain which is, or is homologous to, an antibody binding domain. CDR grafted antibodies are also contemplated by this term. An "antibody" is any immunoglobulin, including antibodies and fragments thereof, that binds a specific epitope. The term encompasses polyclonal, monoclonal, and chimeric antibodies. The term "antibody(ies)" includes a wild type immunoglobulin (Ig) molecule, generally comprising four full length polypeptide chains, two heavy (H) chains and two light (L) chains, or an equivalent Ig homologue thereof (e.g., a camelid nanobody, which comprises only a heavy chain); including full length functional mutants, variants, or derivatives thereof, which retain the essential epitope binding features of an Ig molecule, and including dual specific, bispecific, multispecific, and dual variable domain antibodies; Immunoglobulin molecules can be of any class (e.g., IgG, IgE, IgM, IgD, IgA, and IgY), or subclass (e.g., IgG1, IgG2, IgG3, IgG4, IgA1, and IgA2). Also included within the meaning of the term "antibody" are any "antibody fragment".

[000120] An "antibody fragment" means a molecule comprising at least one polypeptide chain that is not full length, including (i) a Fab fragment, which is a monovalent fragment consisting of the variable light (VL), variable heavy (VH), constant light (CL) and constant heavy 1 (CH1) domains; (ii) a F(ab')2 fragment, which is a bivalent fragment comprising two Fab fragments linked by a disulfide bridge at the hinge region; (iii) a heavy chain portion of an Fab (Fd) fragment, which consists of the VH and CH1 domains; (iv) a variable fragment (Fv), which consists of the VL and VH domains of a single arm of an antibody, (v) a domain antibody (dAb) fragment, which comprises a single variable domain (Ward, E.S. et al., Nature 341, 544-546 (1989)); (vi) a camelid antibody; (vii) an isolated complementarity determining region (CDR); (viii) a Single Chain Fv Fragment wherein a VH domain and a VL domain are linked by a peptide linker which allows the two domains to associate to form an antigen binding site (Bird et al, Science, 242, 423-426, 1988; Huston et al, PNAS USA, 85, 5879-5883, 1988); (ix) a diabody, which is a bivalent, bispecific antibody in which VH and VL domains are expressed on a single polypeptide chain, but using a linker that is too short to allow for pairing between the two domains on the same chain, thereby forcing the domains to pair with the complementarity domains of another chain and creating two antigen binding sites (WO94/13804; P. Holliger et al Proc. Natl. Acad. Sci. USA 90 6444-6448, (1993)); and (x) a linear antibody, which comprises a pair of tandem Fv segments (VH-CH1-VH-CH1) which, together with complementarity light chain polypeptides, form a pair of antigen binding regions; (xi) multivalent antibody fragments (scFv dimers, trimers and/or tetramers (Power and Hudson, J Immunol. Methods 242:

193-204 9 (2000)); (xii) a minibody, which is a bivalent molecule comprised of scFv fused to constant immunoglobulin domains, CH3 or CH4, wherein the constant CH3 or CH4 domains serve as dimerization domains (Olafsen T et al (2004) Prot Eng Des Sel 17(4):315-323; Hollinger P and Hudson PJ (2005) Nature Biotech 23(9):1126-1136); and (xiii) other non-full length portions of heavy and/or light chains, or mutants, variants, or derivatives thereof, alone or in any combination.

[000121] Test agent: Any agent that that is tested for its effects, for example its effects on a cell. In some embodiments, a test agent is a chemical compound, such as a chemotherapeutic agent, antibiotic, or even an agent with unknown biological properties.

[000122] Tissue: A plurality of functionally related cells. A tissue can be a suspension, a semi-solid, or solid. Tissue includes cells collected from a subject such as blood, cervix, uterus, lymph nodes breast, skin, and other organs.

[000123] The term "binding" refers to a direct association between molecules and/or atoms, due to, for example, covalent, electrostatic, hydrophobic, and ionic and/or hydrogen-bond interactions, including interactions such as salt bridges and water bridges. The term "bonding" as used herein generally refers to covalent bonding.

[000124] The term affinity" as used herein generally refers to the strength of non-covalent binding. A lower KD signifies increased binding affinity. Binding between an antibody and antigen have an "affinity" that can be described by the dissociation constant (KD). An antibody or fragment thereof, including for example, Fab and Fv, can have much higher affinity for a biological target molecule than it has for an unrelated amino acid sequence, for example, at least 1, 2, 3, 4, 5, 6, 8, 10, 20, 40, 60, 100 or 1000 fold greater. Affinity of an antibody or a portion or fragment thereof to a biological target molecule can be, for example, from about 100 nanomolar (nM) to about 0.1 nM, from about 100 nM to about 1 picomolar (pM), or from about 100 nM to about 1 femtomolar (fM), or more.

[000125] The term "comprise" has its customary meaning, generally used in the sense of include, that is to say permitting the presence of one or more features or components.

[000126] The term "consisting essentially of" has its customary meaning and, for example, refers to a product, such as a peptide sequence, of a defined number of residues which is not covalently attached to a larger product. The disclosure permits other embodiments substituting "consisting essentially of" for "comprising" or "including."

[000127] As used herein, "pg" means picogram, "ng" means nanogram, "ug" or "µg" mean microgram, "mg" means milligram, "ul" or "µl" mean microliter, "ml" means milliliter, "l" means liter.

[000128] The invention relates generally to reagents, methods, systems and kits for determination of the organizational structure of chromatin and chromosomes in three dimensions, including whereby information as to the structure and sequence is determined and provided. Methods are provided that allow

one skilled in the art to study chromosome architecture at all scales in as many sample types as possible, including rare cell types and precious samples with limited or lower cell numbers. We identified chromatin fragmentation as the primary bottleneck in the 3C methods and we have designed and implemented Caspase Activated DNase or CAD, which evolved to uniformly fragment chromosomes during the onset of apoptosis for 3C methods and determination of chromosome and chromatin organizational structure and higher order structural analysis of nucleic acids. CAD is also known as 40-kDa DNA Fragmentation Factor (DFF40). CAD has the unique property of being a strict endonuclease and retains the end point digestion property of restriction enzymes (as in Hi-C) while keeping the high resolution of MNase, with lesser DNA sequence bias. We developed a novel recombinant nuclease based on Caspase-Activated DNase (CAD). Unlike MNase, CAD has no substantial exonuclease activity and produces blunt-ended, ligatable DNA ends. Further, CAD does not overdigest chromatin over a wide range of concentrations. These unique properties of CAD improve the prior art approaches and methods, including Hi-C and Micro-C, and their resolution without sacrificing sensitivity. Thus, novel methods based on a specific engineered ICAD/CAD system have been developed and are provided.

[000129] Caspase Activated DNase or CAD (also denoted DFF-40, DFFB) is a major apoptotic nuclease in most metazoans and uniformly fragments chromosomes during the onset of apoptosis. Apoptosis is a programmed cell death module resulting in the removal of unwanted cells from an organism and genomic DNA fragmentation accompanies apoptosis. CAD is closely associated with genomic DNA fragmentation that accompanies apoptosis. Caspase 3 activates CAD by proteolytic inactivation of the inhibitor ICAD. Nuclease activity of CAD is restricted by association with its inhibitor ICAD. The association of CAD with ICAD begins with translation of CAD mRNA where ICAD acts as a chaperone to ensure CAD is properly folded. Inactivating cleavage of ICAD by activated caspases (particularly caspase-3) releases CAD from inhibition and facilitates CAD homodimerization (FIGURE 2A; Larsen BD and Sorensen CS (2017) The FEBS Journal 284:1160-1170).

[000130] In accordance with the invention, a controllable and activatable CAD enzyme has been generated by engineering of the inhibitor ICAD to be proteolytically cleaved and inactivated specifically by an alternative enzyme. Ordinarily CAD is activated by proteolytic cleavage of ICAD with caspase. In accordance with the invention, a CAD/ICAD system has been designed wherein ICAD is cleaved by an enzyme distinct from caspase. The caspase cleavage sequences in ICAD are specifically replaced by cleavage sequences and amino acids recognized by an alternative enzyme. This provides a CAD/ICAD system under control of and activated by an alternative enzyme.

[000131] In an embodiment, caspase recognition sequence selected from DEP, DET and/or DAV is deleted and is replaced with a protease recognition sequence for an alternative protease. Alternative proteases may be selected from known sequence specific proteases, particularly those having short

recognition sequences of between three amino acids and ten amino acids. Alternative proteases may be selected from known sequence specific proteases, whereby replacement of the alternative protease sequence in the ICAD generates an engineered ICAD, wherein the engineered ICAD is capable of forming a heterodimer with CAD thereby inactivating CAD enzyme and wherein the engineered ICAD is specifically cleaved by the alternative protease. In an embodiment, the engineered ICAD is specifically cleaved by the alternative protease and is not cleaved by caspase.

[000132] In an exemplary aspect, the enzyme TEV protease (Tobacco Etch Virus nuclear-inclusion-a endopeptidase; TEVp) is utilized herein as an alternative activating enzyme for ICAD and CAD/ICAD system. TEVp demonstrates high sequence specificity and has been used for the controlled cleavage of fusion proteins *in vitro* and *in vivo*. TEVp shows a 10-fold loss of activity at 4°C and loses activity at temperatures above 34°C, therefore it can be readily inactivated by temperature alteration. TEV protease is relatively non-toxic *in vivo* as the recognized sequence scarcely occurs in native proteins.

[000133] The alternative protease can be selected from various proteolytic enzymes, provided the proteolytic enzyme recognizes a specific amino acid sequence for cleavage with stringent specificity. Various proteases which could be suitable are available in the art. Waugh provides a review of commonly used proteases, particularly endoproteases (Waugh, David S. (2011) Prot Expression and Purification 80:283-293). Proteases suitable as alternative proteases in accordance with the invention include TEV protease (which is particularly exemplified herein), enteropeptidase (also called enterokinase), thrombin, Factor Xa, Rhinovirus 3C protease, HRV 3C protease, and also carboxypeptidase A and B, DAPase, PreScission protease. PreScission Protease is a fusion protein of glutathione S-transferase (GST) and human rhinovirus (HRV) type 14 3C protease and recognition site for this enzyme is the sequence Leu-Glu-Val-Leu-Phe-Gln/Gly-Pro (LEVLFQ/GP) and cleavage occurs between the Gln and Gly-Pro residues. Thrombin recognizes the consensus sequence Leu-Val-Pro-Arg-Gly-Ser, cleaving the peptide bond between Arg and Gly. Factor Xa protease preferentially cleaves after the arginine residue in the amino acid sequence Ile-Glu-Gly-Arg and its its preferred cleavage site *Ile-(Glu or Asp)-Gly-Arg*. Enteropeptidase (enterokinase) recognizes a highly specific amino acid sequence 'DDDDK' and cleaves after the lysine (K) residue. HRV 3C Protease cleaves protein substrates with the recognition sequence Leu-Glu-Val-Leu-Phe-Gln-Gly-Pro between the Gln and Gly residues. Other and new sets of highly efficient proteases have also been described (Frey S and Gorlich D (2014) J of Chromatography A 1337:95-105). Additional alternative proteases include the SUMO-specific and the NEDD8-specific protease from *Brachypodium distachyon* (bdSENP1 and bdNEP1), the NEDP1 protease from *Salmo salar* (ssNEDP1), *Saccharomyces cerevisiae* Atg4p (scAtg4) and *Xenopus laevis* Usp2 (xlUsp2).

[000134]These alternative protease recognition amino acid sequences can be incorporated in ICAD sequence to replace caspase recognition amino acids. The thereby engineered ICAD is then evaluated for

activity, complexing with CAD to form a heteodimer, proper folding and expression. The efficiency of activating CAD in a suitably engineered ICAD/CAN system with the selected alternative protease is then further evaluated. Exemplary approaches to select and evaluate are provided herein, including with TEV protease, and can be applied to alternative proteases.

[000135]In an embodiment, the TEV protease recognition sequence of ENLYFQS is inserted to replace caspase recognition amino acids. In an embodiment, the TEV protease recognition sequence of ENLYFQS is inserted to replace a minimal number of caspase recognition amino acids. In an embodiment, the TEV protease recognition sequence of ENLYFQS is inserted to replace three caspase recognition amino acids. In an embodiment, the ENLYFQS sequence is inserted at both caspase recognition locations and replaces DEP in one location and replaces DAV in the other in the mouse ICAD sequence. An exemplary sequence variant is shown in FIGURE 4 and provided in SEQ ID NO: 1. Notably, the replacement variant of ICAD ($ICAD^{2xTEV}$) specifically alters the cleavage location and sequences of the cleaved ICAD protein fragments, i.e. ICAD is cleaved in a different location.

[000136]The ICAD wildtype/native murine sequence is as follows (SEQ ID NO:3):

MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPIT
LVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWIYNDSDGGTAWVSQESFEADEPDSRA
GVKWKNVARQLKEDLSSIILLSEEDLQALIDIPCAELAQELCQSCATVQGLQSTLQQVLD
QREEARQSKQLLELYLQALEKEGNILSNQKESKAALSEELDAVDTGVGREMASEVLLRSQ
ILTTLKEKPAPELSLSSQDLESVSKEDPKALAVALSWDIRKAETVQQACTTELALRLQQV
QSLHSLRNLSARRSPLPGEPQRPKRAKRDSS

[000137]The murine ICAD sequence with the amino acids deleted to eliminate caspase-3 proteolysis underlined and in BOLD is shown below (SEQ ID NO:3):

MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPIT
LVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWIYNDSDGGTAWVSQESFEA**DEP**DSRA
GVKWKNVARQLKEDLSSIILLSEEDLQALIDIPCAELAQELCQSCATVQGLQSTLQQVLD
QREEARQSKQLLELYLQALEKEGNILSNQKESKAALSEEL**DAV**DTGVGREMASEVLLRSQ
ILTTLKEKPAPELSLSSQDLESVSKEDPKALAVALSWDIRKAETVQQACTTELALRLQQV
QSLHSLRNLSARRSPLPGEPQRPKRAKRDSS

[000138]The engineered murine ICAD sequence susceptible to specific and controlled inactivation and cleavage by TEV protease ($ICAD^{2xTEV}$) amino acid sequence is as follows (SEQ ID NO:1) (The amino acids added to provide TEV protease cleavage, specifically ENLYFQS are underlined):

MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPIT
LVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWIYNDSDGGTAWVSQESFEA<u>ENLYFQS</u>DSRA
GVKWKNVARQLKEDLSSIILLSEEDLQALIDIPCAELAQELCQSCATVQGLQSTLQQVLD
QREEARQSKQLLELYLQALEKEGNILSNQKESKAALSEEL<u>ENLYFQS</u>DTGVGREMASEVLLRSQ

ILTTLKEKPAPELSLSSQDLESVSKEDPKALAVALSWDIRKAETVQQACTTELALRLQQV
QSLHSLRNLSARRSPLPGEPQRPKRAKRDSS

[000139] ICAD^{2xTEV} nucleic acid sequence (SEQ ID NO:7) (*E. coli* codon-optimized TEV protease recognition sequences underlined)

ATGGAACTGTCACGCGGAGCTAGTGCACCCGACCCGGACGACGTGCGTCCGCTGAAACCGTGCCT
GCTCCGTCGCAACCACAGTCGTGATCAACATGGTGTGGCCGCGAGCAGCCTGGAAGAGTTGCGCA
GCAAGGCCTGCGAACTGCTGGCGATTGATAAAAGCTTGACCCCGATTACCTTGGTTCTGGCTGAGG
ACGGCACGATCGTGGATGACGACGACTACTTCCTGTGTCTGCCTAGCAACACCAAATTCGTGGCGT
TGGCCTGTAATGAAAAGTGGATTTATAACGATTCGGATGGCGGTACGGCTTGGGTTTCGCAAGAAAG
CTTTGAAGCGGAAAACCTGTATTTTCAGAGCGACAGCCGTGCCGGTGTCAAATGGAAAAACGTCGC
GCGTCAGCTGAAGGAGGACCTGTCAAGCATTATCCTGTTGTCCGAGGAGGACTTGCAAGCTCTGAT
CGACATCCCGTGCGCGGAGCTTGCGCAGGAGCTGTGCCAGAGCTGTGCGACCGTTCAGGGCCTG
CAATCCACCTTACAACAAGTGCTGGATCAGCGCGAAGAAGCACGTCAGAGCAAGCAGCTGCTGGAG
TTGTACCTGCAAGCGCTGGAGAAAGAGGGCAATATCCTGTCTAACCAGAAAGAGTCTAAGGCGGCAT
TAAGCGAAGAGTTAGAAAACCTGTATTTTCAGAGCGATACCGGTGTAGGTCGTGAAATGGCATCCGA
AGTTCTGCTTCGCTCTCAGATTCTGACCACCCTGAAGGAGAAGCCGGCACCGGAACTCTCCCTGTC
CTCTCAAGATTTGGAGAGCGTTAGCAAGGAGGATCCGAAAGCGTTGGCGGTGGCCTTGAGCTGGG
ATATCCGTAAAGCTGAGACTGTGCAGCAGGCGTGCACCACGGAACTTGCTTTACGTCTGCAACAAGT
TCAGTCCCTGCACAGCCTGCGCAATCTGTCTGCACGTCGTAGCCCGCTGCCGGGTGAACCGCAGC
GCCCAAAGCGCGCGAAACGCGACAGCAGT

[000140]As an alternative embodiment, human ICAD is engineered to be cleaved and activated specifically by TEV protease. In an embodiment of the invention, human and murine CAD and ICAD are utilized interchangeably. Human and murine ICAD show 77% amino acid identity in sequence. Human and murine CAD also show 77% amino acid identity in sequence. Meiss et al have previously reported use of expressed murine CAD and human ICAD, for example (Meiss et al (2001) Nucl Acids Res 29(19):3901-3909). Sequences of various CAD enzymes are known and available to one skilled in the art and these could be utilized in the invention. These include mouse/murine (Mus musculus, GenBank accession numbers AB009377, NM_007859), rat (Rattus norvegicus, GenBank accession number AF136598), human (Homo sapiens, GenBank accession numbers AF064019, AF039210, AB013918, NM_004402), zebrafish (Danio rerio, GenBank accession numbers AF286179) and fruitfly (Drosophila melanogaster, GenBank accession numbers AF149797, AB036773). In particular, primate and mammal CAD/ICAD pairs including mixed pairs would be preferred.

[000141]The human ICAD amino acid sequence is as follows (SEQ ID NO:4). Comparable caspase recognition sequences for deletion to eliminate caspase cleavage and replacement with an alternative enzyme recognition sequence, specifically DET and DAV, are in bold and underlined:

>sp|O00273|DFFA_HUMAN DNA fragmentation factor subunit alpha OS=Homo sapiens
MEVTGDAGVPESGEIRTLKPCLLRRNYSREQHGVAASCLEDLRSKACDILAIDKSLTPVT
LVLAEDGTIVDDDDYFLCLPSNTKFVALASNEKWAYNNSDGGTAWISQESFDV**DET**DSGA
GLKWKNVARQLKEDLSSIILLSEEDLQMLVDAPCSDLAQELRQSCATVQRLQHTLQQVLD
QREEVRQSKQLLQLYLQALEKEGSLLSKQEESKAAFGEEV**DAV**DTGISRETSSDVALASH
ILTALREKQAPELSLSSQDLELVTKEDPKALAVALNWDIKKTETVQEACERELALRLQQT
QSLHSLRSISASKASPPGDLQNPKRARQDPT

[000142]An engineered human ICAD sequence susceptible to specific and controlled inactivation and cleavage by TEV protease (hICAD^{2xTEV}) amino acid sequence is as follows (SEQ ID NO:2) (The amino acids added to provide TEV protease cleavage, specifically ENLYFQS are underlined):

MEVTGDAGVPESGEIRTLKPCLLRRNYSREQHGVAASCLEDLRSKACDILAIDKSLTPVT
LVLAEDGTIVDDDDYFLCLPSNTKFVALASNEKWAYNNSDGGTAWISQESFDV<u>ENLYFQS</u>DSGA
GLKWKNVARQLKEDLSSIILLSEEDLQMLVDAPCSDLAQELRQSCATVQRLQHTLQQVLD
QREEVRQSKQLLQLYLQALEKEGSLLSKQEESKAAFGEEV<u>ENLYFQS</u>DTGISRETSSDVALASH
ILTALREKQAPELSLSSQDLELVTKEDPKALAVALNWDIKKTETVQEACERELALRLQQT
QSLHSLRSISASKASPPGDLQNPKRARQDPT

[000143]In one or more embodiments, CAD enzyme is utilized in the methods of the invention and is combined with engineered ICAD. In an embodiment, the altered or engineered ICAD is expressed or otherwise combined with CAD. Sequences of various CAD enzymes are known and available to one skilled in the art and these could be utilized in the invention. An exemplary mouse/murine CAD amino acid sequence is as follows (SEQ ID NO:5):

Mouse CAD DFF-40 (DFFA) >sp|O54788|DFFB_MOUSE DNA fragmentation factor subunit beta OS=Mus musculus
MCAVLRQPKCVKLRALHSACKFGVAARSCQELLRKGCVRFQLPMPGSRLCLYEDGTEVTD
DCFPGLPNDAELLLLTAGETWHGYVSDITRFLSVFNEPHAGVIQAARQLLSDEQAPLRQK
LLADLLHHVSQNITAETREQDPSWFEGLESRFRNKSGYLRYSCESRIRGYLREVSAYTSM
VDEAAQEEYLRVLGSMCQKLKSVQYNGSYFDRGAEASSRLCTPEGWFSCQGPFDLESCLS
KHSINPYGNRESRILFSTWNLDHIIEKKRTVVPTLAEAIQDGREVNWEYFYSLLFTAENL
KLVHIACHKKTTHKLECDRSRIYRPQTGSRRKQPARKKRPARKR

[000144]The human CAD amino acid sequence is as follows (SEQ ID NO:6):

Human CAD (DFF-40)DFFB >sp|O76075|DFFB_HUMAN DNA fragmentation factor subunit beta OS=Homo sapiens
MLQKPKSVKLRALRSPRKFGVAGRSCQEVLRKGCLRFQLPERGSRLCLYEDGTELTEDYF
PSVPDNAELVLLTLGQAWQGYVSDIRRFLSAFHEPQVGLIQAAQQLLCDEQAPQRQRLLA
DLLHNVSQNIAAETRAEDPPWFEGLESRFQSKSGYLRYSCESRIRSYLREVSSYPSTVGA
EAQEEFLRVLGSMCQRLRSMQYNGSYFDRGAKGGSRLCTPEGWFSCQGPFDMDSCLSRHS
INPYSNRESRILFSTWNLDHIIEKKRTIIPTLVEAIKEQDGREVDWEYFYGLLFTSENLK
LVHIVCHKKTTHKLNCDPSRIYKPQTRLKRKQPVRKRQ

[000145]In another embodiment of the invention mutant or variant CAD enzyme may be utilized, provided it is active enzymatically. Meiss et al have described CAD mutants altered in one or more of the

CAD conserved histidine residues, His127, His242, His263, His304, His308 and His313 and that these variants, while they complex stably with ICAD, they exhibit decrease in catalytic activity (Meiss et al (2001) Nucl Acids Res 29(19):3901-3909). Mutants altered at amino acids other than these His residues are contemplated.

[000146]Although ENLYFQS is the optimal sequence, with the protease cleaving between Q and S, the protease is active to a greater or lesser extent on a range of substrates (i.e. shows some substrate promiscuity). The highest cleavage is of sequences closest to the consensus EXLYΦQ\φ where X is any residue, Φ is any large or medium hydrophobe and φ is any small hydrophobic or polar residue.

[000147]In a preferred consensus recognized sequence of ENLYFQX, X is particularly preferred as S, although G is also recognized, as well as A, M, C or H. This provides other variant

[000148]In an embodiment of the invention, the ENLYFQS amino acids replaces each of the caspase recognition sequences in the engineered ICAD. In an embodiment of the invention, the ENLYFQG amino acids replaces each of the caspase recognition sequences in the engineered ICAD. In an embodiment either the ENLYFQS or the ENLYFQG amino acids replace each of the caspase recognition sequences in the engineered ICAD. In an embodiment of the invention, the ENLYFQS amino acids replaces both of the caspase recognition sequences DEP and DAV in the engineered ICAD, in the instance of murine ICAD. In an embodiment of the invention, the ENLYFQS amino acids replaces both of the caspase recognition sequences DET and DAV in the engineered ICAD, in the instance of human ICAD. In an embodiment of the invention, the ENLYFQG amino acids replaces both of the caspase recognition sequences DEP and DAV in the engineered ICAD, in the instance of murine ICAD. In an embodiment of the invention, the ENLYFQG amino acids replaces both of the caspase recognition sequences DET and DAV in the engineered ICAD, in the instance of human ICAD.

[000149]In other embodiments, one or more of the sequences selected from ENLYFQX, wherein X is S, G, A, M, C or H is inserted to replace one or more, either or both caspase recognition sequences.

[000150]Native TEV protease or variant TEV protease and sequences thereof are contemplated, provided they are active and capable of cleaving at the relevant inserted TEVp amino acid sequence, such as any TEVp recognition sequences inserted in an engineered ICAD of the invention. Altered versions and variants of TEVp have been generated. These include an S219V mutation which abolishes self-cleavage (autolysis) and other variant versions which remain active in the presence or absence of reducing agent (Kapust RB et al (2001) Protein Eng 14(12):993-1000). Directed evolution of TEVp has been utilized to improve the catalytic efficiency of TEVp, including TEV-S153N (Sanchez MI and Ting AY (2020) Nature Methods 17:167-174). Other TEVp variants have been generated with multiple mutations to improve solubility and reduce self-cleavage thereby enhancing enzymatic activity, including S219N and

S219V mutants in the background of T17S, N68D and I77V mutations (Nam H et al (2020) FEBS Open Bio 10(4):619-626).

[000151]In an embodiment, engineered ICAD and CAD are combined at about equal amounts to form an inactive ICAD/CAD heterodimer. In an embodiment, engineered ICAD of SEQ ID NO: 1 or 2 is expressed or otherwise combined with CAD. In an embodiment, engineered ICAD of SEQ ID NO: 1 or 2 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5 or 6. In an embodiment of the invention, human and murine CAD and ICAD are utilized interchangeably. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5 or 6. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5. In an embodiment, engineered ICAD of SEQ ID NO: 1 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 6. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5 or 6. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 5. In an embodiment, engineered ICAD of SEQ ID NO: 2 is expressed or otherwise combined with CAD of sequence SEQ ID NO: 6.

[000152]In another embodiment of the invention mutant or variant CAD enzyme may be utilized, provided it is active enzymatically. In an embodiment of the invention mutant or variant CAD enzyme may be utilized, provided it is active enzymatically and capable of forming a heterodimer with ICAD, particularly with the engineered ICAD hereof. Therefore, the CAD enzyme sequence, such as that of SEQ ID NO: 5 or SEQ ID NO:6 may be a variant sequence, particularly a variant sequence having 80%, 85%, 89%, 90%, 95%, 97% or 99% amino acid sequence identity to SEQ ID NO:5 or SEQ ID NO:6. If an alternative animal or mammalian CAD sequence as known, for example a native rat or other mammalian ICAD sequence, is utilized, variants thereof are also contemplated, including variant sequence having 80%, 85%, 89%, 90%, 95%, 97% or 99% amino acid sequence identity to the alternative animal or mammalian CAD sequence.

[000153]In embodiments of the invention, nucleic acids encoding the engineered ICAD and CAD are provided. In an embodiment, the encoding nucleic acid is codon optimized for expression, for example in a particular cell type, for example in bacteria such as E.coli.

[000154]Vectors or plasmids comprising the encoding nucleic acid are also provided as embodiments. IN an embodiment, the vector or plasmid corresponds to vector provided in FIGURE 5 or sequence set out in SEQ ID NO: 14. In an embodiment, the altered or engineered ICAD is co-expressed with CAD in the same vector or plasmid. In an embodiment, the altered or engineered ICAD is co-expressed with CAD

via the same promoter. Suitable promoters will provide ample expression and approximately equal expression of ICAD and of CAD. An exemplary promoter includes an inducible promoter. In an embodiment, an IPTG-inducible promoter is utilized.

[000155]In particular aspects, the ICAD and/or the CAD of the invention and of use herein is tagged or otherwise includes a label or other means for purification, such as by a protein or molecule having affinity for or otherwise recognizing the tag or label. In an embodiment, a label or tag may include strepdavidin, biotin, GFP etc. In an embodiment, one or more linker, such as between the tag or label, may be added and included. In an embodiment, a recognition sequence for cleavage to remove the one or more tag or labels may also be included. This provides for cleavage and specific separation of the ICAD or CAD, such as of the one or more tag or labels, after initial expression so that a purified ICAD and/or CAD is provided free of alternative or added N terminal or C terminal sequence.

[000156]In one embodiment, the proximally ligated, hybridized or joined nucleic acids are labeled and can be detected by detecting one or more labels attached to the sample nucleic acids. The labels can be incorporated by any of a number of methods. In one example, the label is simultaneously incorporated during an amplification step, ligation step, or otherwise in the preparation of the sample nucleic acids. Thus, for example, polymerase chain reaction (PCR) with labeled primers or labeled nucleotides will provide a labeled amplification product. In one embodiment, transcription amplification, as described above, using a labeled nucleotide (such as fluorescein-labeled UTP and/or CTP) incorporates a label into the transcribed nucleic acids.

[000157]In one embodiment, the genomic DNA of the cells used for CAD-C is labeled prior to fixation using labeled nucleotide analogues such as ethinyl-deoxyuridine (EdU) incorporation or radiolabeled nucleotides, and is then labeled with biotin or another label using click chemistry and enriched using streptavidin or other compatible capture following the standard CAD-C protocol for enrichment, or visualized after proximity ligation and DNA isolation.

[000158]Detectable labels suitable for use include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels include biotin for staining with labeled streptavidin conjugate, magnetic beads (for example DYNABEADS™), fluorescent dyes (for example, fluorescein, Texas red, rhodamine, green fluorescent protein, and the like), radiolabels (for example, 3H, 1251, 35S, 14C, or 32P), enzymes (for example, horseradish peroxidase, alkaline phosphatase and others commonly used in an ELISA), and colorimetric labels such as colloidal gold or colored glass or plastic (for example, polystyrene, polypropylene, latex, etc.) beads.

[000159]Means of detecting such labels are also well known. Thus, for example, radiolabels may be detected using photographic film or scintillation counters, fluorescent markers may be detected using a

photodetector to detect emitted light. Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate, and colorimetric labels are detected by simply visualizing the colored label. The label may be added to the target (sample) nucleic acid(s) prior to, or after, the hybridization. So-called "direct labels" are detectable labels that are directly attached to or incorporated into the target (sample) nucleic acid prior to hybridization. In contrast, so-called "indirect labels" are joined to the hybrid duplex after hybridization. Often, the indirect label is attached to a binding moiety that has been attached to the target nucleic acid prior to the hybridization. Thus, for example, the target nucleic acid may be biotinylated before the hybridization. After hybridization, an avidin-conjugated fluorophore will bind the biotin bearing hybrid duplexes providing a label that is easily detected.

[000160]In an embodiment of the method, systems or kits hereof, a or the tag is at least one of a GFP tag, Myc tag, HA tag, V5-tag, CD tag, or FLAG tag, and combinations thereof. In another embodiment of the method, system or kits hereof, a or the tag is a peptide and protein affinity tag including at least one of SPYtag, CBP-tag, GST-tag, poly His-tag, SNAP-tag, CDB tag, Halo tag, Avitag, S-tag, or Strep-tag, and combinations thereof. In an embodiment of the method, system or kits hereof, at least one nanobody, scFv or Fab has affinity for a protein or a tag of the protein, the protein being associated with at least one other protein that bind to DNA. The protein is one of the biological target molecules.

[000161]A purified and isolated engineered ICAD and CAD are provided herein. These may be provided as heterodimers for further activation of CAD by the alternative protease. Compositions of the isolated engineered ICAD, of CAD, and of engineered ICAD/CAD heterodimers are provided herein

[000162]In accordance with the invention, CAD-C methods are provided and utilized for determination of organizational structure of chromatin and chromosomes in three dimensions, including whereby information as to the structure and sequence is determined. In accordance with the invention, kits including components for CAD-C methods are provided for determination of organizational structure of chromatin and chromosomes in three dimensions, including whereby information as to the structure and sequence is determined.

[000163]In an embodiment, the methods are in line with the Hi-C methods or with the Micro-C methods, adapted for use and with use of the engineered ICAD of the invention and CAD enzyme, or of activated CAD which has been activated by an alternative enzyme as provided herein.

[000164]A method is provided for capturing the sequence and structure (three dimensional conformation) of one or more target regions of one or more genome in a sample comprising:

(a) crosslinking chromatin from a sample to preserve the genome sequence and structure;

(b) introducing a CAD enzyme and digesting the crosslinked chromatin with CAD to generate digested ends of DNA, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase;

(c) labeling the digested ends of DNA;

(d) ligating the digested and labeled ends of DNA to generate ligated DNA; and

(e) purifying the ligated DNA to generate proximally-ligated DNA.

[000165]In some embodiments, the digested ends of DNA are not labeled. IN one such embodiment, ligating the digested ends proceeds without first labeling the ends. In an embodiment, the method comprises:

(a) crosslinking chromatin from a sample to preserve the genome sequence and structure;

(b) introducing a CAD enzyme and digesting the crosslinked chromatin with CAD to generate digested ends of DNA, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase;

(d) ligating the digested ends of DNA to generate ligated DNA; and

(e) purifying the ligated DNA to generate proximally-ligated DNA.

[000166]In an embodiment, the method further comprises:

(f) fragmenting the proximally-ligated DNA; and

(g) capturing the proximally-ligated DNA.

[000167] In embodiments of the method, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8) or ENLYFQG (SEQ ID NO:9). In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

[000168] In some embodiments, the ICAD is an engineered murine ICAD or is an engineered human ICAD. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2, or variants thereof retaining the TEV protease recognition sequence and having up to 90% amino acid sequence identity with SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1.

[000169] In some embodiments, the CAD is a murine or human CAD. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6, or variants thereof retaining CAD endonuclease activity and having up to 90% amino acid sequence identity with SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5.

[000170] In an embodiment of the method, in step (g) the capturing is achieved by enriching for proximally-ligated DNA via the label of (c). In an embodiment of the method, in step (g) the capturing is achieved by size selection of the fragments. In other embodiments, the proximally-ligated DNA is sequenced. In other embodiments, the proximally-ligated DNA is size selected. In an embodiment, the crosslinking of step (a) utilizes UV. In an embodiment, the crosslinking of step (a) utilizes formaldehyde. In an embodiment, the crosslinking of step (a) utilizes formaldehyde and disuccinimidyl glutarate (DSG). Other methods for crosslinking are suitable and will be known and available to one skilled in the art.

[000171] In an embodiment, the purified or captured proximally-ligated DNA is used to prepare a library of the proximally-ligated DNA. In an embodiment, the library of the proximally-ligated DNA is sequenced in its entirety or in part. In an embodiment, the library of the proximally-ligated DNA is sequenced to assess the library's quality in terms of library complexity and long-range interaction information prior to proceeding with the capture. Methods for library preparation and suitable library systems are known and available to one skilled in the art.

[000172] In some embodiments, the purified or captured proximally-ligated DNA is enriched using one or more probe directed against or specific for one or more target or sequence of interest.

[000173] In some embodiments, the purified or captured proximally-ligated DNA is enriched using an incorporated modified nucleotide that marks genomic regions with particular replication timing.

[000174] In some embodiments of the method, the fragmentation of (f) uses mechanical methods. In one embodiment, the mechanical means is sonication. Other methods and means for fragmentation, such as enzymatic fragmentation, will be known and available to one skilled in the art. In one such embodiment, for example, frgmentase is utilized for fragmentation such as that provided commercially (such as the NEBNext UltraExpress FS kit). In some embodiments of the method, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of <1 kb. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of 500-800bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of 500-600bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of about 500bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of 300-600 bp. In an embodiment, the fragmentation of (f) fragments DNA to

generate fragments of a fragment size of greater than 300bp and less than 1000 bp. In an embodiment, the fragmentation of (f) fragments DNA to generate fragments of an average fragment size >400bp.

[000175] In some embodiments, the purifying of (e) and/or the enriching of (g) is via beads. In some embodiments, the purifying of (e) and/or the enriching of (g) is via magnetic beads.

[000176] In an aspect of the invention, the engineered ICAD/CAD heterodimers are activated by one or more alternative protease prior to combining with a sample for evaluation. In an aspect of the invention, the engineered ICAD/CAD heterodimers are activated by one or more alternative protease upon or shortly after combining with a sample for evaluation.

[000177] In an aspect, a method is provided for detecting special proximity relationships between nucleic acid sequences comprising contacting a sample containing nucleic acid with a CAD enzyme and digesting the nucleic acid with CAD to generate digested ends of nucleic acid, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase.

[000178] The invention provides a system and kit for determining sequence and structure (three dimensional conformation) of one or more target regions or of one or more genome in a sample comprising nucleic acid, comprising an engineered ICAD and CAD, wherein the engineered ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence, and further comprising an alternative protease to inactivate the engineered ICAD and activate the CAD such that the nucleic acid is specifically digested by the CAD.

[000179] In an embodiment, the system or kit further comprises:

(a) a means for labeling the nucleic acid digested by the CAD;

(b) a ligation enzyme for ligating the labeled digested nucleic acid to generate ligated nucleic acid; and

(c) a means for purifying the ligated nucleic acid to generate proximally-ligated nucleic acid;

(d) a means for capturing the proximally-ligated nucleic acid.

[000180] In an embodiment, the system or kit further comprises reagents for sequencing the proximally-ligated nucleic acid. In an embodiment, the system or kit further comprises reagents for generating a library of the proximally-ligated nucleic acid. In some embodiments, reagents for sequencing the proximally-ligated nucleic acid in the library are also included.

[000181] In embodiments of the methods, systems or kits, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with alternative protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence. In an embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8) or ENLYFQG (SEQ ID NO:9). In an

embodiment, the ICAD is an engineered ICAD wherein caspase enzyme recognition sequence is deleted and replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

[000182] In some embodiments, the ICAD is an engineered murine ICAD or is an engineered human ICAD. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1 or SEQ ID NO:2, or variants thereof retaining the TEV protease recognition sequence and having up to 90% amino acid sequence identity with SEQ ID NO:1 or SEQ ID NO:2. In some embodiments, the ICAD corresponds to a polypeptide having SEQ ID NO:1.

[000183] In some embodiments, the CAD is a murine or human CAD. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5 or SEQ ID NO:6, or variants thereof retaining CAD endonuclease activity and having up to 90% amino acid sequence identity with SEQ ID NO:5 or SEQ ID NO:6. In some embodiments, the CAD corresponds to a polypeptide having SEQ ID NO:5.

[000184] In accordance with the methods, kits and systems provided herein, a sample for evaluation or assessment may include eukaryotic cells derived from cell cultures, tissues, patient samples, or whole organisms. Prokaryotic cells or archae may also be suitable. Samples of cell free DNA are also contemplated.

[000185] In some embodiments, determining the sequence includes using a probe that specifically binds to the junction at the site of the two joined nucleic acid fragments, as in the two ends of the proximally-ligated nucleic acids. In particular embodiments, the probe specifically hybridizes to the junction both 5' and 3' of the site of the join and spans the site of the join. A probe that specifically binds to the junction at the site of the join can be selected based on known interactions, for example in a diagnostic setting where the presence of a particular target junction, or set of target junctions, has been correlated with a particular disease or condition. It is further contemplated that once a target junction is known, a probe for that target junction can be synthesized.

[000186] In some embodiments, the end joined nucleic acids are selectively amplified. In some examples, to selectively amplify the end joined nucleic acids, a 3' DNA adaptor and a 5' RNA, or conversely a 5' DNA adaptor and a 3' RNA adaptor can be ligated to the ends of the molecules can be used to mark the end joined nucleic acids. Using primers specific for these adaptors only end joined nucleic acids will be amplified during an amplification procedure such as PCR. In some embodiments, the target end joined nucleic acid is amplified using primers that specifically hybridize to the adaptor nucleic acid sequences present at the 3' and 5' ends of the end joined nucleic acids. In some embodiments, the non-ligated ends of the nucleic acids are end repaired. In some embodiments attaching sequencing adapters to the ends of the end ligated nucleic acid fragments.

[000187] In some embodiments, the cells are lysed to release the cellular contents, for example after crosslinking. In some examples the nuclei are lysed as well, while in other examples, the nuclei are maintained intact, which can then be isolated and optionally lysed, for example using an reagent that selectively targets the nuclei or other separation technique known in the art.

[000188] In some examples, the sample is a sample of permeablized nuclei, multiple nuclei, isolated nuclei, synchronized cells, (such at various points in the cell cycle, for example metaphase) or acellular. In some embodiments, the nucleic acids present in the sample are purified, for example using ethanol precipitation. In example embodiments of the disclosed method the cells and/or cell nuclei are not subjected to mechanical lysis. In some example embodiments, the sample is not subjected to RNA degradation. In specific embodiments, the sample is not contacted with an exonuclease to remove of biotin from un-ligated ends. In some embodiments, the sample is not subjected to phenol/chloroform extraction.

[000189] In some embodiments of the disclosed method the nucleic acids present in the cell or cells are fixed in position relative to each other by chemical crosslinking, for example by contacting the cells with one or more chemical cross linkers. This treatment locks in the spatial relationships between portions of nucleic acids in a cell. Any method of fixing the nucleic acids in their positions can be used. In some embodiments, the cells are fixed, for example with a fixative, such as an aldehyde, for example formaldehyde or gluteraldehyde. In some embodiments, a sample of one or more cells is cross-linked with a cross-linker to maintain the spatial relationships in the cell. For example, a sample of cells can be treated with a cross-linker to lock in the spatial information or relationship about the molecules in the cells, such as the DNA and RNA in the cell. In other embodiments, the relative positions of the nucleic acid can be maintained without using crosslinking agents. For example the nucleic acids can be stabilized using spermine and spermidine. Other methods of maintaining the positional relationships of nucleic acids are known in the art.

[000190] In some embodiments, nuclei are stabilized by embedding in a polymer such as agarose. In some embodiments, the cross-linker is a reversible cross-linker. In some embodiments, the cross-linker is reversed, for example after the fragments are joined. In specific examples, the nucleic acids are released from the cross-linked three-dimensional matrix by treatment with an agent, such as a proteinase, that degrade the proteinaceous material form the sample, thereby releasing the end ligated nucleic acids for further analysis, such as determination of the nucleic acid sequence. In specific embodiments, the sample is contacted with a proteinase, such as Proteinase K.

[000191] In some embodiments of the disclosed methods, the cells are contacted with a crosslinking agent to provide the crosslinked cells. In some examples, the cells are contacted with a protein-nucleic acid crosslinking agent, a nucleic acid nucleic acid crosslinking agent, a protein-protein crosslinking agent or

any combination thereof. By this method, the nucleic acids present in the sample become resistant to special rearrangement and the spatial information about the relative locations of nucleic acids in the cell is maintained. In some examples, a cross-linker is a reversible-, such that the cross-linked molecules can be easily separated in subsequent steps of the method. In some examples, a crosslinker is a non-reversible cross-linker, such that the crosslinked molecules cannot be easily separated. In some examples, a cross-linker is light, such as UV light. In some examples, a cross linker is light activated. These crosslinkers include formaldehyde, disuccinimidyl glutarate, UV light, psoralens and their derivatives such as aminomethyltrioxsalen, glutaraldehyde, ethylene glycol bis[succinimidysuccinate], bissulfosuccinimidy suberate, 1-Ethy 1-3-[3-dimethy laminopropy] carbodiimide (EDC) bis [sulfosuccinimidyl] suberate (BS3) and other compounds known to those skilled in the art.

[000192] In some embodiments, the proximity ligation step is omitted to create a genome-wide CAD cleavage control that reveals the efficiency of the cleavage reaction, or to perform CAD-based nucleosome mapping or chromatin footprinting experiments.

[000193] The disclosed methods are also particularly suited to monitoring disease states, such as disease state in an organism, for example a plant or an animal subject, such as a mammalian subject, for example a human subject. Certain disease states may be caused and/or characterized by the differential formation of certain target joins. For example, certain interactions may occur in a diseased cell but not in a normal cell. In other examples, certain interactions may occur in a normal cell but not in diseased cell. Thus, using the disclosed methods a profile of the interaction between DNA sequences in vivo, can be correlated with a disease state. The target joined profile correlated with a disease can be used as a "fingerprint" to identify and/or diagnose a disease in a cell, by virtue of having a similar "fingerprint." In addition, the profile can be used to monitor a disease state, for example to monitor the response to a therapy, disease progression and/or make treatment decisions for subjects. The ability to obtain an interaction profile allows for the diagnosis of a disease state, for example by comparison of the profile present in a sample with the correlated with a specific disease state, wherein a similarity in profile indicates a particular disease state.

[000194] Accordingly, aspects of the disclosed methods relate to diagnosing a disease state based on target junction profile correlated with a disease state, for example cancer, or an infection, such as a viral or bacterial infection. It is understood that a diagnosis of a disease state could be made for any organism, including without limitation plants, and animals, such as humans.

[000195] The disclosure now refers to Examples for more specific features, which should not be used to unduly limit the embodiments of the disclosure.

EXAMPLE 1

[000196] The last two decades of work on chromosome conformation in eukaryotic nuclei have revealed a complex and highly regulated hierarchy of architectural features, from self-associating domains and compartmental interactions to locus-specific loops. Recent findings have shown that these structures are dynamic and heterogeneous, with emerging insights into the factors that shape them and implications for the control of transcription and other nuclear processes (Soroczynski J and Risca VI (2023) Curr Opinion in Cell Bio 84:102211; doi.org/10.1016/j.ceb.2023.102211).

[000197] We wanted to develop a method that would allow us to study chromosome architecture at all scales in as many sample types as possible. We identified chromatin fragmentation as the primary bottleneck in the 3C methods. Chromosome fragmentation is a key bottleneck in proximity ligation methods, resolution and sensitivity of measurement relies on fragmentation and generating ligatable ends.

[000198] Hi-C detects compartments well. Micro-C detects loops well. No method measures all chromosome folding features well simultaneously. In contrast to the rapid advancements that have been made in DNA sequencing technologies, MNase has been the workhorse of chromatin fragmentation for decades. However, while MNase provides high resolution, it does so at a cost and its problems and issues include over-digestion, AT-bias, and unligatable DNA ends.

[000199] With Hi-C methods utilizing restriction enzyme fragmentation, no enzyme titration is necessary, and the process can be taken to end-point digestion. Hi-C provides easily ligatable DNA ends. Chromatin SDS solubilization is required. Hi-C generates non-uniform genome fragmentation and is applicable at > 1 kb resolution. Micro-C methods using MNase Fragmentation provide for the possibility of high-resolution (< 1kb), can generate footprinting information and no chromatin solubilization is necessary. Arduous titration is necessary, however, and one cannot use precious samples. Frayed DNA stubs lead to inefficient ligation. MNase is also AT-biased and destroys fragile sites.

[000200] Therefore, we sought to develop an alternative approach using an enzyme and modifying the enzyme to provide an enzyme and system that does not have the issues with MNase. We wanted to identify an approach and nuclease which would provide specific restriction-enzyme-like end-point digestion, that did not require chromatin stabilization, that provided predictable and reproducible footprinting, that gave high resolution with no/without DNA sequence bias, and that generated ready-to-ligate ends to provide high efficiency especially for precious samples.

[000201] We chose to turn to a mammalian Caspase Activated DNase or CAD. CAD (also denoted DFF-40, DFFB) is a major apoptotic nuclease in most metazoans. CAD evolved to uniformly fragment chromosomes during the onset of apoptosis. Apoptosis is a programmed cell death module resulting in the removal of unwanted cells from an organism and genomic DNA fragmentation accompanies apoptosis. Caspase Activated DNase or CAD is closely associated with this function, including in apoptosis. Caspase 3 activates CAD by proteolytic inactivation of the inhibitor ICAD. Nuclease activity of CAD is restricted

by association with its inhibitor ICAD. The association of CAD with ICAD begins with translation of CAD mRNA where ICAD acts as a chaperone to ensure CAD is properly folded. Inactivating cleavage of ICAD by activated caspases (particularly caspase-3) releases CAD from inhibition and facilitates CAD homodimerization (FIGURE 2A; Larsen BD and Sorensen CS (2017) The FEBS Journal 284:1160-1170).

[000202] CAD has various properties that make it unique and ideal for methods and applications to explore, evaluate and characterize chromosome architecture on various scales. CAD has the unique property of being a strict endonuclease, which means it retains the easy end point digestion property of restriction enzymes but the high resolution of MNase, with lesser DNA sequence bias. CAD is a dsDNA endonuclease that produces mostly blunt DNA ends (5'-phoshorylated, 3'-OH). CAD is non-processive and has no exonuclease activity. Further CAD cleaves chromatin strictly between nucleosomes to produce uniformly-sized mono-nucleosomes. These properties are depicted in FIGURE 3.

[000203] CAD is ordinarily activated by caspase cleavage of its inhibitor ICAD. We developed a strategy for engineering the murine Caspase Activated DNase (CAD) proenzyme complex for orthogonal activation by Tobacco Etch Virus protease (TEVp), to enable its easy activation *in vitro*. We re-engineered ICAD to be activated by recombinant TEV protease Specifically, we partially deleted the two caspase cleavage recognition sequences in ICAD and inserted residues to reconstitute TEVp recognition sequences that can be cleaved proximally to the original caspase cleavage sites, referred to as $ICAD^{2xTEV}$.

[000204] We synthesized *E. coli* codon-optimized *Mus musculus* CAD and ICAD coding sequences, and cloned them into a modified pRSFDuet™ dual cassette protein co-expression vector (Schilling et al, 2020). The robust co-expression of CAD and ICAD is essential for formation of the functional proenzyme, as the subunits of heterodimer chaperone each other's folding (Sakahira et al, 1999). We placed the affinity and stability tags on the N-terminus of the CAD ORF, reasoning that we should recover the CAD:ICAD proenzyme heterodimer by pulling down on CAD alone, due to their tight binding affinity (McCarty et al, 1999). We used a dual tag approach, adding a Twin-Strep-tag followed by muGFP, a highly-stable variant of GFP (Scott et al, 2018) and $SUMO^{Eu1}$ (Vera Rodriguez et al, 2019), to promote solubility and facilitate over-expression of CAD in *E. coli*. To this end, we were able to obtain robust soluble over-expression of TwinStrepTag-muGFP-$SUMO^{Eu1}$-CAD and $ICAD^{2xTEV}$ in *E. coli*. Hereafter, for simplicity, we refer to the recombinant TEVp-activatable proenzyme heterodimer as 'CAD'.

[000205] We purified the CAD proenzyme by Strep-tag affinity chromatography, followed by size exclusion chromatography (SEC). A majority of the intact TwinStrepTag-muGFP-$SUMO^{Eu1}$-$CAD:ICAD^{2xTEV}$ complex protein eluted at the void volume of the SEC column, indicating the CAD proenzyme was part of a large supramolecular complex. We analyzed fractions of the SEC elution corresponding to smaller complexes by SDS-PAGE, which showed that the void volume complex contained a majority of intact full-length CAD and ICAD subunits, whereas the other fractions contained

truncated subunits. We next tested the susceptibility of the ICAD^{2xTEV} present in the void volume complex to TEVp cleavage. We envisaged this would serve as a proxy for testing whether the proenzyme heterodimer was natively folded, or instead present in insoluble aggregates which would be expected to be TEVp-inaccessible. Encouragingly, ICAD^{2xTEV} was readily digested by TEVp into expectedly sized digestion fragments (Figure 6). We therefore concluded that the supramolecular complex likely represents binding of CAD proenzyme molecules to *E. coli* DNA, as previously reported (Widlak P (2006) Biochem Cell Biol 84(\$):405-410).

[000206] To verify the intended activity of the engineered proenzyme, we assayed the specific TEVp-activatable DNase activity of the purified CAD. To measure DNase activity, we incubated aliquots of CAD SEC eluate fractions with linear, double-stranded Lambda phage genomic DNA (λ gDNA) to serve as the assay substrate and resolved the DNA products by agarose gel electrophoresis (Methods). CAD proenzyme from void volume fractions showed little background DNase activity, however, upon the addition of TEVp CAD rapidly digested λ gDNA to species shorter than 100 bp in length (Figure 7). Our engineered CAD possessed specific DNase activity comparable to previously reported activity of wild-type murine enzyme, 2 µg of engineered CAD proenzyme digests 25 µg of λ gDNA in 20 minutes incubation (Methods).

[000207] Coordinated expression and use of CAD with a novel protease susceptible mutant ICAD provides a system for controlled and specific activation of CAD endonuclease activity. We have applied this controlled and activatable CAD to the chromatin fragmentation step of a proximity ligation protocol with very good results, including as described below. We have named this novel method CAD-C.

[000208] We optimized the expression and purification of the engineered CAD proenzyme in *E. coli*, to enable orthogonal TEVp activation of CAD DNase activity *in vitro* (FIGURES 2, 3, 4, 6 and 7). CAD was engineered for orthogonal *in vitro* activation by the TEV protease (TEVp). Murine CAD (DFFB) and ICAD (DFFA) was utilized. ICAD is ordinarily cleaved by caspase. The caspase recognition sites in murine ICAD were deleted and TEV cleavage sites were inserted in replacement (FIGURE 4). This is in contrast to previously reported ICAD variants where the caspase recognition amino acids were retained and TEV recognition amino acids added as neighboring amino acids (Xiao, Widlak, and Garrard 2007). The specific amino acid sequence alterations are depicted in FIGURE 4.

[000209] CAD/ICAD^{2XTEV} was purified as a proenzyme heterodimer including from E.coli expression. High quality proenzyme was obtained from small and large scale preparations (50-500ml). The purified enzyme is stable in 4°C for at least two months without an appreciable decline in activity. In vitro CAD activity was tested on naked DNA with a λ phage genomic high molecular weight (HMW) DNA test substrate (48.5kb) (data not shown). We define the TEVp-dependent DNase activity of CAD digestion unit (U) as the amount of enzyme capable of complete digestion of high-molecular weight λ genomic DNA

(48.5 kb size) (Figure 6). One microliter of TST-muGFP-SUMO-CAD-ICAD^{2xTEV} stock at 2 µg/µL has roughly 25 U activity, which is sufficient for completely digesting 25 µg of λ gDNA in 20 minutes, demonstrating the high TEV-p specific activity (Figure 7) (see Material and Methods). TEVp activation increases CAD enzyme activity approximately 100-fold, and there is low non-specific nuclease contamination of the ICAD^{2XTEV} – CAD proenzyme. CAD digestion of 1% formaldehyde-fixed human K562 cell line nuclei produces a sharp nucleosomal laddering pattern, consistent with CAD's lack of exonuclease activity (data not shown). We demonstrated the lack of chromatin over-digestion, a hallmark of MNase (Figure 3), in human crosslinked cells under conditions commonly used for Hi-C and Micro-C assays (3 mM DSG + 1% FA), by digesting hTERT-RPE1 nuclei using a 100-fold range of [CAD] (Figure 9). CAD activity on native and fixed chromatin in human K562 and GM12878 cells was also evaluated and demonstrated with similar results (Figure 8 and data not shown).

[000210] The effects of CAD concentration on digestion properties was evaluated and digestion behavior across a 100 fold range of CAD concentration was assessed. Results from hTERT-RPE-1 cells are shown in FIGURE 9 and 10.

[000211] We detect low intrinsic GC-sequence-content cleavage site bias when assayed from sequencing data using CAD under-digestion conditions, including across 100 fold concentration range of CAD (FIGURE 11 and 12). CAD has preference for periodic pyrimidine/purine sequences and balanced GC preferences.

[000212] CAD digestion sensitivity was evaluated for various chromosome milieus, again across 100 fold concentration range of CAD. ATAC-sequence narrow peaks – enhancers, transcribed loci – were evaluated. Transcription start sites (TSSs) were evaluated. Also, epigenetic domains annotated by histone modification CUT&Tag were assessed. The results and profiles across 100 fold concentrations were consistent and similar (data not shown).

[000213] One issue with MNase is its use and application for footprinting chromosome bound proteins, because no single MNase digestion condition can capture all footprints. MNase digestion ordinarily proceeds until no DNA is left. Typically, this requires sequencing multiple MNase digestion point samples to provide applicable data and either CTCF is preserved and nucleosomes are poorly resolved, or mono-nucleosomes are resolved but CTCF-bound DNA is completely digested and lost. In sharp contrast, we have demonstrated that CAD digestion provide fine footprinting details and both nucleosome and CTCF footprints are distinguished (FIGURE 13). Also, ATAC-seq peaks, TSS and TES features (in <100bp fragments and 100-200bp fragments) are evaluated and identified (data not shown). CAD was also shown to be applicable and consistent for use in measuring nucleosome positioning. Uniform mono-nucleosome fragment size improves identification of well-positioned nucleosomes and quantifying their phasing with CAD. In contrast, MNase produces a wider size distribution of mono-nucleosome fragments, is non-

uniform across chromosome AT isochores, and introduces noise which requires deep sequencing (data not shown).

[000214] CAD shows efficient proximity ligation, particularly in comparison to the ligation of Micro-C. Micro-C ligation does not proceed beyond di-nucleosomes. Therefore in a Micro-C protocol it is necessary to size select for di-nucleosomes before pulldown (e.g. with biotin pulldown) and library preparation. CAD-C ligation proceeds beyond di-nucleosomes, producing concatemers. This provides a modified protocol more like with classic Hi-C in CAD-C, where in the protocol one sonicates to ~200bp, then pulldown (e.g. with biotin pulldown), library preparation, and direct sequencing 2x150 across ligation junction to recover footprinting information.

[000215] CAD-C contact probability curves are also unique from Micro-C and Hi-C. In CAD-C we have detected more ultra-short (<1kb) and ultra-long range (>10 Mb) contacts. Parallel processed ICE balanced 500 bp resolution maps demonstrate this important difference and distinction (see FIGURE 14).

[000216] We have recently obtained the first deep chromosome conformation capture data using CAD-C. Data show that we capture the entire dynamic range from 500 bp to inter-chromosomal contacts, with a streamlined protocol. Iteratively-corrected (ICE balanced) contact probability curve and its derivative (FIGURE 14) were calculated and plotted as described in Oksuz et al. 2021, showing the unique ability of CAD-C to capture contact features typically uniquely seen separately in Micro-C and Hi-C assays.

[000217] CAD-C detects strong inter-chromosomal contacts between active loci. ATAC-seq and H3K27 ac-enriched compartments are demonstrated and evaluable on mumerous chromosomes (Chr 10, 11, 12, 15, 16, 18 and 19 – data not shown).

[000218] CAD-C detects strong intra-chromosomal constitutive heterochromatin compartments. CAD-C contact matrix map of chromosome 5 at 200 kb resolution shows strong chequering, indicative of compartmentalization. H3K9me3-modified chromatin domains form unique dense contacts.

[000219] CAD-C finely captures contacts at intermediate genomic sites and the interface between cohesion-mediated looping and block compartmentalization. Evaluation of Chromosome 6 95-125 Mb at 50 kb resolution demonstrates fine compartment and topologically associated domain (TAD) structures at intermediate chromosomal scales. CAD-C finely captures contacts at intermediate genomic sites and the interface between cohesin-mediated looping and block compartmentalization (data not shown).

[000220] CAD-C captures loop domains, topologically associated domains (TADs), dots, flames and provides complex transcription and cohesion loop extrusion-mediated features, including at various resolutions. Chromosome 2 (Chr2) 233-238 Mb at 10kb resolution demonstrates this (data not shown).

[000221] CAD-C captures regulatory enhancer-gene contacts, including with high signal-to-noise down to 500bp resolution (~2.5 nucleosome/bin). CAD-C compared with Micro-C at 500bp resolution of chr8

evaluation of human hTERT-RPE1 cells demonstrates regulatory enhancer gene contacts, while Micro-C fails to capture them (data not shown).

[000222] CAD as a tool for *in vitro* chromatin digestion

[000223] The use of recombinant CAD for digestion of chromatin, including high-throughput DNA sequencing, was reported previously in a small number of publications, with varying degrees of documentation and benchmarking (Xiao, Widlak, and Garrard 2007; Allan et al. 2012; Luse et al. 2020; Spector et al. 2022; Naughton et al. 2022; Piotr Widlak and Garrard 2006b). (Xiao, Widlak, and Garrard 2007) report use of a murine CAD and insert TEVP cleavage sites immediately downstream of the two caspase sites within ICAD. To generate TEV protease activated ICAD, the caspase sites were retained and a TEV protease site added adjacent to each of the caspase sites by Xiao. Xiao reported that maintenance of the second caspase-3 cleavage site was important for maintaining ICAD chaperone function. They reported that with the caspase-3 sites being important for ICAD (DFF45) chaperone activity, they created ICAD (DFF45) mutants in which the TEVP cleavage sequence was inserted immediately downstream of the caspase-3 cleavage sites. In Xiao's mutants, the DEPD and DAVD caspase recognition sites were maintained. This provided mutant ICAD that retained chaperone activity, retained caspase cleavage sites, and also had TEV protease cleavage sites. The Xiao mutant ICAD, however, was not cleaved by caspase-3, likely because of the adjacent TEV site sequences. Co-expression of TEVP with the modified ICAD (denoted DFF-T) under galactose control resulted in nucleosomal DNA laddering and cell death in *S. cerevisiae*.

[000224] Luse et al. 2020 and Spector et al 2022 use human CAD (DFF40) and ICAD (DFF45) inserted into a vector bi-cistronically for coexpression with a 6x His-tag at the C-terminus of DFF40 and TEV sites (ENLYFQS) inserted immediately downstream of the two Caspase-3 (117-118, 224-225) digestion sites within DFF40 as previously described above by Xiao. Luse et al and Spector et al bulk pre-activate the enzyme with TEVp. Luse et al assesses the sequence and functional organization of the human RNA polymerase promoter in confluent cultured HeLa cells, providing a footprint of RNA Polymerase II. Spector et al further assessed data generated from and promoter architecture in cultured HeLa cells and the human cytomegalovirus (HCMV) genome promoter.

[000225] CAD expression vector design An optimized IPTG-inducible promoter design adapted from (Shilling et al. 2020) was inserted into the common pRSF1-Duet vector. Murine CAD was utilized instead of human CAD, including in as much as it has been previously reported that murine CAD can be expressed at a much higher yield than its human homolog, for example in E. coli (Meiss et al. 2001). TEV site insertion into ICAD follows the scheme described in (Ageichik et al. 2007) The TEV site insertion schema we utilize in murine CAD is unique.

[000226] The exemplary modified pRSFDuet-1 vector for co-expressing CAD and ICAD^{2XTEV} is depicted in FIGURE 5. A VR221 (T7::ICAD2xTEV-IRES-TwinStrepTag-muGFP-SUMOEu1-CAD) and also a VR227 (T7::ICAD2xTEV_T7::TwinStrepTag-muGFP-SUMOEu1-CAD) were generated. These express CAD that is uniquely activatable with a specific enzyme, the CAD has an affinity tag, it can be purified via the one or more tag, and the tag can be cleanly removed to generate active enzyme. The VR227 construct promotes expression of TEVp-activatable CAD, affinity-tagged CAD, which is isolatable or purifiable using GFP nanobody or TwinStrepTag pulldown, and the tag can be removed (scarless removal) using SUMO.

[000227] CAD purification Buffer compositions and purification strategy is in line with that earlier reported (Meiss et al. 2001; Korn et al. 2002), though buffers were modified to be compatible with the commercial TwinStrepTag system.

[000228] CAD digestion CAD digestion buffer and reaction conditions are in line with (P. Widlak and Garrard 2001; Piotr Widlak and Garrard 2006a).

[000229] CAD-C protocol CAD-C protocol was based upon and modified from reported RCMC (Goel, Huseyin, and Hansen 2023) and MCC (Hamley et al. 2023) protocols. The CAD-C protocol employs proximity ligation conditions based on RCMC (Goel, Huseyin, and Hansen 2023) whilst the capture of ligation junctions by sonication is based on MCC (Hamley et al. 2023).

[000230] CAD-C Materials and Methods

[000231] Design of CAD-ICAD expression construct

[000232] Choosing source sequences for wild-type CAD and ICAD. Amino acids sequences corresponding to full open reading frames for *Mus musculus* CAD (Dffb/DFF-40) and *Mus musculus* ICAD (Dffa/DFF-45) and were retrieved from the uniprot database and are shown below (CAD=DFFB, ICAD=DFFA). Full coding sequences were used.

[000233] >sp|O54788|DFFB_MOUSE DNA fragmentation factor subunit beta OS=Mus musculus OX=10090 GN=Dffb PE=1 SV=1 CAD sequence (SEQ ID NO:5)

MCAVLRQPKCVKLRALHSACKFGVAARSCQELLRKGCVRFQLPMPGSRLCLYEDGTEVTD
DCFPGLPNDAELLLLTAGETWHGYVSDITRFLSVFNEPHAGVIQAARQLLSDEQAPLRQK
LLADLLHHVSQNITAETREQDPSWFEGLESRFRNKSGYLRYSCESRIRGYLREVSAYTSM
VDEAAQEEYLRVLGSMCQKLKSVQYNGSYFDRGAEASSRLCTPEGWFSCQGPFDLESCLS
KHSINPYGNRESRILFSTWNLDHIIEKKRTVVPTLAEAIQDGREVNWEYFYSLLFTAENL
KLVHIACHKKTTHKLECDRSRIYRPQTGSRRKQPARKKRPARKR

[000234] >sp|O54786|DFFA_MOUSE DNA fragmentation factor subunit alpha OS=Mus musculus OX=10090 GN=Dffa PE=1 SV=2 ICAD sequence (SEQ ID NO:3)

MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPIT
LVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWIYNDSDGGTAWVSQESFEADEPDSRA

GVKWKNVARQLKEDLSSHLLSEEDLQALIDIPCAELAQELCQSCATVQGLQSTLQQVLD
QREEARQSKQLLELYLQALEKEGNILSNQKESKAALSEELDAVDTGVGREMASEVLLRSQ
ILTTLKEKPAPELSLSSQDLESVSKEDPKALAVALSWDIRKAETVQQACTTELALRLQQV
QSLHSLRNLSARRSPLPGEPQRPKRAKRDSS

[000235] Codon optimization of CAD and ICAD cDNA sequences for E. coli expression

[000236] The amino acid sequences were reverse translated using *E. coli* optimal codons using the free GenSmart™ Codon Optimization tool online. The sequences were then further codon optimized by manual inspection of *E. coli* codon frequency usage in Snapgene. The resulting DNA sequences were purchased from IDT and synthesized as gene blocks.

[000237] >MmDFF40_EcOpt/wt_CAD Codon Optimized CAD encoding nucleic acid sequence (SEQ ID NO:12)

ATGTGTGCTGTACTGCGCCAACCCAAATGCGTCAAACTCCGCGCTCTGCACTCTGCGTGTAAATTTG
GTGTGGCGGCGCGTTCGTGCCAAGAATTGTTGCGGAAGGGGTGCGTGCGTTTCCAGCTGCCGAT
GCCGGGTTCTCGTCTGTGCCTGTACGAGGACGGCACCGAAGTCACCGACGACTGTTTTCCGGGCC
TGCCGAATGATGCAGAGCTGTTGTTGTTGACGGCGGGTGAAACCTGGCATGGTTACGTTAGCGACA
TCACGCGTTTCCTGAGCGTTTTTAACGAACCGCATGCGGGCGTTATTCAAGCTGCACGTCAATTGCT
GTCCGACGAGCAGGCCCCACTGCGTCAGAAGCTGCTTGCTGATTTGCTGCACCACGTTAGCCAAA
ACATTACCGCGGAGACGCGTGAGCAGGACCCGAGCTGGTTCGAAGGCCTGGAGAGCCGCTTCCG
CAACAAAAGCGGTTACCTGCGCTACAGCTGCGAATCCCGTATTCGTGGTTACCTCCGCGAGGTGTC
TGCGTATACCAGCATGGTGGATGAGGCCGCTCAGGAGGAATACCTGCGCGTGCTGGGTTCGATGTG
CCAAAAGCTTAAGTCTGTGCAGTATAACGGCAGCTATTTCGACCGCGGCGCAGAGGCGTCCAGCCG
TCTGTGTACCCCGGAAGGCTGGTTTAGCTGCCAGGGTCCGTTTGATCTCGAGAGCTGCCTGAGCA
AACACTCCATCAACCCGTATGGTAATCGTGAATCCCGCATCCTGTTCTCCACCTGGAATCTGGATCAC
ATCATCGAAAAAAAACGTACTGTAGTTCCGACTCTGGCGGAGGCAATTCAAGACGGCCGTGAGGTTA
ACTGGGAATACTTCTATAGTCTGCTGTTTACCGCGGAGAACCTGAAATTGGTTCATATCGCCTGCCAC
AAAAAGACCACCCATAAACTGGAATGTGATCGTTCTCGCATTTATCGTCCGCAGACCGGTTCCCGCC
GTAAGCAGCCGGCGCGTAAGAAGCGCCCTGCTCGCAAGCGC

(ATG start codon underlined, no stop codon synthesized)

[000238] >MmDFF45_EcOpt/wt_ICAD Codon Optimized ICAD encoding nucleic acid sequence (SEQ ID NO:13)

ATGGAACTGTCACGCGGAGCTAGTGCACCCGACCCGGACGACGTGCGTCCGCTGAAACCGTGCCT
GCTCCGTCGCAACCACAGTCGTGATCAACATGGTGTGGCCGCGAGCAGCCTGGAAGAGTTGCGCA
GCAAGGCCTGCGAACTGCTGGCGATTGATAAAAGCTTGACCCCGATTACCTTGGTTCTGGCTGAGG
ACGGCACGATCGTGGATGACGACGACTACTTCCTGTGTCTGCCTAGCAACACCAAATTCGTGGCGT
TGGCCTGTAATGAAAAGTGGATTTATAACGATTCGGATGGCGGTACGGCTTGGGTTTCGCAAGAAAG
CTTTGAAGCGGATGAACCGGACAGCCGTGCCGGTGTCAAATGGAAAAACGTCGCGCGTCAGCTGA
AGGAGGACCTGTCAAGCATTATCCTGTTGTCCGAGGAGGACTTGCAAGCTCTGATCGACATCCCGT

GCGCGGAGCTTGCGCAGGAGCTGTGCCAGAGCTGTGCGACCGTTCAGGGCCTGCAATCCACCTTA
CAACAAGTGCTGGATCAGCGCGAAGAAGCACGTCAGAGCAAGCAGCTGCTGGAGTTGTACCTGCA
AGCGCTGGAGAAAGAGGGCAATATCCTGTCTAACCAGAAAGAGTCTAAGGCGGCATTAAGCGAAGA
GTTAGATGCAGTTGATACCGGTGTAGGTCGTGAAATGGCATCCGAAGTTCTGCTTCGCTCTCAGATT
CTGACCACCCTGAAGGAGAAGCCGGCACCGGAACTCTCCCTGTCCTCTCAAGATTTGGAGAGCGT
TAGCAAGGAGGATCCGAAAGCGTTGGCGGTGGCCTTGAGCTGGGATATCCGTAAAGCTGAGACTGT
GCAGCAGGCGTGCACCACGGAACTTGCTTTACGTCTGCAACAAGTTCAGTCCCTGCACAGCCTGC
GCAATCTGTCTGCACGTCGTAGCCCGCTGCCGGGTGAACCGCAGCGCCCAAAGCGCGCGAAACG
CGACAGCAGT

(ATG start codon underlined, no stop codon synthesized)

[000239] Modifying ICAD sequence to insert TEV protease sites

[000240] Unless otherwise stated, amino acid coordinates include the methionine start codon as 'aa 1' (1-indexed). There are two uniprot-annotated caspase cleavage sites in ICAD. Recognition sites are: site 1 = aa 112-119 inclusive, site 2 = aa 219-226 inclusive. The caspase cleavage positions are: site 1 between D117|S118, site 2 D224|T225. A section of the caspase recognition sequences were first deleted before insertion of TEV protease recognition sequences (Ageichik et al. 2007). For site 1 'DEP' residues at position 114-116 were deleted, for site 2 'DAV' residues at positions 221-223 were deleted. Subsequently the TEV protease recognition sequence (ENLYFQS) was inserted into the positions of the deleted amino acids for caspase site 1 and caspase site 2. Overall 3x2=6 amino acids were deleted from the wild-type ICAD amino acid sequence and 7x2=14 amino acids were added to give a modified ICAD ^{2xTEV} protein of length 339 aa.

[000241] > ICAD^{2xTEV} nucleic acid sequence (SEQ ID NO:7)

ATGGAACTGTCACGCGGAGCTAGTGCACCCGACCCGGACGACGTGCGTCCGCTGAAACCGTGCCT
GCTCCGTCGCAACCACAGTCGTGATCAACATGGTGTGGCCGCGAGCAGCCTGGAAGAGTTGCGCA
GCAAGGCCTGCGAACTGCTGGCGATTGATAAAAGCTTGACCCCGATTACCTTGGTTCTGGCTGAGG
ACGGCACGATCGTGGATGACGACGACTACTTCCTGTGTCTGCCTAGCAACACCAAATTCGTGGCGT
TGGCCTGTAATGAAAAGTGGATTTATAACGATTCGGATGGCGGTACGGCTTGGGTTTCGCAAGAAAG
CTTTGAAGCGGAAAACCTGTATTTTCAGAGCGACAGCCGTGCCGGTGTCAAATGGAAAAACGTCGC
GCGTCAGCTGAAGGAGGACCTGTCAAGCATTATCCTGTTGTCCGAGGAGGACTTGCAAGCTCTGAT
CGACATCCCGTGCGCGGAGCTTGCGCAGGAGCTGTGCCAGAGCTGTGCGACCGTTCAGGGCCTG
CAATCCACCTTACAACAAGTGCTGGATCAGCGCGAAGAAGCACGTCAGAGCAAGCAGCTGCTGGAG
TTGTACCTGCAAGCGCTGGAGAAAGAGGGCAATATCCTGTCTAACCAGAAAGAGTCTAAGGCGGCAT
TAAGCGAAGAGTTAGAAAACCTGTATTTTCAGAGCGATACCGGTGTAGGTCGTGAAATGGCATCCGA
AGTTCTGCTTCGCTCTCAGATTCTGACCACCCTGAAGGAGAAGCCGGCACCGGAACTCTCCCTGTC
CTCTCAAGATTTGGAGAGCGTTAGCAAGGAGGATCCGAAAGCGTTGGCGGTGGCCTTGAGCTGGG

ATATCCGTAAAGCTGAGACTGTGCAGCAGGCGTGCACCACGGAACTTGCTTTACGTCTGCAACAAGT
TCAGTCCCTGCACAGCCTGCGCAATCTGTCTGCACGTCGTAGCCCGCTGCCGGGTGAACCGCAGC
GCCCAAAGCGCGCGAAACGCGACAGCAGT

(*E. coli* codon-optimized TEV protease recognition sequences underlined)

[000242] _> ICAD^{2xTEV} amino acid sequence (SEQ ID NO:1)

MELSRGASAPDPDDVRPLKPCLLRRNHSRDQHGVAASSLEELRSKACELLAIDKSLTPIT
LVLAEDGTIVDDDDYFLCLPSNTKFVALACNEKWIYNDSDGGTAWVSQESFEAENLYFQSDSRA
GVKWKNVARQLKEDLSSIILLSEEDLQALIDIPCAELAQELCQSCATVQGLQSTLQQVLD
QREEARQSKQLLELYLQALEKEGNILSNQKESKAALSEELENLYFQSDTGVGREMASEVLLRSQ
ILTTLKEKPAPELSLSSQDLESVSKEDPKALAVALSWDIRKAETVQQACTTELALRLQQV
QSLHSLRNLSARRSPLPGEPQRPKRAKRDSS

(The altered additional amino acids in the engineered ICAD are underlined)

[000243] Expression plasmid backbone design

[000244] A custom vector backbone design was assembled based on modification of the pRSFDuet-1 vector which allows for IPTG-inducible expression of two different open reading frames from within the same backbone. The IPTG-inducible T7 promoter cassettes were modified to incorporate changes that enhance protein overexpression by optimizing transcription and translation, based on publication by Shilling and colleagues (Shilling et al. 2020) . Specifically, the T7 promoter sequence was extended and an ORF translation 'TIR-2' leader sequence was placed immediately downstream of the Shine-Dalgarno sequence. A vector map of the expression plasmid is provided in FIGURE 5. The expression plasmid (VR226 pRSFDuet_TIR2-ICAD(2xTEV)_TIR2-TwinStrepTag-muGFP-SUMOEu1-CAD) sequence is provided (SEQ ID NO:14).

[000245] CAD is known to form a tight heterodimer with ICAD to form the CAD proenzyme. ICAD acts as both an inhibitor of CAD nuclease activity, as well as a chaperone which promotes the native folding state of CAD in its proenzyme state. CAD expressed alone in *E. coli* does not fold into a native state and does not form an active nuclease. Therefore, an optimal ratio of expression of ICAD to CAD is required to occur within the same *E. coli* cell. Expression of too little ICAD will result in poor folding of the excess pool of CAD, and therefore lead to poor solubility overall yield of CAD proenzyme. On the other hand, expressing too large of an excess of ICAD relative to CAD will lead to a poor yield of CAD proenzyme. We found the placement of the ICAD ORF upstream of the CAD ORF in the modified pRSFDuet provides an optimal yield ratio of ICAD:CAD expression.

[000246] Addition of affinity purification tags to CAD

[000247] We placed protein affinity purification tags on the N-terminus of the CAD ORF, reasoning that we should recover the CAD-ICAD proenzyme heterodimer by pulling down on CAD alone, due to their tight binding affinity. We used a dual tag approach, adding a Twin-Strep-tag followed by muGFP, a highly-

stable variant of GFP (Scott et al. 2018) , followed by $SUMO^{Eu1}$, a variant of SUMO resistant to cleavage by endogenous eukaryotic SUMO proteases (Vera Rodriguez, Frey, and Görlich 2019) . The Twin-Strep-tag, muGFP and $SUMO^{Eu1}$ are separated by disordered protein linkers including GGSGGSGGS and GDGAGLIN sequences, the CAD is placed directly after $SUMO^{Eu1}$ without extraneous linker aa residues to enable optional scarless tag removal.

[000248] >Leader-TwinStrepTag-muGFP-SUMOEu1-CAD_ORF (SEQ ID NO: 15)

MQLSAWSHPQFEKGGGSGGGSGGSAWSHPQFEKGDGAGLINSKGEELFTGVVPILVELDG
DVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLVTTLTYGVLCFSRYPDHMKRHD
FFKSAMPEGYVQERTISFKDDGTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNF
NSHNVYITADKQKNGIKAYFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLS
KDPNEKRDHMVLLEDVTAAGITHGMDELYKGGSGGSGGSGDGAGLINSAAGGEEDKKPAGG
EGGGAHINLKVKGQDGNEVFFRIKRSTQLKKLMNAYCDRQSVDMKAIAFLFKGRRLRAERTP
DELEMEDGDEIDAMLHQTGGCAVLRQPKCVKLRALHSACKFGVAARSCQELLRKGCVRFQL
PMPGSRLCLYEDGTEVTDDCFPGLPNDAELLLLTAGETWHGYVSDITRFLSVFNEPHAGVIQA
ARQLLSDEQAPLRQKLLADLLHHVSQNITAETREQDPSWFEGLESRFRNKSGYLRYSCESRI
RGYLREVSAYTSMVDEAAQEEYLRVLGSMCQKLKSVQYNGSYFDRGAEASSRLCTPEGWF
SCQGPFDLESCLSKHSINPYGNRESRILFSTWNLDHIIEKKRTVVPTLAEAIQDGREVNWEYF
YSLLFTAENLKLVHIACHKKTTHKLECDRSRIYRPQTGSRRKQPARKKRPARKR

(disordered linker sequences underlined)

[000249] Plasmid propagation

[000250] The modified pRSFDuet-1 plasmid containing the $ICAD^{2xTEV}$ TwinStrepTag-muGFP-$SUMO^{Eu1}$ -CAD ORFs was propagated in NEB Stable *E. coli* strain at 37 °C, as the construct proved toxic in a routine NEB 5-alpha cloning strain, likely due to the high background protein cargo expression. NEB Stable strain expresses $LacI^{q}$ to shut off leaky transcription of the plasmid cargo.

[000251] CAD-ICAD expression and purification

[000252] Culture induction, scale-up and induction pRSFDuet-1_$ICAD^{2xTEV}$ TwinStrepTag-muGFP-$SUMO^{Eu1}$ -CAD plasmid was transformed into NEB C3013 strain (T7 Express lysY/I^{q}), grown on agar plates with 50 µg/mL kanamycin selection. Starter cultures were grown from fresh glycerol stock streak outs, single colony was picked and grown in 20 mL complete Terrific Broth (TB) supplemented with 0.4% glycerol, 50 µg/mL kanamycin, 0.005% antifoam 204 agent, grown at 37 °C, ambient atmosphere, with 300-350 rpm shaking in a baffled 250 mL culture flask. Starter culture was grown for 18-24 h, at which point it was used to inoculate full-scale induction cultures 1:100-fold, and grown out at 280-320 rpm, at 37 °C, as for starter culture. Typical prep scale is 500 mL culture in a 2 L double-baffled shaker flask, sufficient for a preparative CAD purification. Cultures were grown at 37 °C until OD_{600} = 1.00-1.40, approximately 3 hours under optimal aeration in double-baffled flasks. After cultures reached the desired

OD_{600} , cultures were placed briefly on ice for 5-10 minutes and induced with 1 mM IPTG, final. Cultures were placed into a pre-cooled shaker set to 18 °C and grown with shaking at 280-320 rpm for 18-24 hours.

[000253] Cell harvesting *E. coli* was harvested by pelleting at 5'000-10'000 xg at 4 °C. Cell pellets were washed in 1x ice-cold DPBS, and snap frozen with liquid nitrogen and stored at – 80 °C.

[000254] Cell lysis *E. coli* pellet from 500 mL was resuspended on ice in 4 mL 1x lysis buffer (CLysB) per gram of wet pellet. Cells were lysed by sonication on ice (6 min total ON time, 10 sec ON/OFF cycle, 80% powder). Sonicated lysate was clarified by centrifugation at 21'000 xg at 4 °C for 30 minutes.

[000255] Affinity purification Clarified lysate from 4 g wet pellet-equivalent was 0.22 µm filtered and loaded onto one 1 mL Strep-TactinXT 4Flow column, and washed using 1x strep wash buffer (CSWB). Bound protein was eluted with strep elution buffer supplemented with 20 mM βME (IBA-BXT) into five fractions of 600 µL, 800 µL, 800 µL, 800 µL, 1400 µL. Fraction 3 yields most muGFP fluorescence. Fractions 2 to 5, inclusive, were pooled and dialyzed against 2 L of 1x SEC (CSEC) in a pre-wetted 10 MWCO 15 mL-capacity Slide-A-Lyzer G2 dialysis cassette overnight at 4 °C with stirring.

[000256] Size exclusion chromatography and storage Dialyzed eluate was 0.22 µm filtered and injected onto a S200 16/600 on an Akta Pure instrument and separated with 1 mL/min flow of 1x CSEC into 0.8 mL fractions in 96 deep-well collection plates. Majority of intact-length TwinStrepTag-muGFP-$SUMO^{Eu1}$ -CAD-$ICAD^{2xTEV}$ complex eluted in peak 1 (peak at 49.48 mL) corresponding to the column's void volume. 19x0.8 mL Peak 1 fractions were pooled and concentrated using a pre-equilibrated 15 mL-capacity Amicon Ultra-15 30 MWCO centrifuge tube filter-concentrator down to 1.1 mL volume 1:14 of original volume. 100% UltraPure glycerol was added to a final concentration of 25%. Protein was aliquoted into 1.5 mL Protein LoBind tubes and placed in an enzyme block in the – 20 °C (not snap-frozen). Once thawed, aliquots were kept at 4°C, avoiding freeze-thaw cycles.

[000257] Buffer Table

1x Lysis Buffer (CLysB)	20 mM HEPES-NaOH pH 7.8, 100 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.01% Triton X-100, 10 mM βME, 1x c0mplete EDTA-free, 200 U/mL Dr. Nuc
1x Strep Wash Buffer (CSWB)	20 mM HEPES-NaOH pH 7.8, 150 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.01% Triton X-100, 10 mM βME
1x Strep Elution Buffer (IBA-BXT)	100 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 50 mM biotin, 20 mM βME
1x SEC Buffer (CSEC)	20 mM HEPES-HCl pH 7.5, 100 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.01% CHAPS, 10 mM DTT

[000258] CAD-ICAD *in vitro* activity assays

[000259] Naked DNA

[000260] DNase activity was determined using purified genomic Lambda phage DNA purchased from NEB. Digestions were carried out in 1x Widlak Digestion Buffer 70 (WDB-70). Typical reactions used DNA at a final concentration of 1 µg / 10 µL of reaction volume, or 2.5 µg λ DNA per 25 µL volume. Typical reactions used 2 µL or approx 2 µg muGFP-tagged CAD, and were activated with 2 µL (20 U) of TEV protease (NEB). TEVp-specific CAD Reactions were incubated for 5 minutes at 30 °C (optimal for TEV protease activity), followed by 15 minutes at 37 °C (optimal murine CAD temperature), followed by heat inactivation at 65 °C for 1 minute. For longer reaction times this 30 °C to 37 °C incubation time ratio was kept constant. To release DNA from any bound CAD protein, reactions were quenched with final 0.77% SDS and diluted 1:10 with ultrapure water to a final 20 µL volume for separation on the E-Gel system using 1% EX E-Gel agarose gels. Under these reaction conditions, 1 µg of λ DNA is completely digested to < 100 bp using 40 ng of muGFP-tagged CAD, at a CAD concentration of 20 nM (molecular weight of TwinStrepTag-muGFP-SUMO-CAD = 81.9 kDa).

[000261] CAD unit definition

[000262] 1 unit of CAD is defined as the amount of enzyme which can digest 1 µg of λ DNA to mostly < 100 bp species. One CAD unit typically corresponds to 40 ng of muGFP-tagged CAD, calculated at a typical CAD concentration of 20 nM (molecular weight of TwinStrepTag-muGFP-SUMO-CAD = 81.9 kDa).

[000263] Human cells

[000264] Chromatin digestion assays were performed on native, or chemically-crosslinked cell nuclei as for naked DNA, with a modified 1x Widlak Digestion Buffer 10 (WDB-10). Reactions were quenched with either final 1% SDS, 50 °C heat denaturation or final 1 mM $ZnCl_2$. Protein was digested by proteinase K digestion, crosslink reversal was carried out overnight at 56 °C, and DNA was purified using Zymo genomic DNA Clean and Concentrator kit. 1 µg muGFP-CAD was sufficient to digest 200'000 hTERT-RPE1 nuclei previously fixed with 3 mM DSG + 1% FA to completion in a reaction volume of 250 µL in 1 hour (15 min 30 °C, 45 min 37 °C). Complete digestion was defined as where no apparent additional digestion is observed after increasing CAD concentration 10-fold under otherwise identical conditions.

[000265] Buffer Table Both WDB-70 and WDB-10 buffers are prepared as 10x stocks.

1x WDB-70	70 mM KCl, 3 mM $MgCl_2$, 1 mM DTT, 10 mM HEPES-KOH pH 7.5
1x WDB-10	10 mM KCl, 3 mM $MgCl_2$, 1 mM DTT, 10 mM HEPES-KOH pH 7.5

[000266] CAD-C protocol

[000267] Cell Harvesting

[000268] Contact-inhibited, quiescent hTERT-RPE1 cells were harvested using TrypLE reagent, washed with DPBS and counted. GM12878 cells were harvested by centrifugation and counted. Cells were first washed in pre-warmed Dulbecco's Phosphate-Buffered Saline (DPBS) at 37 °C and then resuspended in 37 °C DPBS at a concentration of 2×10^6 cells/mL. The final cell concentration at the time of fixation was adjusted to 1×10^6 cells/mL. For cell harvesting, no more than 8 mL of cells (at 2×10^6 cells/mL) or 16 mL of cells (at 1×10^6 cells/mL) were used per 50 mL conical tube to allow sufficient space for adding Tris-HCl for quenching the fixative. Cells were maintained at 37 °C in a pebble bath until the Di(N-succinimidyl) glutarate (DSG) fixative was prepared.

[000269] DSG Fixation

[000270] DSG fixation was performed on the lab bench. A 300 mM DSG stock solution was freshly prepared by dissolving DSG powder in 100% dimethyl sulfoxide (DMSO). For example, 30 mg of DSG was dissolved in 300 µL of DMSO to create the 100× concentrate. The 300 mM DSG stock solution was diluted 1:50 in pre-warmed DPBS to generate a 6 mM working solution. This solution was added to the cells at a 1:1 ratio to achieve a final concentration of 3 mM DSG and 1×10^6 cells/mL. To minimize cell loss, gentle mixing by inversion on a rocker was used rather than pipetting. The cell suspension was rocked gently at room temperature (RT) for 35 min.

[000271] Formaldehyde (FA) Fixation

[000272] After DSG fixation, cells were subjected to crosslinking with formaldehyde (FA) inside a fume hood. A fresh ampule of 16% methanol-free FA was used to prepare a 2% FA working solution in pre-warmed DPBS. The 2% FA solution was then added to the DSG-treated cells at a 1:1 ratio, resulting in a final concentration of 1×10^6 cells/mL, 3 mM DSG, and 2% FA. The suspension was rocked gently at RT for exactly 10 min.

[000273] Fixative Quenching with Tris-HCl

[000274] Fixation was quenched using Tris-HCl, as described previously (Bylino et al, 2021; deJonge et al, 2020). A 2.25 M Tris-HCl stock solution was prepared, and 0.5 volumes were added to the cell suspension to reach a final concentration of 750 mM Tris-HCl. For example, 16 mL of 2.25 M Tris-HCl was added to 32 mL of cell suspension containing DSG and FA. The suspension was gently rocked at RT for 5 min. Cells were then pelleted by centrifugation at 1,000 × g for 5 min at 4 °C, and the supernatant was discarded. Cells were washed twice with ice-cold DPBS (30 mL per 50 mL tube), pelleted, and finally resuspended to a concentration of 4×10^6 cells/mL.

[000275] Cell Storage

[000276] Aliquots of 500 µL of the final cell suspension were transferred into pre-labeled 1.5 mL tubes (each containing 2 × 10^6 cells or more). After removing excess DPBS, cells were snap-frozen in liquid nitrogen and stored at −80 °C.

[000277] Nuclear isolation and CAD digestion for CAD-C

[000278] Cells were grown, harvested and fixed in large batches to support multiple experiments, typically 5E6 cells were used as input per technical replicate. Frozen fixed cell pellets were placed on ice and resuspended with 1 mL ice-cold 1x WDB-10 supplemented with final 0.2% NP-40 surfactant, pipetting intermittently during 20 minute incubation on ice. Cells were spun for 5 minutes at 1750 xg at 4 °C, supernatant removed and resuspended in 500 µL 1x WDB-10+0.2% NP-40, washed by pipetting and again pelleted by spinning for 5 minutes at 1750 xg at 4 °C. Supernatant was removed and resuspended in 200 µL 1x WDB-10, with RNAse. For hTERT-RPE1 samples 7.5 µL 10 µg/µL RNase A (Thermo Scientific EN0531) was added to a final concentration of 0.36 µg/µL RNAse, mixed by pipetting and incubated in a thermomixer at 1000 rpm at 30 °C for 15 minutes. For GM12878 and K562 samples RNase T1 (Thermo Scientific EN0542) was added to a final concentration of 10 U/µL RNase T1, mixed by pipetting and incubated in a thermomixer at 500 rpm at 37 °C for 15 minutes. Following RNAse treatment, 20 µL (40 µg) TST-muGFP-SUMO^{Eu1}-CAD:ICAD^{2xTEV} proenzyme and 5 µL (50 U) of TEV protease (NEB P8112S) was added to each reaction. CAD digestion proceeded for 20 minutes at 30 °C and 1 hour at 37°C, shaking at 1000 rpm in a thermomixer with heated lid on. CAD was quenched with 1 mM final ZnCl2 and incubated for 5 minutes at room temperature. Nuclei were pelleted for 5 minutes at 2'000 xg for 5 minutes at 4°C, supernatant removed and washed with 1 mL 1xWDB-10+0.2% NP-40. At this point 50µL samples out of 250µL (20%) were taken to serve as digestion-only controls. To 50µL control aliquots added 2.50µL 20% SDS and 2.50µL proteinase K and incubated overnight at 65°C at 750 rpm in a thermomixer with heated lid on. The remaining 200µL of resuspended nuclei were stored at 4°C until further processing.

[000279] Nuclear isolation and CAD digestion for CAD-seq

[000280] For GM12878 CAD titration, used 4E6 nuclei per technical replicate, specifically 200 µL nuclei resuspended in WDB-10 buffer at concentration of 1E6 nuclei/50µL = 2E5 nuclei/µL.

[000281] CAD-seq DNA library preparation

[000282] After CAD digestion, samples were deproteinated and crosslinks were reversed by addition of 6.50 µL 20% SDS, 5.20 µL 5 M NaCl, 13.0 µL Proteinase K (NEB P8107S, 20 µg/µL) incubated in a Thermomixer (Eppendorf) with a heated lid at 65 °C, 500 rpm overnight. DNA was subsequently purified using a Genomic DNA Clean & Concentrator-10 (Zymo D4011) using a 1:5 volume ratio of sample to ChIP DNA Binding Buffer, following manufacturer's instructions with minor modifications: after washes, column was dry-spinned for 2 minutes at 18'000 xg and samples were eluted in 70 µL of Zymo DNA

elution buffer previously heated to 65 °C. Purified CAD digest DNA was analyzed by 1% Agarose EX E-Gel and Cell-free DNA ScreenTape on an Agilent Tapestation instrument to quantify fraction of sample DNA below 700 bp. Purified DNA concentration was quantified using Qubit dsDNA Broad Range kit (Thermo Fisher Scientific Q32853). Sample input volumes for library preparation were adjusted to give an equal input of 500 ng DNA within the 50-700 bp size range, to allow uniform library preparation conditions across CAD titration samples.

[000283] DNA from CAD-digested cell nuclei was prepared for sequencing (CAD-seq) using NEB Ultra II (NEB E7645L) Illumina DNA library preparation with modifications to ensure capture of short sub-nucleosomal DNA fragments, using manufacturer incubation times unless otherwise stated. To 50 µL concentration-adjusted input sample 7 µL end prep buffer, 3 µL end prep enzyme mix were added and pipetted with P200 >10 times. 2.5 µL undiluted NEBNext 15 µM hairpin adapter was added, ligation was carried out as per manufacturer instructions, followed by treatment with 3 µL USER enzyme. Free and self-ligated adapters were cleaned up with an increased 150 µL volume of Ampure XP beads per sample to ensure recovery of sub-nucleosomal DNA fragments. Beads were washed twice with 80% EtOH and air dried for 5 minutes. Samples were eluted with 17 µL 10 mM Tris-HCl pH 8 and incubated for 20 minutes at RT.

[000284] Indexing PCR was carried out as per manufacturer's instruction (NEB E6441A) using 4 PCR cycles using the manufacturer's thermocycling conditions. Amplified library PCR reactions were cleaned using 1.2x volume of Ampure XP beads and eluted in 52 µL IDTE 7.5 buffer. Average size and DNA concentration of libraries was quantified by Agilent HS D1000 Tapestation system and by Qubit dsDNA HS kit before pooling for sequencing.

[000285] Biotin labeling

[000286] DNA ends were recessed and repaired using T4 PNK and Large Klenow Fragment. 55µL H 2 O, 10µL 10x NEB 2.1 buffer, 20µL 10 mM ATP, 5µL 100 mM DTT were added to 200µL nuclei in 1xWDB-10+0.2% NP-40, mixed, then 5µL (50 U) T4 PNK (NEB M0201L) and 10µL (50 U) Large Klenow Fragment (NEB M0210L). Reaction was incubated for 30 minutes at 37°C, 1000 rpm in a thermomixer with heated lid on. Biotin fill-in end labeling was performed by adding 10µL 1 mM Bio-dATP (Jena #NU-835-BIO14-S), 10µL 1mM Bio-dCTP (Jena #NU-809-BIOX-S), 1µL 10 mM dTTP, 1µL 10 mM dGTP, 5µL 10x T4 DNA ligase buffer, 0.25µL 20 mg/mL BSA, 22.75µL H_2O. Total volume added to sample = 50µL. Reaction was carried out for 1 hour at 25°C in a thermomixer with interval mixing (1 minute 1000 rpm, 3 minutes 0 rpm, cycled) followed by 4°C hold.

[000287] Proximity ligation

[000288] 442.5µL H_2O, 50µL 10x T4 DNA ligase buffer, 2.5µL 20 mg/mL BSA and 25µL (10'000 U) T4 DNA ligase were added. Ligation reaction was carried out at 25°C in a thermomixer, 300 rpm, overnight.

[000289] Biotin removal from unligated DNA ends

[000290] 170 H_2O, 20µL 10x NEB 1 buffer, 10µL (1'000 U) Exonuclease III was added to and incubated for 15 minutes at 37°C in thermomixer using interval mixing, followed by 4°C hold.

[000291] DNA isolation

[000292] To digest protein and reverse crosslinks 26µL of 20 mg/mL proteinase K, 26µL 10% SDS, 10.4µL 5M NaCl, 2.6µL 10 mg/mL RNase A was added. Samples were incubated for 48 hours at 65°C in a thermomixer, 1000 rpm. 200 µL phenol-chloroform-isoamyl alcohol was added, samples were vortexed, transferred to Qiagen MaXtract tube, centrifuged for 15 minutes, 18'000 xg at 25°C. Supernatant was transferred to fresh tube. 20 µL 3 M sodium acetate, 500 µL 200-proof EtOH and 15 µg GlycoBlue was added, vortexed well and precipitated at – 80°C for 2 hours. Samples were then centrifuged for 15 minutes at 18'213 xg at 0°C, DNA pellets were washed with 1 mL ice-cold 200-proof EtOH. Samples were centrifuged for 5 minutes at 18'213 xg at 0°C and supernatant removed and discarded. Pellets were then air-dried for 25 minutes and DNA pellet was resuspended in 55 µL IDTE 7.5 buffer (IDT). The above procedure was followed for hTERT-RPE1 CAD-C samples. Alternatively for GM12878 and K562 CAD-C samples, DNA was purified using Genomic DNA Clean & Concentrator-10 (Zymo D4011) using a 1:5 volume ratio of sample to ChIP DNA Binding Buffer, following manufacturer's instructions with minor modifications: after washes, column was dry-spinned for 2 minutes at 18'000 xg and samples were eluted in 70 µL of Zymo DNA elution buffer previously heated to 65 °C.

[000293] DNA sonication

[000294] DNA concentration was measured using Qubit BR, and 1µg of DNA in 14µL volume was sonicated using Covaris in micro-tubes to an average size of ~220 bp.

[000295] CAD-C DNA library prep

[000296] DNA library preparation was carried out on streptavidin after enrichment for DNA molecules containing biotinylated ligation junctions. 10µL of MyOne C1 Dynabeads per technical replicate sample was washed in 1 mL of 1x BW buffer (1M NaCl, 5 mM Tris-HCl pH 7.5, 0.5 M EDTA pH 8.0). 120µL of previously sonicated DNA was diluted with 120µL of 2x BW buffer (2 M NaCl, 10 mM Tris-HCl pH 7.5, 1 M EDTA pH 8.0) and bound to beads at room temperature, rocking for 20 minutes. Beads containing bound DNA were washed twice times using 1x TBW buffer (1M NaCl, 5 mM Tris-HCl pH 7.5, 0.5 M EDTA pH 8.0, 0.1% Tween-20) with thermomixer shaking for 5 minutes at 1200 rpm each time.

[000297] Illumina DNA library preparation was carried out on-bead using NEB Ultra II (NEB E7645L) with minimal modifications, using manufacturer incubation times unless otherwise stated. Dynabeads

were resuspended in 50µL IDTE 7.5 buffer, 7µL end prep buffer, 3µL end prep enzyme mix and pipetted with P200 >10 times. 2.5µL undiluted NEBNext adapter (NEB E6441A) was added, ligation was carried out as per manufacturer instructions. Excess adapter was removed by washing three times in 150µL 1x TBW buffer, followed by two washes in 150µL IDTE 7.5 buffer, and finally resuspending the dynabeads in 15µL IDTE 7.5 buffer. Indexing PCR was carried out as per manufacturer's instruction (NEB E6441A) using 12 cycles. Amplified library PCR reactions were cleaned using 1.2x volume Ampure XP beads and eluted in 52µL IDTE 7.5 buffer. Library average size and DNA concentration was quantified by Agilent HS D1000 Tapestation system and by Qubit dsDNA HS kit.

[000298] DNA sequencing

[000299] CAD-seq libraries were Illumina sequenced using 2x60 paired-end reads on a NextSeq 1000 instrument or 2x150 paired-end reads on a NovaSeq X instrument. CAD-C DNA libraries were Illumina sequenced using 2x150 paired-end reads on a NovaSeq X instrument.

[000300] CAD-C alignment

[000301] Paired-end DNA sequencing reads were aligned to the human reference genome (hg38) using the bwa-mem2 aligner with 16 CPU cores. Raw FASTQ files were processed using a custom Bash script executed in a high-performance computing environment. Reads were aligned using bwa-mem2 with the -SP5M parameter, optimized for paired-end data, and the resulting output was piped to samtools for conversion to BAM format.

[000302] In silico fragment size selection

[000303] Fragment size selection for CAD-seq data was performed using custom bash scripts utilizing samtools and awk. For 100-200 bp fragments, the script retained fragments with lengths between 100 and 200 bp based on the $9 field, resulting in a size-filtered BAM file. Similarly, sub-100 bp fragments were isolated by excluding reads with lengths ≥100 bp. Filtered BAM files were generated for downstream analysis and further processing.

[000304] Genome coverage

[000305] Genome coverage of CAD-seq fragments was calculated using the bamCoverage tool from the deepTools suite. BAM files were normalized using the Reads Per Genomic Content (RPGC) method to account for variations in sequencing depth, with an effective hg38 genome size set to 2,913,022,398 bp. The chromosome chrX was excluded from normalization using the --ignoreForNormalization parameter. The --extendReads parameter was used to extend paired-end reads to the exact fragment length defined by the two read mates, ensuring that the true fragment size between paired reads was represented in the coverage profile. Output coverage files were generated in bigWig format with a bin size of 10 bp.

[000306] Genome coverage of fragment midpoints from CAD-seq was calculated using the bamCoverage tool from the deepTools suite. BAM files were normalized using the Reads Per Genomic Content (RPGC)

method to account for sequencing depth variations, with an effective genome size set to 2,913,022,398 bp. The chromosome chrX was excluded from normalization using the --ignoreForNormalization parameter. The --MNase option was used to calculate the midpoint signal of each fragment, rather than extending the reads across the entire fragment length. Output coverage files were generated in bigWig format with a bin size of 1 bp.

[000307] Fragment Length Distribution Analysis

[000308] Fragment length distributions (FLDs) were computed for both whole-genome and ChIP-enriched regions using a custom Python script. The script relies on pysam to parse aligned read data from BAM files and to subset read counts based on genomic coordinates defined in BED files. Fragment lengths were extracted using the fetch method and normalized to frequency per million reads to account for differences in sequencing depth. Normalized FLDs were plotted using matplotlib with both linear and logarithmic scales to visualize enrichment patterns across fragment sizes.

[000309] For analyzing pipeseq alignment pipeline .log files fragment length distributions from DNA sequencing experiments were analyzed using a custom Python script. Insert size histograms were extracted from log files by identifying the relevant histogram section and converting the data into a tabular format. Counts were normalized to counts per million (CPM) to facilitate comparison across technical replicates and experimental conditions. Merging of technical replicates was performed by summing insert size counts, followed by re-normalization to CPM. Analyses were implemented using pandas for data processing and matplotlib and seaborn for visualizations.

[000310] Nucleotide Composition Analysis

[000311] Nucleotide composition was analyzed using a custom Python script to calculate and visualize positional biases within paired-end read fragments. The reference genome was parsed using Bio.SeqIO, and regions of interest were defined by a BEDPE file corresponding to sequenced DNA fragments. For each genomic interval, sequences were extracted for the first 10 bases and the subsequent 20 bases relative to the start and end coordinates, respectively. Nucleotide counts were computed separately for each position, generating position-specific nucleotide profiles for the initial 10 bases (position_counts_10) and the subsequent 20 bases (position_counts_20).

[000312] To account for genome-specific nucleotide composition rather than assuming a uniform distribution, the observed frequency of each nucleotide at a specific position in the first 10 bases was normalized against the average nucleotide frequency across the next 20 bases. The observed/expected ratio was obtained using the formula:

Observed/Expected RatioPosition i=Count of NucleotideiTotal Nucleotidesi∑j=1130Count of Nucleotidej Total Nucleotides20−1\text{Observed/Expected Ratio}_{\text{Position } i} = \frac{\frac{\text{Count of Nucleotide}_{i}}{\text{Total Nucleotides}_{i}}}{\frac{\sum_{j=11}^{30} \text{Count of

Nucleotide}_{j}}{\text{Total Nucleotides}_{20}}} - 1Observed/Expected RatioPosition i =Total Nucleotides20∑j=1130Count of NucleotidejTotal NucleotidesiCount of Nucleotidei−1

Where:

- Count of Nucleotidei\text{Count of Nucleotide}_{i}Count of Nucleotidei is the count of the given nucleotide at position i within the first 10 positions.
- Total Nucleotidesi\text{Total Nucleotides}_{i}Total Nucleotidesi is the total nucleotide count at position i in the first 10 positions.
- ∑j=1130Count of Nucleotidej\sum_{j=11}^{30} \text{Count of Nucleotide}_{j}∑j=1130 Count of Nucleotidej is the sum of nucleotide counts across the subsequent 20 positions.
- Total Nucleotides20\text{Total Nucleotides}_{20}Total Nucleotides20 is the total number of nucleotides across these 20 positions.

[000313] This calculation produces a normalized measure of enrichment or depletion for each nucleotide at each position in the first 10 bases, relative to the genome-specific background composition observed in the subsequent 20 bases. The normalization by the genome-specific baseline ensures that compositional biases inherent to the genome are accounted for, providing a more accurate reflection of positional sequence biases. The analysis was performed for individual nucleotides (A, T, C, G) as well as grouped categories (e.g., AT/GC and purines/pyrimidines). The resulting observed/expected ratios were visualized as bar graphs using matplotlib. Separate visualizations were generated for individual nucleotides and grouped nucleotide categories, showing deviations from the genome-specific background at each position in the first 10 bases.

[000314] GC Content and Coverage Analysis

[000315] Genome-wide GC content and coverage profiles for sequenced fragments were generated using a custom Python script that leverages several specialized libraries, including bioframe for GC content calculation, pyBigWig for extracting coverage values from bigWig files, and pyfaidx for handling FASTA sequences. Chromosome sizes were loaded from a reference file and filtered for standard chromosomes (e.g., chr1, chr2, chrX, chrY) using regular expression matching (re package). For each chromosome, non-overlapping bins of 1 kb were generated, and corresponding nucleotide sequences were extracted using pyfaidx. The GC content for each bin was computed using bioframe.frac_gc, which calculates the fraction of G and C bases for each interval based on the extracted nucleotide sequences.

[000316] Coverage values were obtained from multiple datasets stored in bigWig format, specified as inputs through a configuration file or command-line arguments. For each bigWig file, coverage was extracted in parallel for all defined bins using pyBigWig and joblib for efficient multi-threaded processing. Bins with zero coverage were excluded to focus on regions with meaningful signal. The mean coverage

for each GC content value was calculated, and the data were grouped and smoothed using a rolling window approach (window size = 20 bins) to reduce noise and emphasize trends.

[000317] The results were visualized using matplotlib, with GC content (%) on the x-axis and smoothed mean coverage on the y-axis for each dataset. Each sample was color-coded using a predefined colormap to ensure consistency across visualizations. An optional y-axis constraint was applied to cap coverage values for better comparability.

[000318] CTCF Motif Occupancy Analysis

[000319] CTCF motif occupancy and enrichment were analyzed using a custom Python script that utilizes several specialized packages for different operations. pandas was used to handle BED file inputs and organize the motif intervals into a pandas.DataFrame format using pd.read_csv(). pyBigWig was employed to extract CTCF ChIP-seq or CTCF Cut&Run signal values from bigWig files using pyBigWig.open() and bigWigFile.values() functions, storing the outputs as numpy.array objects for each motif. These arrays were processed using numpy functions such as numpy.sort() and numpy.cumsum() to compute the empirical cumulative distribution function (ECDF) of CTCF binding intensities. The genomic intervals were managed and manipulated using bioframe, which facilitated interval overlaps and metadata integration. Visualizations were created using matplotlib, specifically matplotlib.pyplot.scatter() for ECDF plots.

[000320] To identify the top 5% of motifs based on occupancy, #using AO Cut&Run or ENCODE ChIP-seq# #ENCODE-cite ENCODE paper and list the CTCF ChIP-seq data set number# motifs (JASPAR MA0139.1) were first sorted by their extracted signal values (pandas.DataFrame.sort_values()). The top 5% were then selected using the head() function based on a dynamically calculated threshold (total_motifs * 0.05), which was computed using the total number of motifs present. The resulting subset was formatted as a BED6 file using the columns (chrom, start, end, name, score, strand), where the signal value was assigned as the score. This subset was stored as a new pandas.DataFrame for potential downstream use and export.

[000321] Pairtools

[000322] BAM files were processed through a custom pairtools pipeline to convert raw sequencing data into refined Hi-C contact pairs. Each BAM file was parsed into .pairs.gz format using pairtools parse2, with a minimum MAPQ score of 30 (--min-mapq 30) and additional columns (--add-columns mapq,cigar,matched_bp,algn_ref_span). Parsed pairs were then flipped (pairtools flip), sorted (pairtools sort), and deduplicated (pairtools dedup) in an initial round to remove PCR duplicates. After selecting only "direct" pairs (pairtools select '(walk_pair_type == "R1" or walk_pair_type == "R2" or walk_pair_type == "R1&2")'), flipped and sorted pairs were saved, and a second selection step was performed to isolate unique-unique "UU" pairs. A final round of deduplication was applied to generate high-confidence UU

pairs. Each step was executed as a separate SLURM job with dependencies (--dependency=afterok), ensuring sequential execution. All jobs were run using 16 CPU cores (--cpus-per-task=16), and the results were stored in timestamped directories with detailed log files for reproducibility.

[000323] Matrix File Generation

[000324] Pairtools generated .pairs.gz files were processed using a custom shell script that generates multi-resolution .mcool files using the cooler command-line utilities. Each .pairs.gz file was converted into a 500 bp .cool file using the cooler cload pairs command, with genomic coordinates defined by hg38 chromosome sizes. ICE balancing was applied to the .cool files using cooler balance with 16 CPU cores and the parameter --ignore-diags 1 to exclude interactions along the diagonal. The resulting balanced .cool files were then converted to multi-resolution .mcool files using cooler zoomify with base resolution of 500 bp and zoom level "500N" setting, applying the --balance argument at each resolution level. All scripts were executed on a high-performance computing (HPC) cluster. Output files were logged and documented in a timestamped report.

[000325] P(s) plots

[000326] Contact probability (P(s)) curves were generated using custom Python scripts built on the `cooler` and `cooltools` libraries. Hi-C datasets were processed from `.mcool` files at 500 bp resolution, using `cooltools.expected_cis()` to compute contact value decay (CVD) across cis-chromosomal arms. Chromosome size information (`hg38.chrom.sizes`) was obtained from UCSC Genome Browser, and centromere positions were annotated using `bioframe.fetch_centromeres` (`bioframe` version 0.7.2). Chromosome arm coordinates were generated with `bioframe.make_chromarms` and filtered to exclude non-canonical scaffolds and the mitochondrial genome (chrM) using the chromosomal names present in the CADC dataset (`CADC_RPE1_1_merged_dedup_final_20240822_500_bal_500N_igdiag1.mcool`).

[000327] Balanced Hi-C matrices were extracted at 500 bp resolution, and `cooltools.expected_cis` was used to calculate intra-chromosomal contact probabilities (P(s)) using the chromosome arms as defined view regions. Smoothing and aggregation were applied (`smooth=True`, `aggregate_smoothed=True`) to generate smoothed CVD curves, and the data were merged across replicates by aligning genomic distances and dropping duplicate rows based on base-pair resolution (`s_bp`). The derivative of the smoothed aggregated P(s) curve was computed on the natural log scale for both genomic distance (`ln(s_bp)`) and contact frequency (`ln(balanced.avg.smoothed.agg)`) to assess contact decay patterns.

[000328] Output plots were generated in Jupyter Notebooks (`20240917_Ps_CADC_RPE1_GM12878_K562_1.ipynb` and `20240914_Ps_2.ipynb`) and saved as PDFs. Intermediate data were stored in a serialized format with version-controlled documentation.

[000329] CADwalk PacBio Library Preparation for Sequel IIe

[000330] CAD digestion, proximity ligation, DNA purification

[000331] CAD digestion of 1% FA + 3 mM DSG chemically crosslinked GM12878 cells was performed as described in methods section for CAD-C and CAD-seq. Proximity ligation-generated concatemer DNA products were isolated as in methods section for CAD-C and CAD-seq. Proximity ligation was carried out as for CAD-C samples, skipping end repair and biotin labelling steps.

[000332] **Concatemer size selection**

[000333] Proximity ligation concatemer DNA was separated on a 2% E-Gel EX. Two technical replicates; proximity ligation concatemer DNA: 25.1, 25.9 ng/µL, as determined by Qubit dsDNA BR kit. 500 ng of each technical replicate was loaded per well.

[000334] Concatemers > 1kb in size were excised and purified using Zymo DNA Gel recovery kit, using 3 volumes of ADB buffer to 1 volume gel slice, additionally 1 volume of H2O was added, DNA isolated following the manufacturers protocol. DNA was eluted by sequential elution using 2x15 µL = 30 µL total Zymo DNA elution buffer pre-heated to 65 °C. Isolated concatemer DNA was analyzed by Agilent genomic DNA Tapestation kit and Qubit dsDNA HS assay.

[000335] **NEBNext Ultra II Library Preparation**

[000336] Sample volume was adjusted to 50 µL with 10 mM Tris-HCl pH 8.0. Material was amplified by carrying out NEBNext Ultra II Library Prep with minor modifications detailed below. NEBNext hairpin adapter was diluted 1:10 with 10 mM Tris-HCl pH 8.0, and 2.50 µL of diluted adapter was used per technical replicate. Adapter-ligated DNA was purified by a one-sided Ampure XP cleanup (45 µL beads) to remove adapter-dimers. NEBNext dual-unique indexing primers were used for sample barcoding. PCR protocol was modified; 2x Watchmaker polymerase mastermix was used (15 µL adapter-ligated DNA, 10 µL indexing primers (10 µM), 25 µL 2xWatchmaker PCR mix). Extension time = 5 minutes, 10 cycles, final extension 10 minutes. PCR reactions were purified by 1.2x Ampure XP bead cleanup.

[000337] **PacBio Library Preparation**

[000338] The library was prepared for PacBio Sequel IIe sequencing using a SMRTbell prep kit 3.0 with an amplicon preparation protocol. The library was sequenced with a Hifi workflow.

[000339] **References**

Akgol Oksuz, Betul, Liyan Yang, Sameer Abraham, Sergey V. Venev, Nils Krietenstein, Krishna Mohan Parsi, Hakan Ozadam, et al. 2021. "Systematic Evaluation of Chromosome Conformation Capture Assays." *Nature Methods* 18 (9): 1046–55.

Ageichik, Alexander V., Kumiko Samejima, Scott H. Kaufmann, and William C. Earnshaw. 2007. "Genetic Analysis of the Short Splice Variant of the Inhibitor of Caspase-Activated DNase (ICAD-S) in Chicken DT40 Cells." *The Journal of Biological Chemistry* 282 (37): 27374–82.

Allan, James, Ross M. Fraser, Tom Owen-Hughes, and David Keszenman-Pereyra. 2012. "Micrococcal Nuclease Does Not Substantially Bias Nucleosome Mapping." *Journal of Molecular Biology* 417 (3): 152–64.

O. V. Bylino, A. N. Ibragimov, A. E. Pravednikova, Y. V. Shidlovskii, Investigation of the basic steps in the chromosome conformation capture procedure. *Front. Genet.* **12** (2021).

Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002 Feb 15;295(5558):1306-11. doi: 10.1126/science.1067799. PMID: 11847345.

W. J. de Jonge, M. Brok, P. Kemmeren, F. C. P. Holstege, An optimized chromatin immunoprecipitation protocol for quantification of protein-DNA interactions. *STAR Protoc.* **1**, 100020 (2020).

Fujita Y, Pather SR, Ming GL, Song H. 3D spatial genome organization in the nervous system: From development and plasticity to disease. Neuron. 2022 Sep 21;110(18):2902-2915. doi: 10.1016/j.neuron.2022.06.004. Epub 2022 Jun 30. PMID: 35777365; PMCID: PMC9509413.

Goel, Viraat Y., Miles K. Huseyin, and Anders S. Hansen. 2023. "Region Capture Micro-C Reveals Coalescence of Enhancers and Promoters into Nested Microcompartments." *Nature Genetics* 55 (6): 1048–56.

Hamley, Joseph C., Hangpeng Li, Nicholas Denny, Damien Downes, and James O. J. Davies. 2023. "Determining Chromatin Architecture with Micro Capture-C." *Nature Protocols* 18 (6): 1687–1711.

Hörz W, Altenburger W. Sequence specific cleavage of DNA by micrococcal nuclease. Nucleic Acids Res. 1981 Jun 25;9(12):2643-58. doi: 10.1093/nar/9.12.2643. PMID: 7279658; PMCID: PMC326882.

Hsieh TH, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell. 2015 Jul 2;162(1):108-19. doi: 10.1016/j.cell.2015.05.048. Epub 2015 Jun 25. PMID: 26119342; PMCID: PMC4509605.

Hsieh TS, Fudenberg G, Goloborodko A, Rando OJ. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. Nat Methods. 2016 Dec;13(12):1009-1011. doi: 10.1038/nmeth.4025. Epub 2016 Oct 10. PMID: 27723753.

Hsieh TS, Cattoglio C, Slobodyanyuk E, Hansen AS, Rando OJ, Tjian R, Darzacq X. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell. 2020 May 7;78(3):539-553.e8. doi: 10.1016/j.molcel.2020.03.002. Epub 2020 Mar 25. PMID: 32213323; PMCID: PMC7703524.

Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh TS, Parsi KM, Yang L, Maehr R, Mirny LA, Dekker J, Rando OJ. Ultrastructural Details of Mammalian Chromosome Architecture. Mol Cell. 2020 May 7;78(3):554-565.e7. doi: 10.1016/j.molcel.2020.03.003. Epub 2020 Mar 25. PMID: 32213324; PMCID:

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A,

Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289-93. doi: 10.1126/science.1181369. PMID: 19815776; PMCID: PMC2858594.

Mirny L, Dekker J. Mechanisms of Chromosome Folding and Nuclear Organization: Their Interplay and Open Questions. Cold Spring Harb Perspect Biol. 2022 Jul 1;14(7):a040147. doi: 10.1101/cshperspect.a040147. PMID: 34518339; PMCID: PMC9248823.

Soroczynski J, Risca VI. Technological advances in probing 4D genome organization. Curr Opin Cell Biol. 2023 Oct;84:102211. doi: 10.1016/j.ceb.2023.102211. Epub 2023 Aug 7. PMID: 37556867; PMCID: PMC10588670.

Slobodyanyuk E, Cattoglio C, Hsieh TS. Mapping Mammalian 3D Genomes by Micro-C. Methods Mol Biol. 2022;2532:51-71. doi: 10.1007/978-1-0716-2497-5_4. PMID: 35867245.

Tettey TT, Rinaldi L, Hager GL. Long-range gene regulation in hormone-dependent cancer. Nat Rev Cancer. 2023 Oct;23(10):657-672. doi: 10.1038/s41568-023-00603-4. Epub 2023 Aug 3. PMID: 37537310. Korn, Christian, Sebastian Richard Scholz, Oleg Gimadutdinow, Alfred Pingoud, and Gregor Meiss. 2002. "Involvement of Conserved Histidine, Lysine and Tyrosine Residues in the Mechanism of DNA Cleavage by the Caspase-3 Activated DNase CAD." *Nucleic Acids Research* 30 (6): 1325–32.

Luse, Donal S., Mrutyunjaya Parida, Benjamin M. Spector, Kyle A. Nilson, and David H. Price. 2020. "A Unified View of the Sequence and Functional Organization of the Human RNA Polymerase II Promoter." *Nucleic Acids Research* 48 (14): 7767–85.

J. S. McCarty, S. Y. Toh, P. Li, Multiple domains of DFF45 bind synergistically to DFF40: roles of caspase cleavage and sequestration of activator domain of DFF40. *Biochem. Biophys. Res. Commun.* **264**, 181–185 (1999).

Meiss, G., S. R. Scholz, C. Korn, O. Gimadutdinow, and A. Pingoud. 2001. "Identification of Functionally Relevant Histidine Residues in the Apoptotic Nuclease CAD." *Nucleic Acids Research* 29 (19): 3901–9.

Naughton, Catherine, Covadonga Huidobro, Claudia R. Catacchio, Adam Buckle, Graeme R. Grimes, Ryu-Suke Nozawa, Stefania Purgato, Mariano Rocchi, and Nick Gilbert. 2022. "Human Centromere Repositioning Activates Transcription and Opens Chromatin Fibre Structure." *Nature Communications* 13 (1): 1–16.

H. Sakahira, M. Enari, S. Nagata, Functional differences of two forms of the inhibitor of caspase-activated DNase, ICAD-L, and ICAD-S. *J. Biol. Chem.* **274**, 15740–15744 (1999).

Scott, Daniel J., Natalie J. Gunn, Kelvin J. Yong, Verena C. Wimmer, Nicholas A. Veldhuis, Leesa M. Challis, Mouna Haidar, Steven Petrou, Ross A. D. Bathgate, and Michael D. W. Griffin. 2018. "A Novel Ultra-Stable, Monomeric Green Fluorescent Protein for Direct Volumetric Imaging of Whole Organs Using CLARITY." *Scientific Reports* 8 (1): 1–15.

Shilling, Patrick J., Kiavash Mirzadeh, Alister J. Cumming, Magnus Widesheim, Zoe Köck, and Daniel O. Daley. 2020. "Improved Designs for pET Expression Plasmids Increase Protein Production Yield in Escherichia Coli." *Communications Biology* 3 (1): 214.

Spector, Benjamin M., Mrutyunjaya Parida, Ming Li, Christopher B. Ball, Jeffery L. Meier, Donal S. Luse, and David H. Price. 2022. "Differences in RNA Polymerase II Complexes and Their Interactions with Surrounding Chromatin on Human and Cytomegalovirus Genomes." *Nature Communications* 13 (1): 2006.

Vera Rodriguez, Arturo, Steffen Frey, and Dirk Görlich. 2019. "Engineered SUMO/protease System Identifies Pdr6 as a Bidirectional Nuclear Transport Receptor." *The Journal of Cell Biology* 218 (6): 2006–20.

Widlak, P., and W. T. Garrard. 2001. "Ionic and Cofactor Requirements for the Activity of the Apoptotic Endonuclease DFF40/CAD." *Molecular and Cellular Biochemistry* 218 (1-2): 125–30.

Widlak, Piotr, and William T. Garrard. 2006a. "The Apoptotic Endonuclease DFF40/CAD Is Inhibited by RNA, Heparin and Other Polyanions." *Apoptosis: An International Journal on Programmed Cell Death* 11 (8): 1331–37.

Widlak, P., and W. T. Garrard. 2006b. "Unique Features of the Apoptotic Endonuclease DFF40/CAD Relative to Micrococcal Nuclease as a Structural Probe for Chromatin." *Biochemistry and Cell Biology = Biochimie et Biologie Cellulaire* 84 (4): 405–10.

Xiao, Fei, Piotr Widlak, and William T. Garrard. 2007. "Engineered Apoptotic Nucleases for Chromatin Research." *Nucleic Acids Research* 35 (13): e93.

EXAMPLE 2

[000340] To further evaluate CAD, we conducted a comparison of the CAD enzyme versus micronuclase (MNase). Mieczkowski et al have previously reported MNase titration studies in evaluating differences between nucleosome occupancy and chromatin assembly (Mieczkowski J et al (2016) Nat Commun 7:11485; doi: 10.1038/ncomms11485). In their studies, they demonstrated that some nucleosomes are seen preferentially at high MNase and some at low MNase, and coined this MNase accessibility (MACC). Digestion assays with MNase were noted as being prone to technical issues that can hinder interpretation and sample-to-sample comparison of the results (Tolstorukov, M. Y., Kharchenko, P. V. & Park, P. J. Epigenomics 2, 187–197 (2010); Zhang, Z. & Pugh, B. F. Cell 144, 175–186 (2011)). Further, MNase displays sequence preference and its output is sensitive to even minor variation in enzyme activity (Dingwall, C., Lomonossoff, G. P. & Laskey, R. A. Nucleic Acids Res. 9, 2659–2673 (1981); Chung, H. R. et al. PLoS ONE 5, e15754 (2010); Weiner, A. et al Genome Res. 20, 90–100 (2010)). They evaluated and quantified chromatin accessibility across genomes of various complexity by utilizing different MNase concentrations.

[000341] We conducted CAD digestion and generated data in hTERT-RPE1 and compared our results to the reported results and titration provided in Mieczkowski et al. The sequenced DNA fragment coverage of all length-sized fragments around the top 5% CTCF-occupied CTCF motifs in the human hg38 genome was evaluated. The 'Reads Per Genome Coverage" (RPGC) was normalized with 1.0 being equal to sequence coverage expected from overall sequencing depth if coverage is perfectly uniform genome wide. The results are depicted in FIGURE 15. CAD was titrated from 2.5U to 250U. MNase was titrated and compared at 5.4U to 304 U. A noticeable and significant fluctuation in the baseline coverage was demonstrated with MNase titration in contrast to CAD.

[000342] The sequenced DNA fragment coverage of sub-100 basepair length fragments around the top 5% CTCF-occupied CTCF motifs in the human hg38 genome was evaluated. The 'Reads Per Genome Coverage" (RPGC) was normalized with 1.0 being equal to sequence coverage expected from overall sequencing depth if coverage is perfectly uniform genome wide. The results are depicted in FIGURE 16. CAD was titrated from 2.5U to 250U. MNase was titrated and compared at 5.4U to 304 U. Preservation of sub-100 bp CTCF-bound protected DNA was noted in CAD titration in contrast to MNase.

[000343] Fragment length distributions were assessed with CAD vs. MNase. Normalized fragment length distributions (log scale) for each are shown in FIGURE 17. CAD digestion produces more uniformly-sized chromatin fragments across 100-fold CAD concentration range and practically identical results over a 10-fold CAD concentration range, reflecting CAD's lack of exonuclease activity. This is in contrast to MNase which progressively erodes protein-associated DNA in a sensitive concentration-dependent manner.

[000344] Genome coverage as a function of GC content was evaluated. The smoothed mean coverage of CAD or MNase digest fragments as a function of GC% is provided in FIGURE 18 with CAD and MNase at different titrated Unit concentrations as before. CAD digested chromatin sequencing shows a much more uniform coverage than MNase, even across a 100-fold change in enzyme concentration and provides indistinguishable results over a 10-fold change in enzyme concentration. Further, there is an equilibration of coverage upon more complete digestion using higher CAD concentrations, in contrast to MNase titration where coverage profiles are distinct. This reflects MNase's preferential fragmentation of AT-rich DNA at lower MNase concentration, followed by over-digestion of AT-rich DNA and its under-representation in higher MNase concentrations. GC-rich DNA sequence coverage follows a converse relationship with MNase concentration. This exemplifies the tolerance of CAD digestion to experimental variations in its concentration relative to cell input, which is advantageous for robust, reproducible data generation from various sample inputs.

EXAMPLE 3

[000345] We are continuing to optimize and benchmark CAD-C against existing Hi-C and Micro-C technologies, including in commonly-used mammalian cell lines (including GM12878 LCLs). Additional studies are applying CAD digestion and CAD-C to understand local effects of cohesin loop extrusion and further benchmarking and fine-tuning CAD-C against Hi-C and Micro-C. Also, applying CAD-C to small cell number inputs and primary tissues, including low input primary tissue samples.

[000346] For instance, single-cell micro-C has been recently reported (Wu, H. et al *bioRxiv* 2023.02.18.529096 (2023) doi:10.1101/2023.02.18.529096). Given the advantages of CAD-C single-cell CAD-C is possible. In this regard, we are profiling single-cell heterogeneity enabled by lack of overdigestion, ligation efficiency, higher signal-to-noise with CAD-C. This has application in view of many existing Tn5 indexing approaches. Also, a recent scNanoHi-C (single cell nano Hi-C) paper described combining RE Hi-C concatemer nanopore sequencing and single cell indexing (Li, W. et al. *Nat. Methods* **20**, 1493–1505 (2023)).

[000347] We are sequencing ligated fragment concatemers directly using CAD-C and long-read sequencing. Long-read sequencing can be used to study multi-way contacts, to uncover complex regulatory neighborhoods within chromosomes. Long read sequencing also aids in de novo genome assembly sequencing and can provide a direct DNA methylation readout. This is intrinsically single-cell, single-allele even if done in bulk due to low interchromosomal contact rate (~3%). This can provide haplotype-phased contacts, useful for understanding effect of eQTLs, structural variants, epigenetic mechanisms and imprinting disorders. Utility of these methods, approaches and applications has already shown for RE Hi-C HiPore-C (Chen, Y. et al. *bioRxiv* 2023.08.29.555243 (2023); doi:10.1101/ 2023.08.29.555243). It has been reported that hTERT-RPE1 cells have haplotype resolved SNPs to extract Hi-C contacts for Xi and Xa (Darrow, E. M. et al. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4504–12 (2016)). Pore-C is a protocol for multi-way contact mapping using nanopore long-read sequencing, described by the Imielinski Lab (Deshpande AS et al Nat Biotechnol. 2022 Oct;40(10):1488-1499. doi: 10.1038/s41587-022-01289-z. Epub 2022 May 30. PMID: 35637420)

[000348] Other CAD-C based applications are in footprinting of nucleosomes, TFs. CAD-C can be used to connect TF occupancy to nucleosome positioning, chromosome folding within a single allele. This is depicted in Figure 25. CADwalk fragments reproduce nucleosome positioning and chromatin footprinting patterns. This permits for more information to be extracted from a single chromosome. CAD-C can be used to build better mechanistic models, understanding mutually exclusive chromosome structure states.

[000349] CAD-C can be combined with single-protein mapping using m6A, as done in DiMeLo-seq, which is a long-read, single-molecule method for mapping protein-DNA interactions genome wide (Altemose, N. et al. *Nat. Methods* **19**, 711–723 (2022)). CAD-C may be used in labeling the positions of single cohesin molecules within the concatemer that otherwise could not be unambiguously inferred from

footprinting data. It can also be used for orienting local folding and footprinting information within the concatemers in context of nuclear bodies labeling.

[000350] CAD-C can be combined with single-cell indexing and long read sequencing. This is in line with the reported Tn5-based PacBio library prep by Ramani lab (Nanda AS et al (2024) Nat Genet 56(6):1300-1309). Single-cell, single-allele, DNA-methylation, nucleosome+TF footprints, and single protein localisation can be combinatorially assessed and determined within one measurement. Previous attempts have been undertaken with SCA-seq to combine SAMOSA with regular pore-C and restriction enzymes for digestion, but the data and results achieved were noisy and the scientists noted that a much higher sequencing throughput is required to achieve analysis at a single-molecule level in a specific spatial location (Xie, Y. et al. *Elife* (2023) doi:10.7554/elife.87868).

[000351] CAD-C in an embodiment in which biotin labeling of ligated ends is not used can be combined with newly replicated DNA labeling using EdU incorporation (KR Stewart-Morgan *et al.*, 2019, *Mol. Cell)* followed by conjugation of biotin to the incorporated EdU using click chemistry. This can be used to analyze the three-dimensional DNA contacts within newly replicated chromatin.

[000352] CAD-C has additional applications, including in cell-free DNA analysis. Researchers have reported using cell-free DNA to map nucleosomes and infer tissue of origin (Snyder MW et al. Cell. 2016 Jan 14;164(1-2):57-68. doi: 10.1016/j.cell.2015.11.050. PMID: 26771485; PMCID: PMC4715266; Underhill HR et al. PLoS Genet. 2016 Jul 18;12(7):e1006162. doi: 10.1371/journal.pgen.1006162. PMID: 27428049; PMCID: PMC4948782). Control activated CAD enzyme and CAD-C methodologies would provide an alternative and improved approach for cell-cell free DNA analysis and evaluation of low concentration and rare DNAs in samples.

[000353] In addition, CAD has applicability in other chromatin research applications, including for example in cryoEM. Cryo-EM is widely used to determine high-resolution three-dimensional (3D) structures of proteins and other protein-bound complexes such as protein-DNA complexes and protein-RNA complexes (Arimura, Y., Shih, R.M., Froom, R., and Funabiki, H. (2021). Structural features of nucleosomes in interphase and metaphase chromosomes. Mol. Cell *81*, 4377-4397; Carragher, B., Cheng, Y., Frost, A., Glaeser, R.M., Lander, G.C., Nogales, E., and Wang, H.W. (2019). Current outcomes when optimizing 'standard' sample preparation for single-particle cryo-EM. J. Microsc. *276*, 39–45).

EXAMPLE 4

Analysis of CAD-C properties and application to multiple cell lines.

[000354] The above studies describe a CAD-C library produced from hTERT-RPE-1 retinal pigment epithelium cells. We have performed further analysis comparing CAD digestion against MNase digestion,

particularly against titrated MNase data from MACC (Mieczkowski et al., Nature Communications, 2016, PMID 27151365). We observed that the structure of CAD, which produces blunt double-stranded DNA cleavage events (Figure 3) gives rise to a fragment size distribution that is less dependent on the enzyme concentration than MNase (Figure 10), has less size bias at fragment ends (Figure 11), and generates more even genomic coverage as a function of the GC nucleotide content of the genomic locus (Figure 12). We see that nucleosome-sized fragments are more efficiently generated at open chromatin near CTCF-bound sites (Figure 13), but sub-nucleosomal sized fragments are better retained with CAD at different concentrations of enzyme (Figure 15, data not shown). Even at sparse coverage, we see nucleosome-sized reads forming clear peaks at single loci, which can be useful for nucleosome mapping. Transcription start sites and CTCF sites show small fragments being retained at nucleosome-free regions (data not shown).

[000355] To determine how CAD cleavage is affected by epigenetic states, we mapped CAD fragments over regions marked with four histone modifications representative of heterochromatin (H3K9me3 or H3K27me3) or active regulatory regions or genes in euchromatin (H3K27ac or H3K4me1). The histone marks were mapped using CUT&Tag. Rescaling the regions to determine CAD-generated fragment coverage relative to the adjacent genomic baseline, we found that CAD cleavage for high enzyme concentrations is depleted by ~16% in H3K9me3-marked heterochromatin, enhanced by ~15% in active regulatory chromatin, and nearly unchanged in H3K27me3-marked facultative heterochromatin (Fig. 16).

[000356] We have compared CAD-C to Micro-C read pair statistics and contact matrices. We observe similar performance when read pairs are filtered for only the ones that have a detectable ligation junction . CAD-C has a higher percentage of contacts below 1 kb distance when detectable ligation junctions are not required (data not shown).

[000357] In addition to hTERT-RPE-1 cells, we have also applied CAD cleavage controls and CAD-C to GM12878 (lymphoblastoid cell line) and K562 (leukemia) cells. We produced usable libraries on both cell lines without further titration of the CAD concentration and the libraries were of good enough quality in terms of complexity and fragment size distribution to sequenced on the first attempt. GM12878 and K562 cells are grown in suspension and represent very different cell types as compared to hTERT-RPE-1 cells, demonstrating the robustness of the protocol to different substrates. The characterization of CAD cleavage in GM12878 cells shows that although the CAD-digested fragment coverage is more skewed toward GC-rich parts of the genome as compared to hTERT-RPE-1 cells (data not shown), the coverage distribution does not depend on the amount of enzyme used. We therefore interpret the difference in genomic region representation as most likely a feature of different chromatin landscapes or different responses to fixation in the two cell types, rather than being an artifact of the enzyme to cell ratio as is the case for MNase . As with hTERT-RPE-1 cells, we also observe that sub-nucleosomal DNA fragments (<100 bp) are recovered even at high CAD enzyme concentrations (data not shown).

[000358] Having shown that the CAD cleavage works well in GM12878 cells as well, we proceeded to prepare CAD-C libraries from GM12878 and K562 cells, and compared their contact probability curves between cell types as well as comparing GM12878 cells against available Omni-C (a variation on Hi-C with restriction enzymes) and Micro-C data from the same cell line (Figure 20). When applying a standard alignment and filtering pipeline that does not require the direct detection of proximity ligation junctions (Figure 17), we find that the contact probability curves differ between methods primarily in the sub-kilobase regime, where CAD-C retains a larger number of fragments (Figure 18A). This is in line with the non-processive cleavage of CAD and the generation of blunt DNA ends. Both activities are likely to increase the retention of nucleosomal and sub-nucleosomal fragments with ligatable ends and to enhance the efficiency of ligation of nearby DNA fragment ends—most of which are likely to lie within a kilobase of each other on the genome. Comparing CAD-C contact probability curves between cell lines, we find that the curves are largely consistent, with the primary difference occurring at the megabase length scale. K562 cells are a cancer cell line, whereas the other two cell lines are immortalized normal tissue. The discrepancy at the megabase length scale may represent a different heterochromatin/euchromatin organization.

[000359] Lastly, we investigated how the orientation-specific contact probability curve at the kilobase length scale, which reflects local nucleosome-nucleosome contacts driven by the folding of the chromatin fiber and the availability of CAD-cleaved DNA ends for ligation (data not shown). Ligations between nucleosomes can occur in four possible orientations with respect to the nucleosomal DNA fragment's position in the genome. Nucleosome entry-to-entry (back flip or tandem entry), exit-to-exit (front flip or tandem exit), entry-to-exit (back), or exit-to-entry (front). We observe an alternation of peaks in the fragment distribution between nucleosome fragment multiples (of ~150 bp) and peaks with lengths consisting of nucleosome multiples plus a shorter sub-nucleosomal fragment, which we interpret to be a part of the inter-nucleosome linker DNA ligated to a nucleosomal fragment. Peaks that consist of a separation distance that corresponds to a nucleosomal linker must be formed by the ligation of a nucleosome edge to another nucleosome edge. Their 10 bp periodic modulation in read occurrence frequency as a function of fragment length is consistent with the nucleosomal DNA "breathing" away from the nucleosome's histone core in 10 bp increments. In contrast to Micro-C, therefore, we are able to detect different classes of ligation events: between nucleosomal fragments (jagged, 10-bp-periodic modulation peaks) and between linker DNA, with a nucleosome or nucleosome multiple in between (smooth peaks) (data not shown). Genome-wide separation distances between the ligated fragment ends (determined by identifying the ligation junction in each chimeric concatamer) were plotted as a function of orientation. Library molecules in which the ligation junction could not be directly detected were excluded. Front flip and back flip curves were nearly identical and coincide (data not shown). We are

investigating how these orientation-specific contact probability curves differ between epigenetic states in which we expect the chromatin fiber to vary in its folding.

EXAMPLE 5

Variation on the CAD-C method without end-repair and biotinylation

[000360] The CAD-C method described above and in the prior examples was performed using biotinylation of cleaved fragment ends prior to proximity ligation, following the approach used in Micro-C and Hi-C. Unlike MNase, CAD cleavage gives rise to blunt ends that should not require DNA end-repair prior to ligation. We hypothesized that skipping the end-repair step may increase the efficiency of CAD-C. Indeed, we demonstrated (Figure 18) that the end-repair step is not necessary and proximity-ligated DNA fragments, which are shifted toward longer fragment sizes, appear with nearly equal probability with and without end-repair.

[000361] There are two conclusions we can draw from this finding. First, the blunt ends produced by CAD are readily ligatable and do not require end repair. Second, the high cleavage density of CAD between nucleosomes produces such a high percentage of short fragments that can be ligated into informative concatamers (whose ligation junctions report on 3-D contacts), that most fragments in the final ligated library contain ligation junctions. Therefore, it is no longer necessary to enrich for those ligation junctions using biotin fill-in and biotin pull-down of the junctions, as in Hi-C. These two advantages greatly enhance the efficiency and applicability of the method.

EXAMPLE 6

Successful long-read "nucleosome walks" showing multi-chromosome contacts

[000362] We size-selected the CADwalk libraries (shown in Figure 21A) to retain nucleosome fragment concatamers above 1 kb and prepared them for sequencing using PacBio Hifi sequencing (Figure 18B).

[000363] We aligned HiFi reads using local alignment with minimap2 to determine the individual concatemer segments that align to the genome (schematic shown in Figure 19). We defined a CADwalk to be the series of consecutive steps between genome-aligned ligated fragments, which we take to represent a series of consecutive or adjacent three-dimensional chromatin contacts. The resulting data showed that the fragment size distribution retained the "ladder" distribution dominated by nucleosome-sized fragments we observed after CAD cleavage, prior to proximity ligation. We characterized the cis-chromosomal and trans-chromosomal contacts represented by the steps in a CADwalk and found that most contacts were cis-chromosomal, with the probability of a trans-chromosomal jump increasing gradually, with an inflection point of walks of 6 segments or more (data not shown). The high fraction of walk lengths that consisted of 2-4 alignments indicates that medium-length sequencing achievable on the Illumina platform

(250 x 250 bp paired end) should also be enough to observe three-mer concatamers and obtain high-throughput coverage of short CADwalks. We are further pursuing this option via the sequencing data.

[000364] We predicted that we should be able to determine loops from a walk by finding adjacent steps that alternate direction but have a similar size. The conditional contact probability of adjacent CADwalk steps shows that certain loop sizes are preferred when step n and step n+1 have similar sizes, indicating a loop contact. We observe an enrichment of such loops around 20,000 bp and 300,000 bp, likely representing different classes of loops (tens of kb) or topologically associating domains (hundreds of kb) within the genome. Our studies and findings using and applying CADwalks are depicted in Figures 20-25.

[000365] Conclusions

We conclude that:

1. CAD-C is a robust method that is applicable to multiple cell types for the determination of three-dimensional DNA contacts,
2. CAD-C is particularly information-rich and efficient for the mapping of orientation-specific chromatin contacts on the kilobase scale, and
3. CAD-C can be optimized to remove end-repair and biotinylation steps, creating an efficient workflow for the generation of multi-contact CADwalk libraries that can be read out with long-read sequencing to reveal cis- and trans-chromosomal contact chains and to provide structural information about the nature of chromatin contacts at loop anchors via the conditional step size distributions between adjacent steps in a walk.

EXAMPLE 7

[000366] In the above examples, we have demonstrated the applicability of CAD-C to multiple cell types, both cancer-derived and immortalized normal tissue, and both adherent and suspension cell lines.

[000367] The demonstration that the efficiency and high density of CAD cleavage and ligation obviate the need for both end repair and biotin-based capture of proximity-ligated junctions encouraged us to apply CAD to another application: the development of a CAD-C based protocol for the mapping of three-dimensional chromatin contacts in nascent chromatin. In a first step, we incorporated ethinyldeoxyuridine (EdU) in replicating cells using a short pulse, to label newly replicated DNA. A standard CAD protocol is then followed, with the modification of a click reaction to couple biotin to the incorporated EdU and a biotin-capture step to pull down concatamers from the labeled nascent chromatin. This is made possible because biotin does not need to be used for ligation junction enrichment.

[000368] This invention may be embodied in other forms or carried out in other ways without departing from the spirit or essential characteristics thereof. The present disclosure is therefore to be considered as

in all aspects illustrated and not restrictive, the scope of the invention being indicated by the appended Claims, and all changes which come within the meaning and range of equivalency are intended to be embraced therein.

[000369] Various references are cited throughout this Specification, each of which is incorporated herein by reference in its entirety.

What is claimed is:

1. A controllably activated Caspase Activated DNase (CAD), wherein the CAD is activated by an enzyme other than caspase, including other than caspase-3.

2. The CAD of claim 1, wherein the CAD is activated by protease digestion of the inhibitor of CAD (ICAD) with a protease with high sequence specificity, wherein the protease is not caspase and wherein the protease recognized sequence is replaced for caspase recognized amino acid sequence in ICAD.

3. The CAD of claim 1 or 2, wherein the CAD is activated by protease digestion of the CAD inhibitor ICAD with TEV protease.

4. An engineered ICAD, having caspase enzyme recognition sequence deleted and replaced with alternative protease recognition sequence whereby the engineered ICAD is cleaved by a specific alternative protease and is not cleaved by caspase.

5. The engineered ICAD of claim 4, wherein caspase recognition sequence selected from DEP, DET and/or DAV is deleted and is replaced with a protease recognition sequence for an alternative protease.

6. The engineered ICAD of claim 4 or 5, wherein caspase recognition sequence is replaced with protease recognition sequence for a protease selected from TEV protease, enteropeptidase, thrombin, Factor Xa, Rhinovirus 3C protease, HRV 3C protease, and also carboxypeptidase A and B, DAPase, PreScission protease, bdSENP1, bdNEP1, ssNEDP1, scAtg4 and xlUsp2.

7. The engineered ICAD of any of claims 4-6, wherein caspase recognition sequence is replaced with TEV protease recognition sequence.

8. The engineered ICAD of any of claims 4-7, wherein caspase recognition sequence selected from DEP, DET and/or DAV is replaced with TEV protease recognition sequence ENLYFQX, wherein X is S, G, A, M, C or H (SEQ ID NO:10).

9. The engineered ICAD of any of claims 4-8, wherein caspase recognition sequence selected from

DEP, DET and/or DAV is replaced with TEV protease recognition sequence ENLYFQX, wherein X is S or G (SEQ ID NO:9).

10. The engineered ICAD of any of claims 4-9, wherein caspase recognition sequence is replaced with TEV protease recognition sequence ENLYFQS (SEQ ID NO:8) or SEQ ID NO:9 or SEQ ID NO:10.

11. The engineered ICAD of any of claims 4-8, comprising the amino acid sequence of SEQ ID NO:1 or SEQ ID NO:2, or a variant thereof capable of forming a heterodimer with CAD and having amino acid sequence at least 90% identical to SEQ ID NO: 1 or SEQ ID NO:2 and retaining the TEV protease recognition sequence ENLYFQS (SEQ ID NO:8).

12. A controllably activatable engineered ICAD/CAD heterodimer comprising the CAD of any of claims 1-3 and the engineered ICAD of any of claims 4-11.

13. A composition comprising the engineered ICAD/CAD heterodimer of claim 12.

14. The heterodimer of claim 12 or the composition of claim 13, comprising a CAD comprising SEQ ID NO: 5 or SEQ ID NO: 6 or a variant CAD having at least 90% amino acid sequence identity to SEQ ID NO:5 or SEQ ID NO:6, and an engineered ICAD comprising SEQ ID NO: 1 or SEQ ID NO: 2 or a variant CAD having at least 90% amino acid sequence identity to SEQ ID NO: 1 or SEQ ID NO: 2.

15. An isolated nucleic acid encoding the CAD of any of claims 1-3, the engineered ICAD of any of claims 4-11, and/or the controllably activatable engineered ICAD/CAD heterodimer of claim 12.

16. A vector or plasmid comprising nucleic acid of claim 15.

17. The vector of claim 16, wherein CAD and engineered ICAD are co-expressed.

18. The vector of claim 17, wherein CAD and engineered ICAD are co-expressed from a single inducible promoter.

19. The CAD of claims 1-3 or the engineered ICAD of claims 4-11, wherein the CAD or ICAD is covalently attached to or expressed having a tag or label.

20. The CAD or ICAD of claim 19, wherein the tag or label is biotin, streptavidin, poly-His or GFP.

21. A method for capturing the sequence and structure (three dimensional conformation) of one or more target regions of one or more genome in a sample comprising:
(a) crosslinking chromatin from a sample to preserve the genome sequence and structure;
(b) introducing a CAD enzyme and digesting the crosslinked chromatin with CAD to generate digested ends of DNA, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase;
(c) labeling the digested ends of DNA;
(d) ligating the digested and labeled ends of DNA to generate ligated DNA; and
(e) purifying the ligated DNA to generate proximally-ligated DNA.

22. The method of claim 21 further comprising:
(f) fragmenting the proximally-ligated DNA; and
(g) capturing the proximally-ligated DNA.

23. The method of claim 21 or 22, comprising introducing in step (b) the controllably activatable engineered ICAD/CAD heterodimer of claim 12 or 14.

24. The method of claim 21 or 22, comprising further introducing in step (b) an engineered ICAD of any of claims 4-11.

25. The method of any of claims 21-24, wherein the step (c) labeling and/or repairing the digested ends of DNA is omitted.

26. The method of any of claims 21-24, wherein in step (g) the capturing is achieved by enriching for proximally-ligated DNA via the label of (c).

27. The method of any of claims 21-26, wherein the proximally-ligated DNA is sequenced.

28. The method of any of claims 21-27, wherein the purified or captured proximally-ligated DNA is used to prepare a library of the proximally-ligated DNA.

29. The method of claim 28, wherein the library of the proximally-ligated DNA is sequenced in its entirety or in part.

30. The method of claim 21 or 22, wherein the fragmentation of (f) fragments DNA to generate fragments of an average fragment size of <1 kb, 500-800bp, or of an average fragment size >400bp.

31. The method of claim 21 or 22, wherein the purifying of (e) and/or the enriching of (g) is via beads, optionally via magnetic beads.

32. The method of claim 21 or 22, wherein the purified proximally ligated DNA, comprising a chimeric molecule ("CADwalk") of multiple proximity-ligated genomic DNA fragments is size selected for fragments of >1kb and sequenced with long-read methods.

33. The method of any of claims 21-32, wherein the engineered ICAD/CAD heterodimers are activated by one or more alternative protease prior to combining with a sample for evaluation.

34. A method for detecting special proximity relationships between nucleic acid sequences comprising contacting a sample containing nucleic acid with a CAD enzyme and digesting the nucleic acid with CAD to generate digested ends of nucleic acid, wherein the CAD enzyme is controllably activated by digestion of ICAD with an enzyme other than caspase.

35. The method of claim 34, wherein the ICAD is an engineered ICAD and comprises an engineered ICAD of any of claims 4-11.

36. A system or kit for determining sequence and structure (three dimensional conformation) of one or more target regions or of one or more genome in a sample comprising nucleic acid, comprising an engineered ICAD and CAD, wherein the engineered ICAD is an engineered ICAD of any of claims 4-11.

37. The system or kit of claim 36, further comprising:
(a) a ligation enzyme for ligating the digested nucleic acid to generate ligated nucleic acid; and
(b) a means for purifying the ligated nucleic acid to generate proximally-ligated nucleic acid;
(c) a means for capturing the proximally-ligated nucleic acid.

38. The system or kit of claim 36 or 37, further comprising a means for labeling the nucleic acid digested by the CAD.

39. The system or kit of any of claims 36, 37 or 38, further comprising reagents for sequencing the proximally-ligated nucleic acid.

40. The system or kit of claim 35-39, further comprising reagents for generating a library of the proximally-ligated nucleic acid.

FIGURE 1

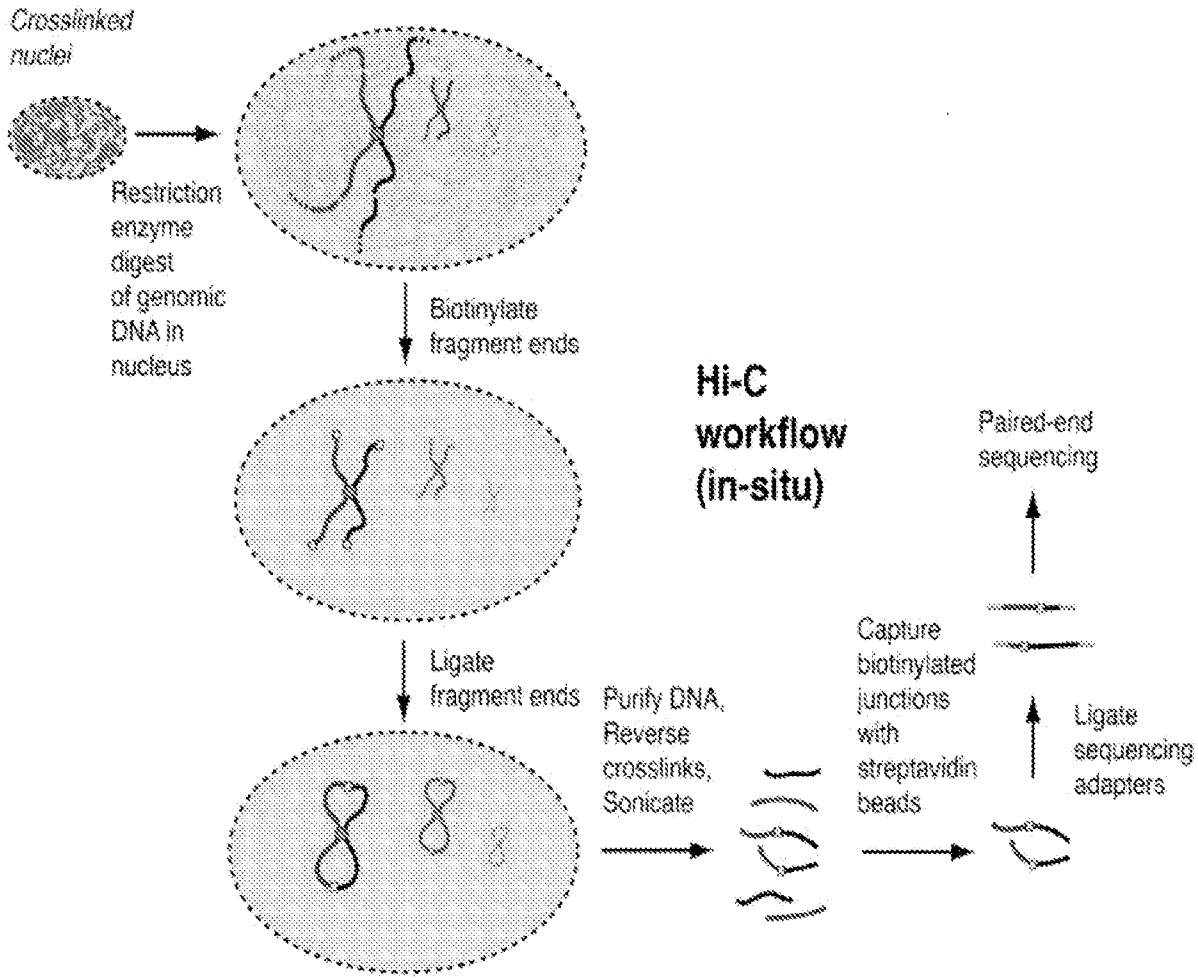
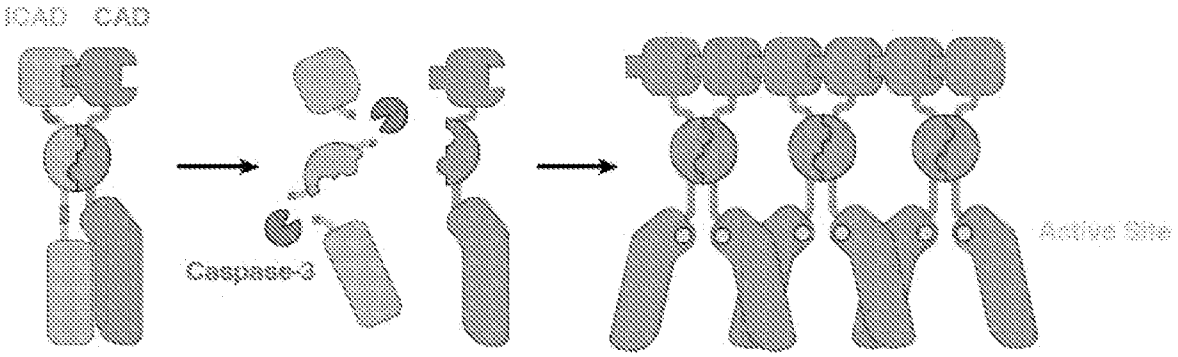

Crosslinked nuclei
Restriction enzyme digest of genomic DNA in nucleus
Biotinylate fragment ends
Ligate fragment ends
Purify DNA, Reverse crosslinks, Sonicate
Hi-C workflow (in-situ)
Capture biotinylated junctions with streptavidin beads
Ligate sequencing adapters
Paired-end sequencing

FIGURE 2

A



B

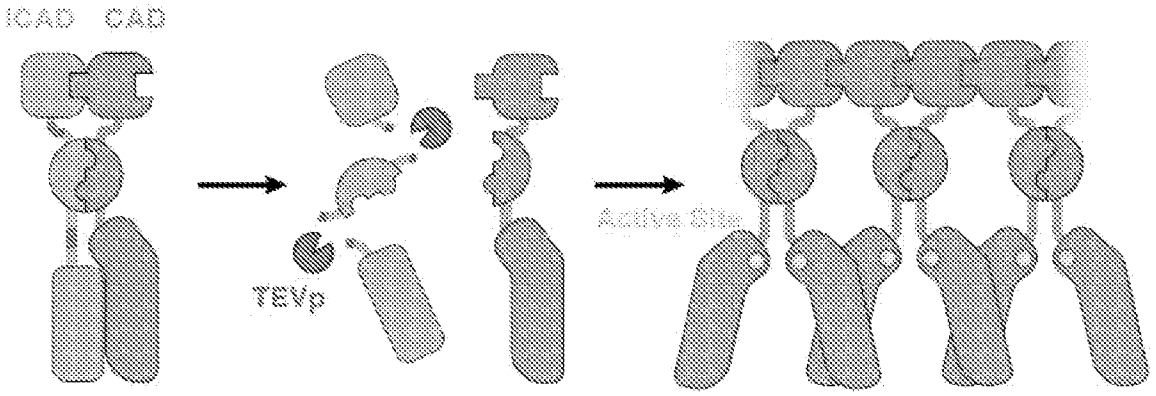


FIGURE 3

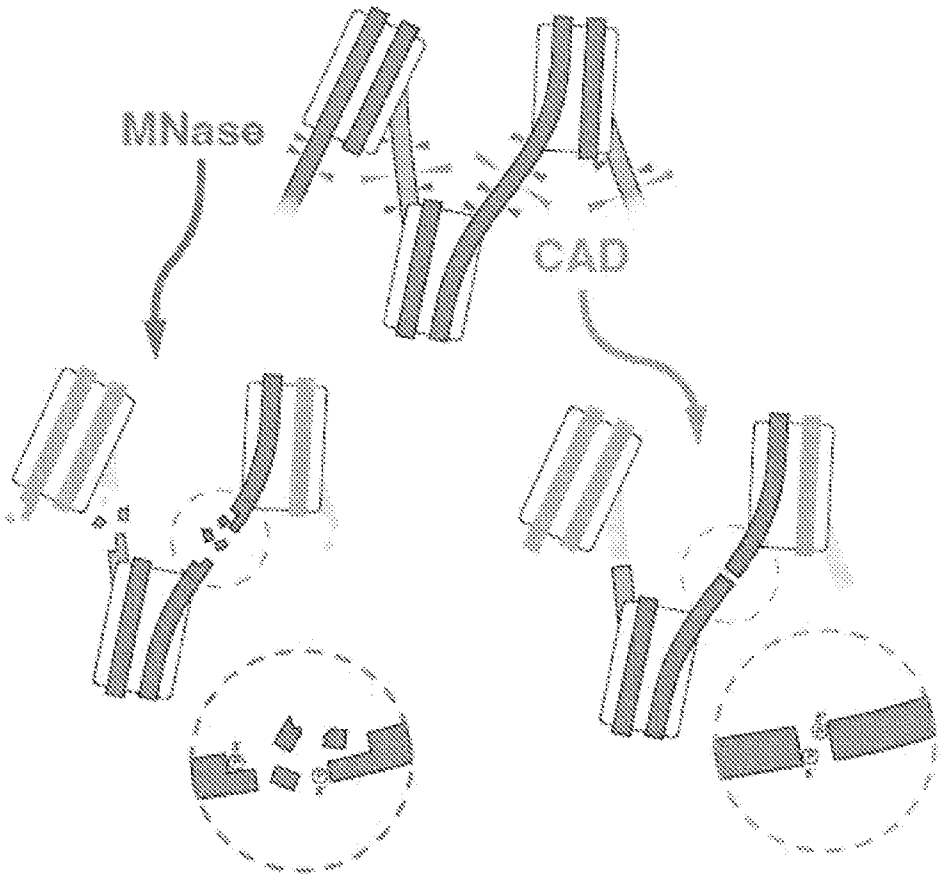


FIGURE 4

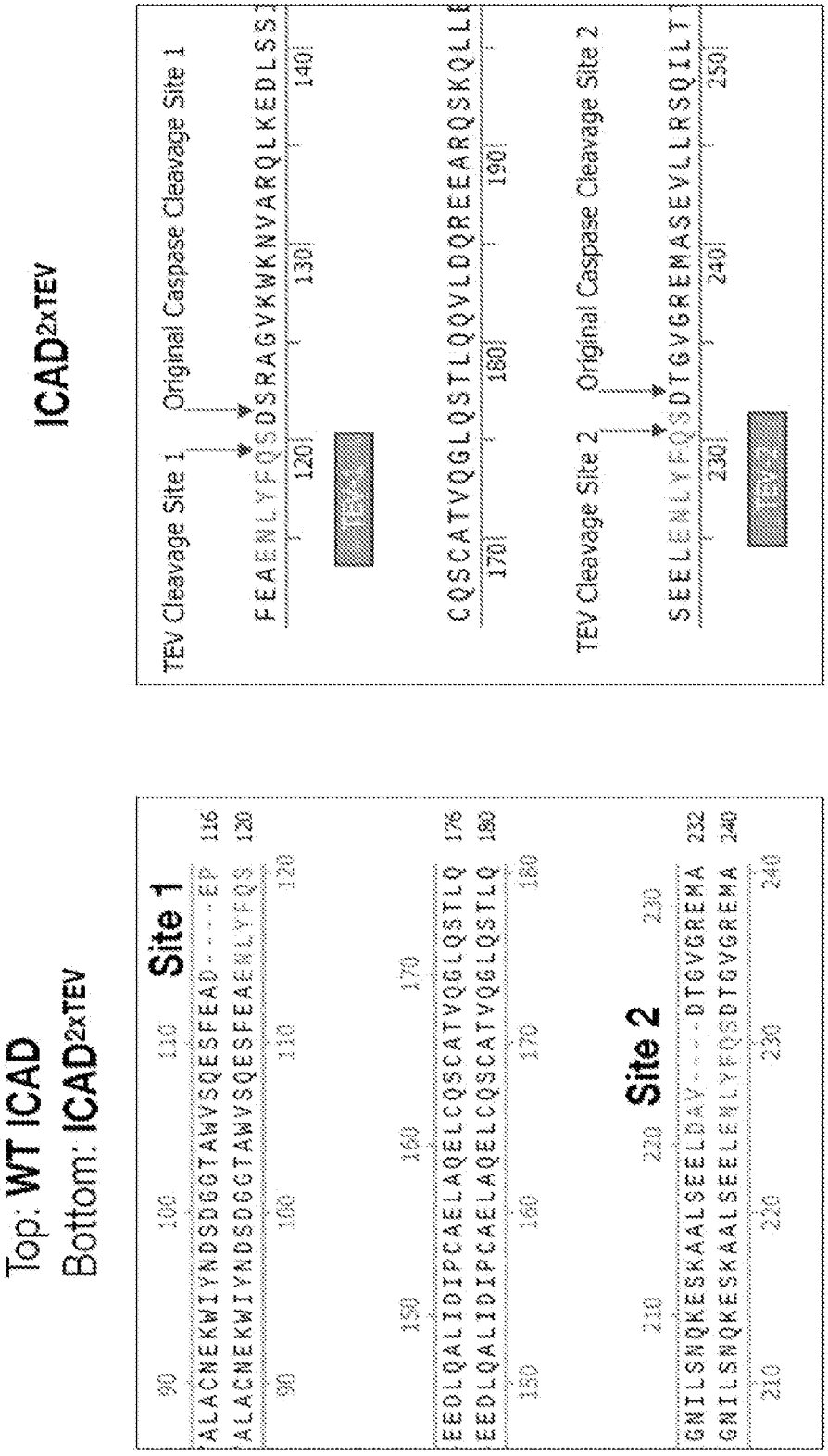

ICAD2xTEV
TEV Cleavage Site 1
Original Caspase Cleavage Site 1
TEV Cleavage Site 2
Original Caspase Cleavage Site 2
Top: WT ICAD
Bottom: ICAD2xTEV
Site 1
Site 2

FIGURE 5

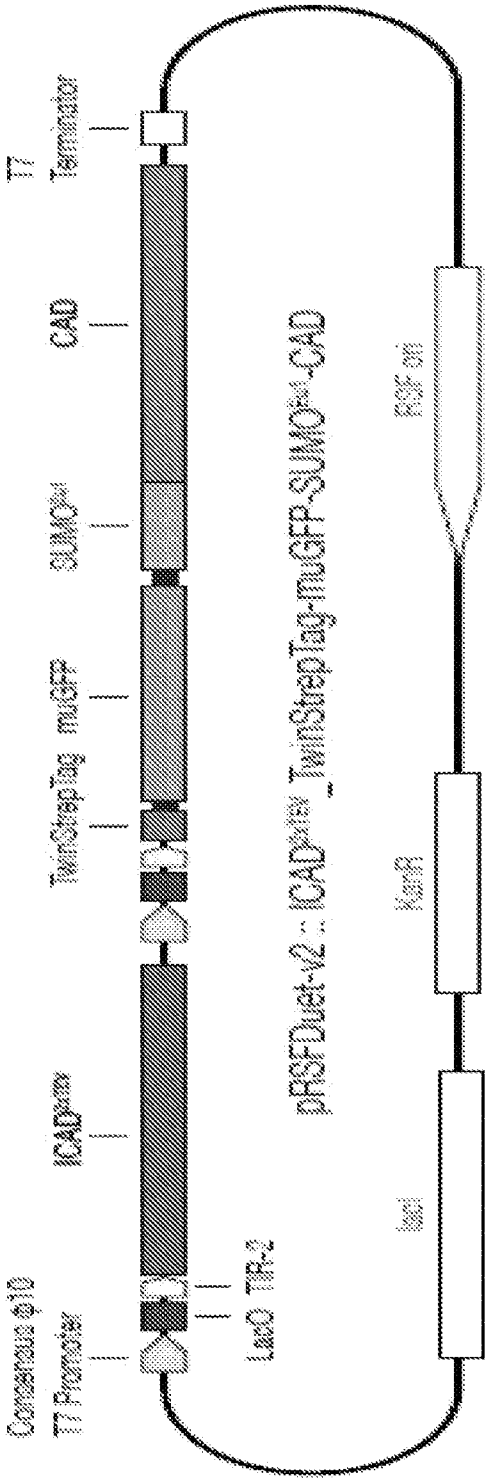
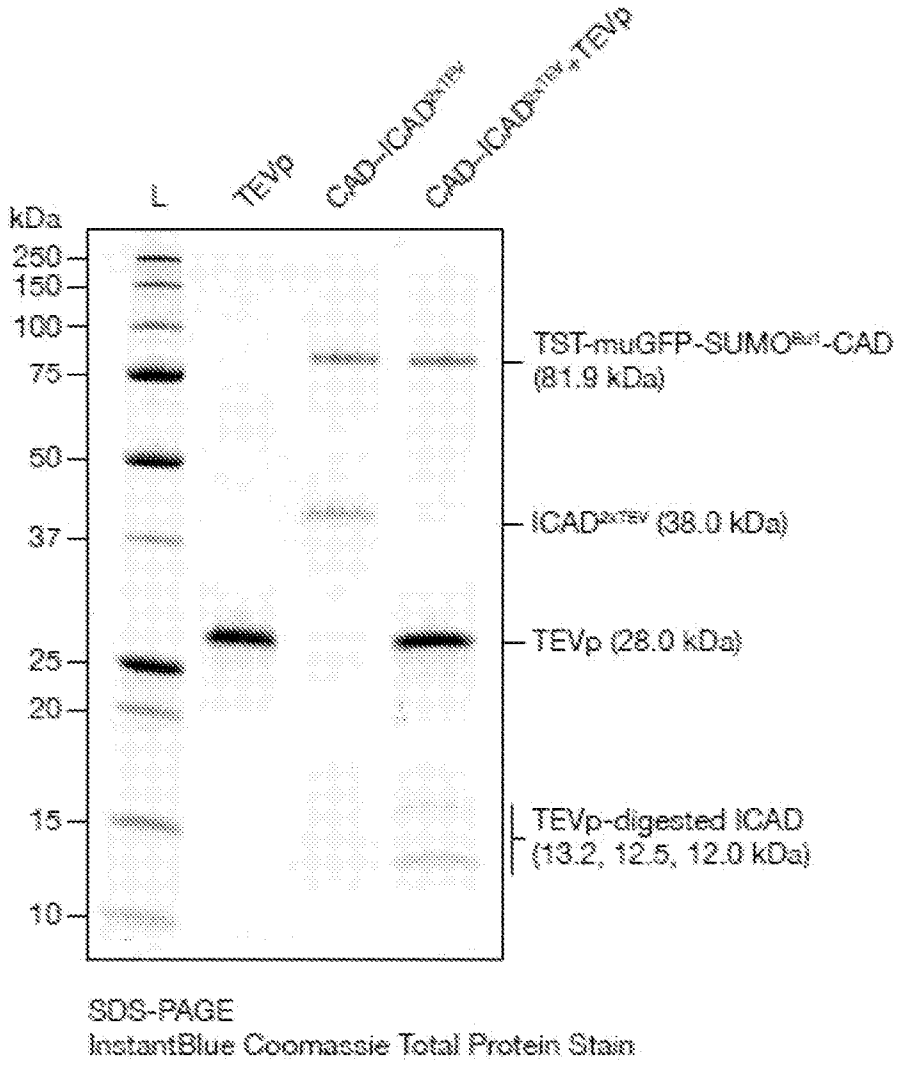

Consensus φ10
T7 Promoter
ICAD-m
LacO TR-2
TwinStrepTag
muGFP
SUMO-1
CAD
T7
Terminator
pRSFDuet-v2 :: ICADnew_TwinStrepTag-muGFP-SUMO-1-CAD
lacI
KanR
RSF ori

FIGURE 6



SDS-PAGE
InstantBlue Coomassie Total Protein Stain

FIGURE 7

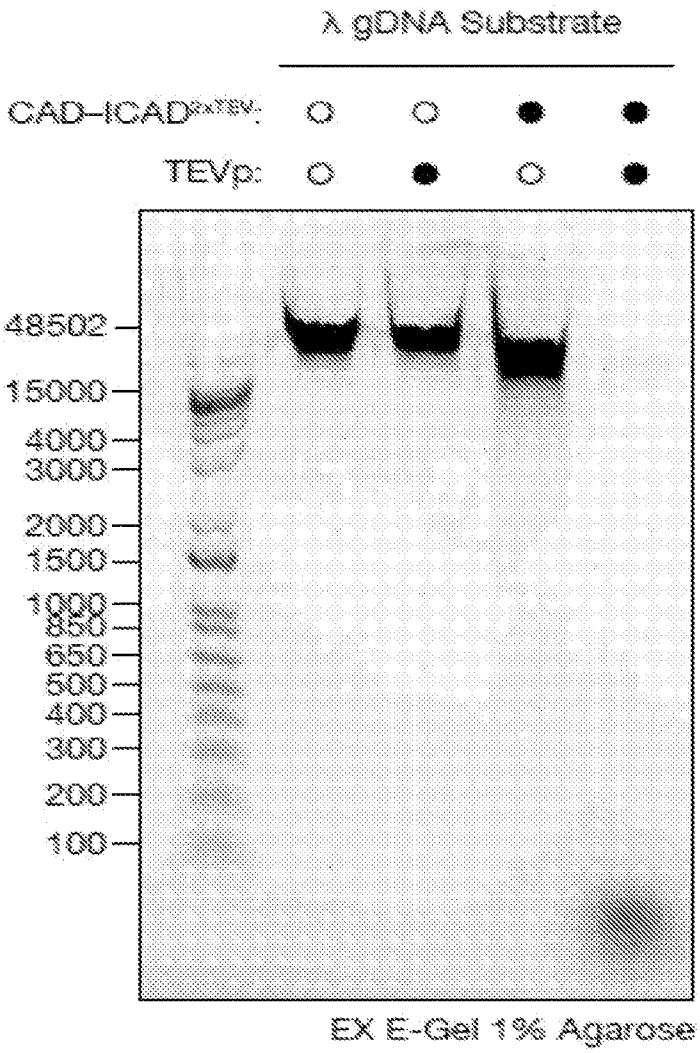

λ gDNA Substrate
CAD–ICAD^RaTEV:
TEVp:
48502
15000
4000
3000
2000
1500
1000
850
650
500
400
300
200
100
EX E-Gel 1% Agarose

FIGURE 8

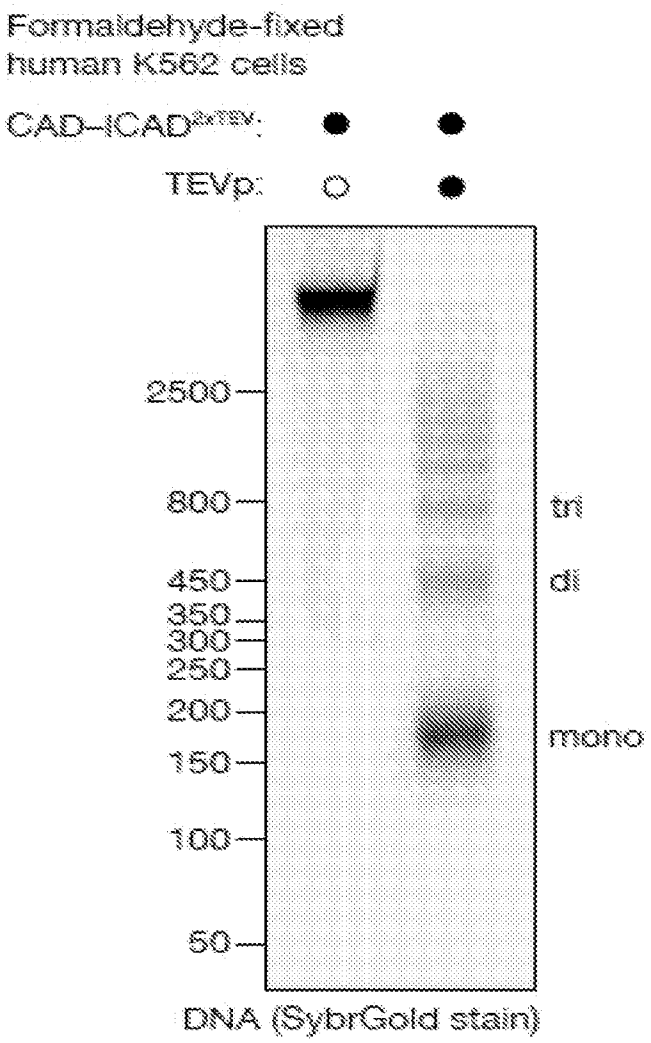

Formaldehyde-fixed
human K562 cells
CAD–ICAD^{3xTEV}:
TEVp:
2500
800
450
350
300
250
200
150
100
50
tri
di
mono
DNA (SybrGold stain)

FIGURE 9

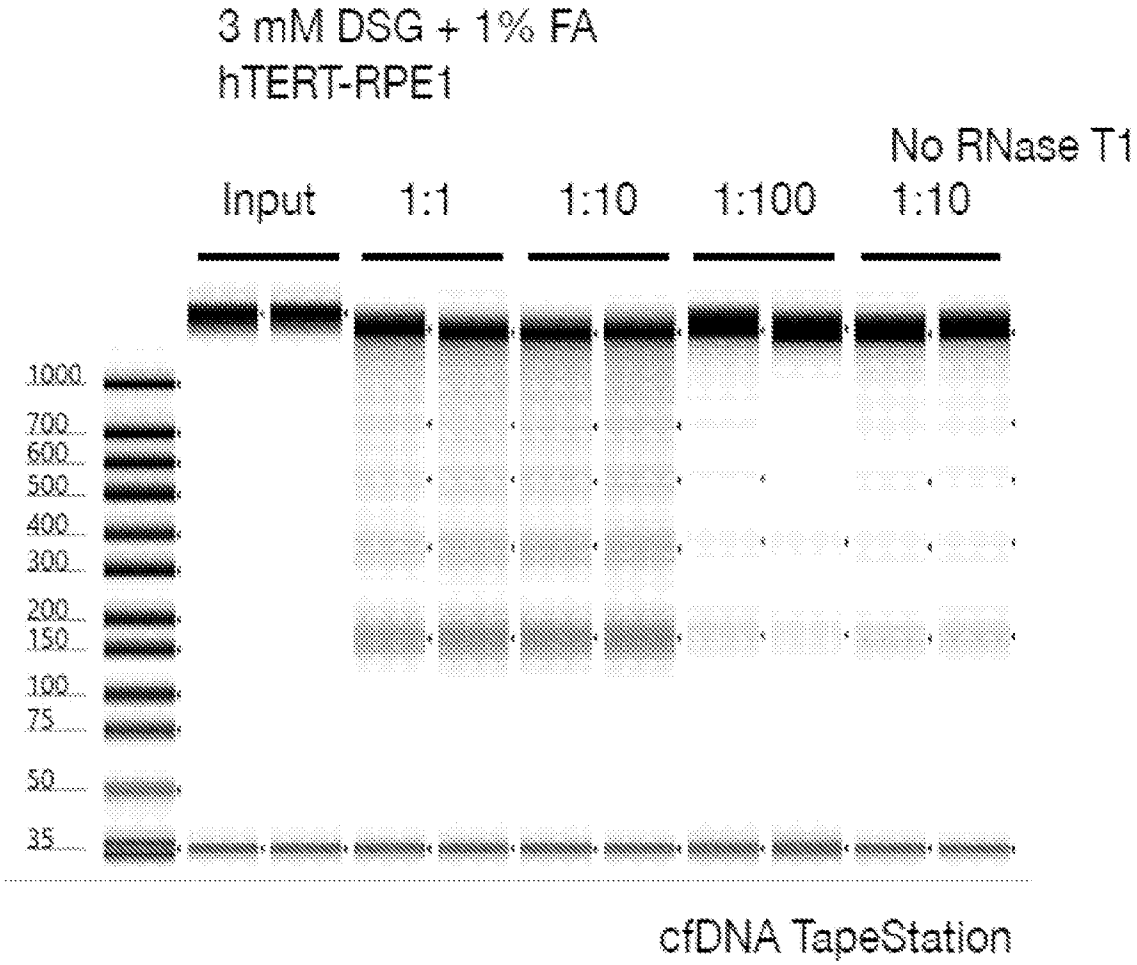

3 mM DSG + 1% FA
hTERT-RPE1
No RNase T1
Input
1:1
1:10
1:100
1:10
1000
700
600
500
400
300
200
150
100
75
50
35
cfDNA TapeStation

FIGURE 10

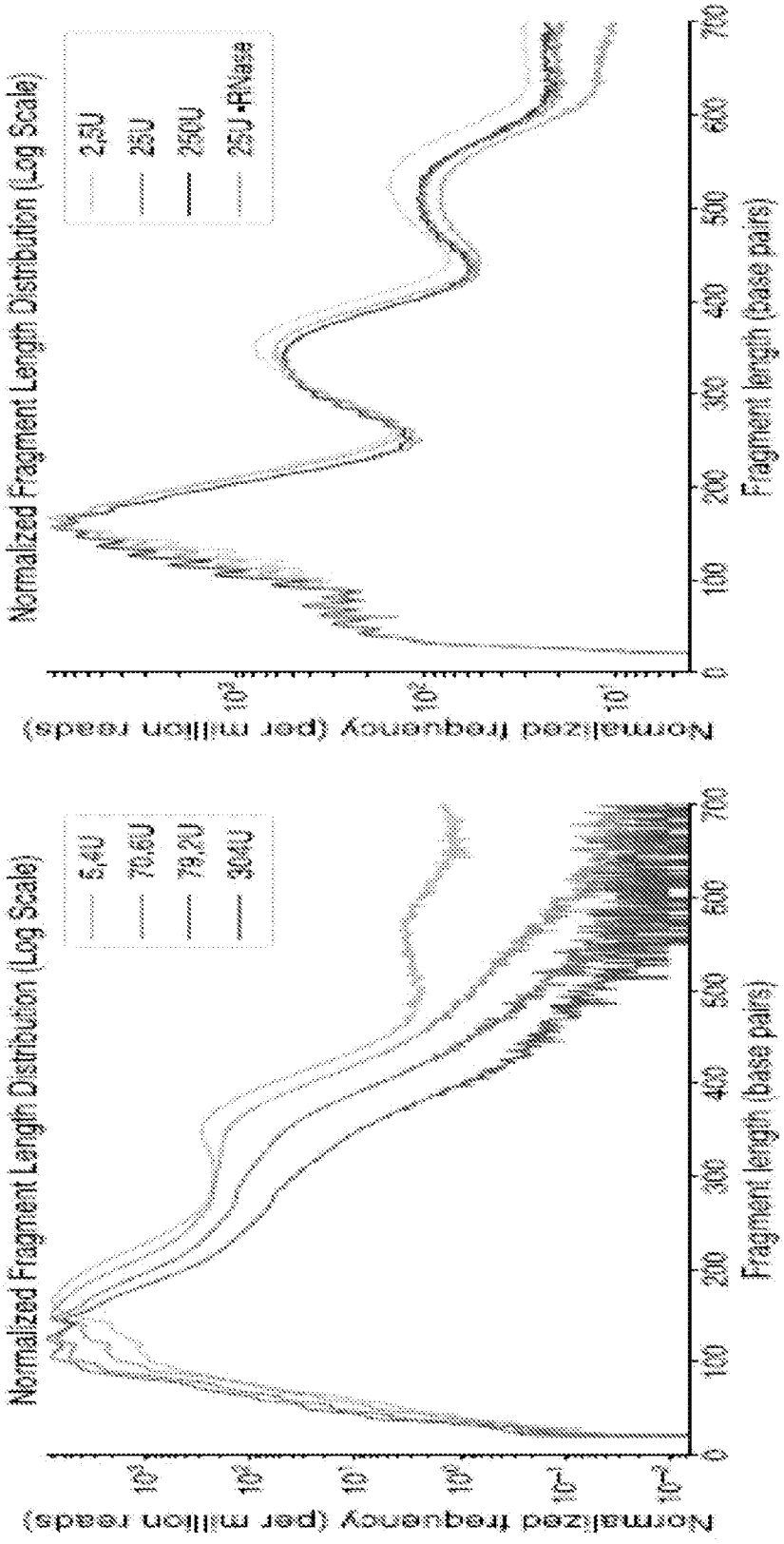

Normalized Fragment Length Distribution (Log Scale)
2,5U
25U
250U
25U +RNase
Normalized frequency (per million reads)
Fragment length (base pairs)
Normalized Fragment Length Distribution (Log Scale)
6,4U
70,6U
79,2U
304U
Normalized frequency (per million reads)
Fragment length (base pairs)

FIGURE 11

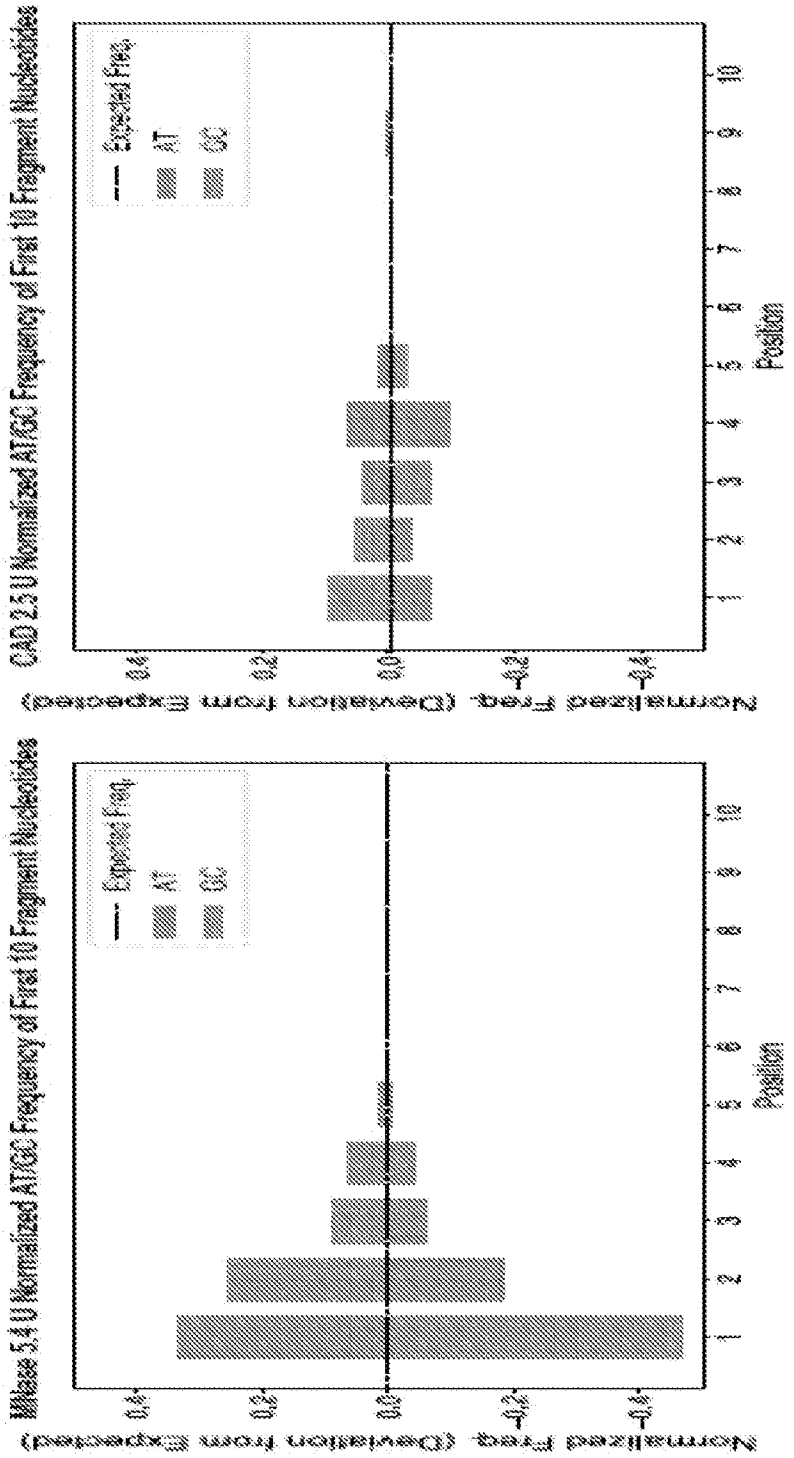

CAD 2.5 U Normalized AT/GC Frequency of First 10 Fragment Nucleotides
Normalized Freq. (Deviation from Expected)
Position
Expected Freq.
AT
GC
Illkase 5.4 U Normalized AT/GC Frequency of First 10 Fragment Nucleotides
Normalized Freq. (Deviation from Expected)
Position
Expected Freq.
AT
GC

FIGURE 12

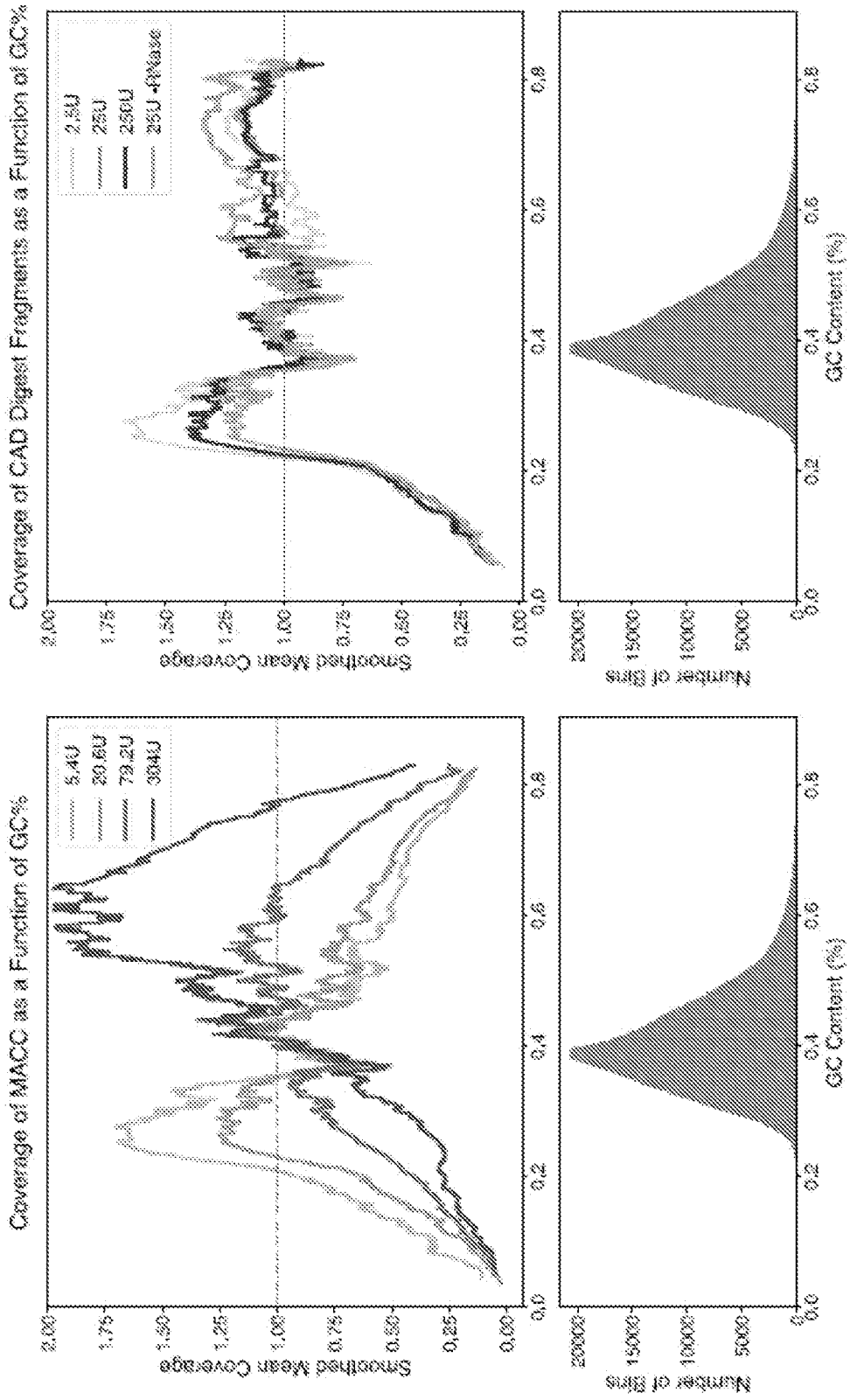

Coverage of MACC as a Function of GC%
Coverage of CAD Digest Fragments as a Function of GC%
Smoothed Mean Coverage
Number of Bins
GC Content (%)
5.4U
20.6U
79.2U
304U
2.5U
25U
250U
25U +Ppase

FIGURE 13

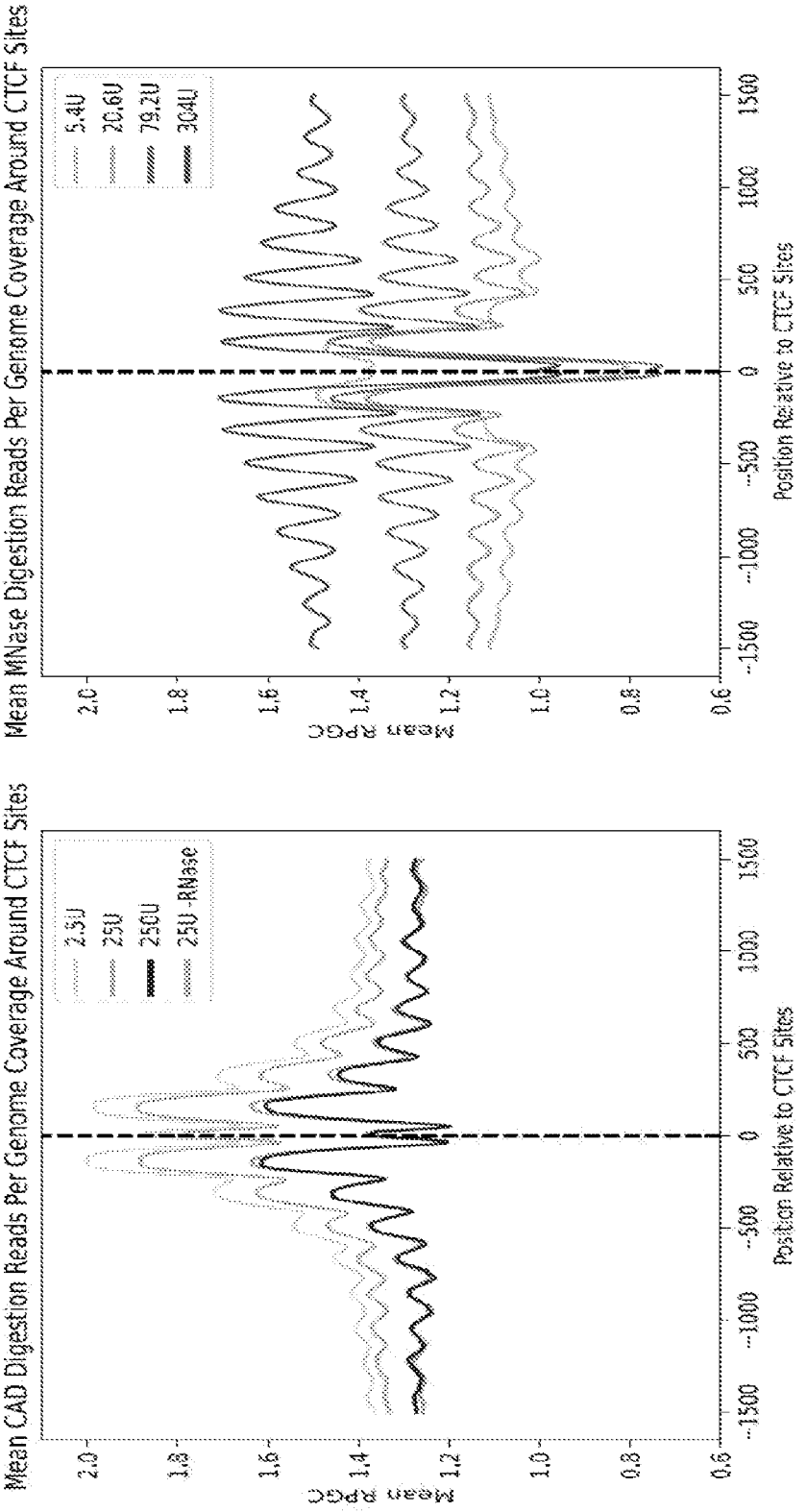

Mean MNase Digestion Reads Per Genome Coverage Around CTCF Sites
5.4U
20.6U
79.2U
304U
Mean RPGC
Position Relative to CTCF Sites
Mean CAD Digestion Reads Per Genome Coverage Around CTCF Sites
2.5U
25U
250U
25U -RNase
Mean RPGC
Position Relative to CTCF Sites

FIGURE 14

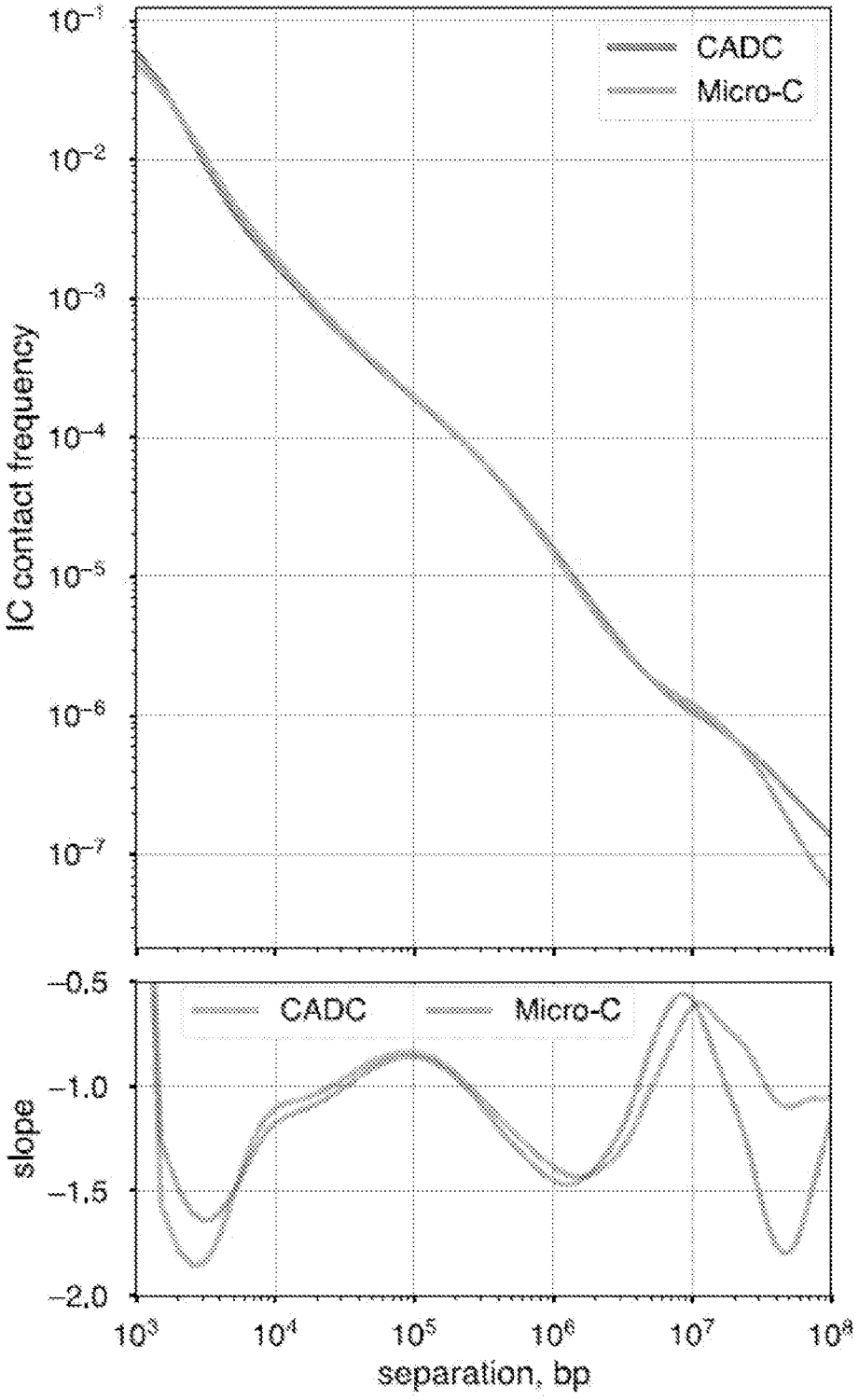


FIGURE 15

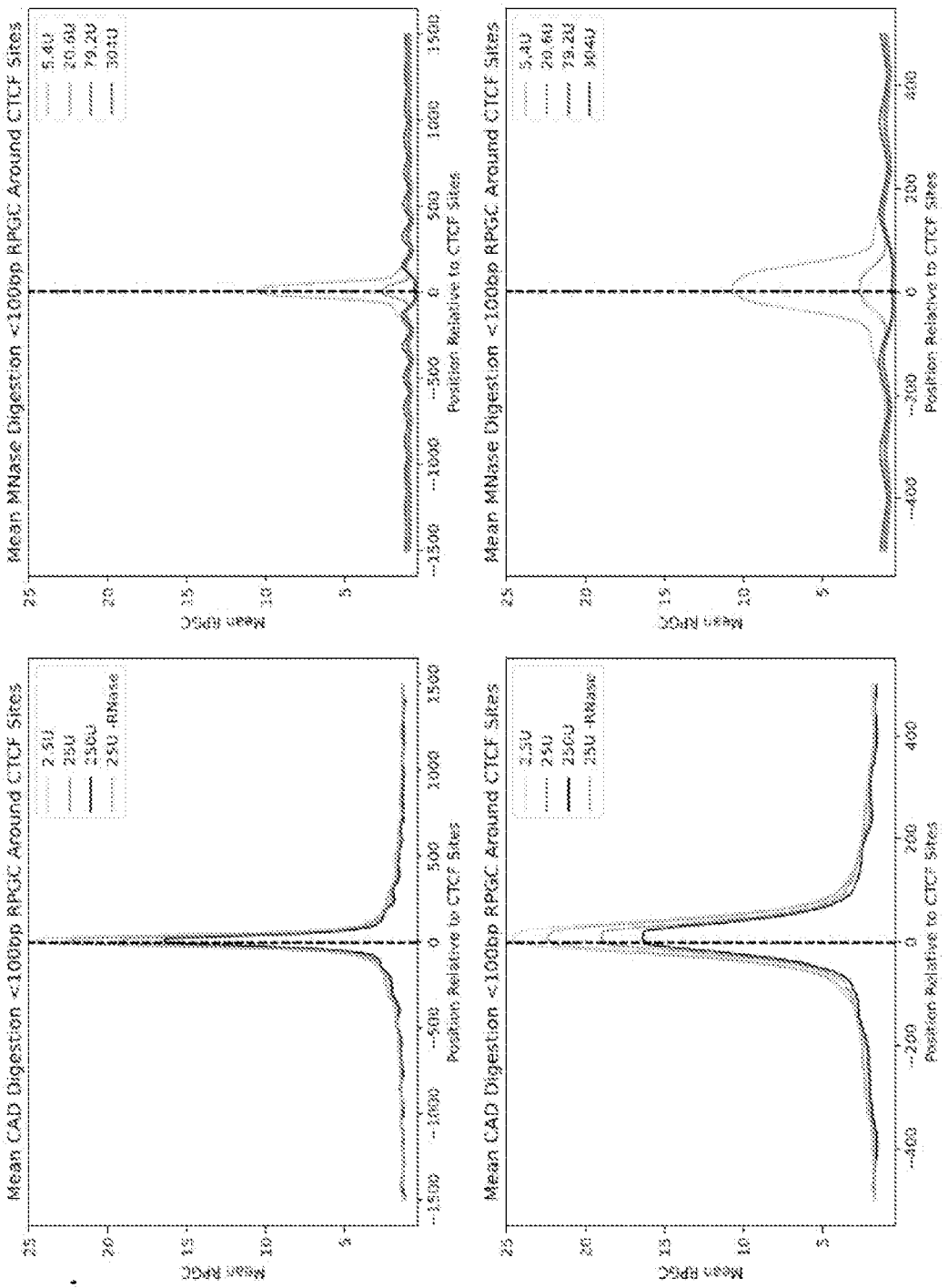

Mean CAD Digestion <1000bp RPGC Around CTCF Sites
Mean MNase Digestion <1000bp RPGC Around CTCF Sites
Mean RPGC
Position Relative to CTCF Sites
2.5U
25U
250U
25U -RNase
5.4U
20.6U
79.2U
304U

FIGURE 16

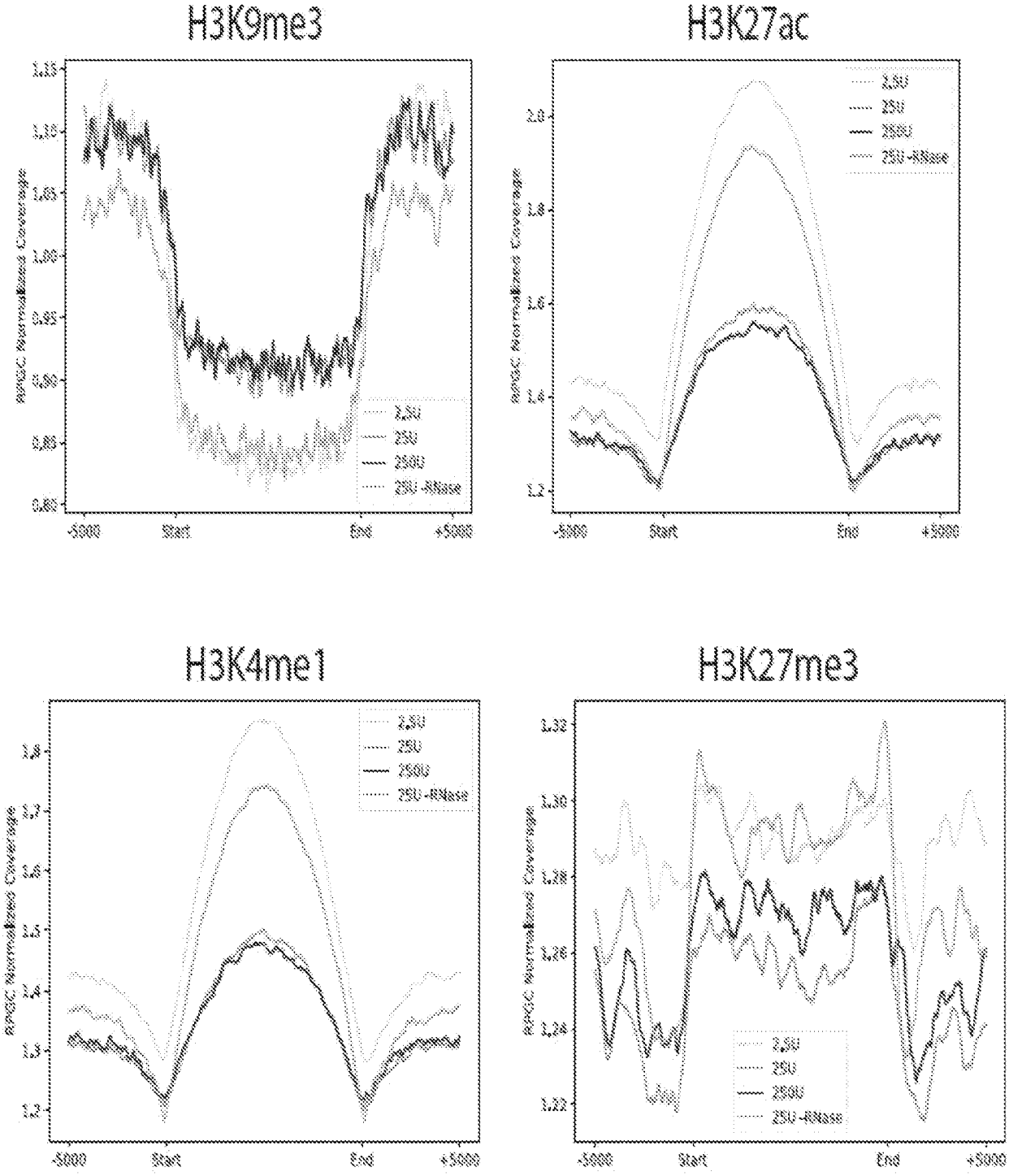

H3K9me3
H3K27ac
H3K4me1
H3K27me3
RPGC Normalized Coverage
-5000
Start
End
+5000
2.5U
25U
250U
25U-RNase

FIGURE 17

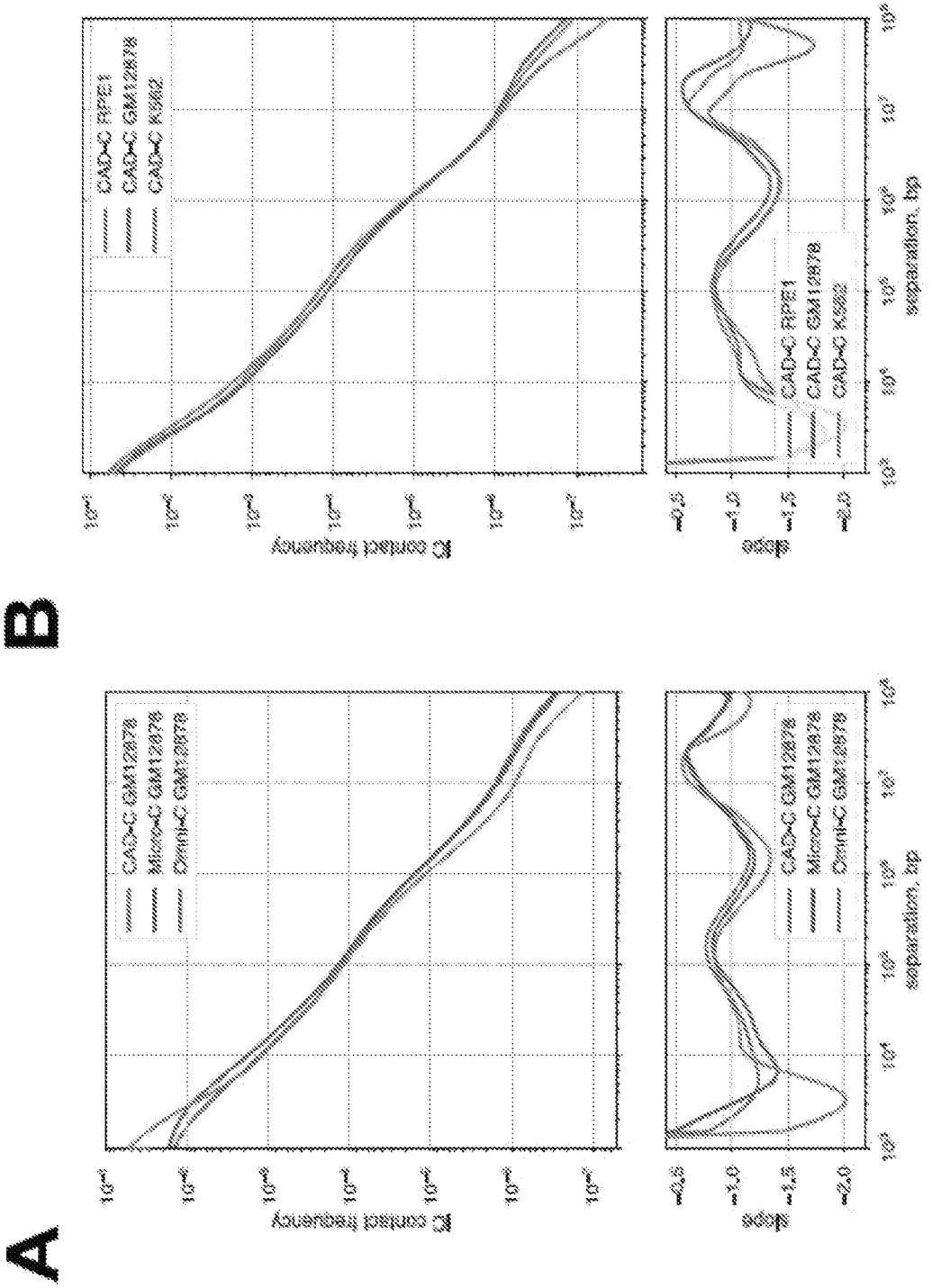

A
B
CAD-C RPE1
CAD-C GM12878
CAD-C K562
Micro-C GM12878
Omni-C GM12878
IC contact frequency
slope
separation, bp

FIGURE 18A

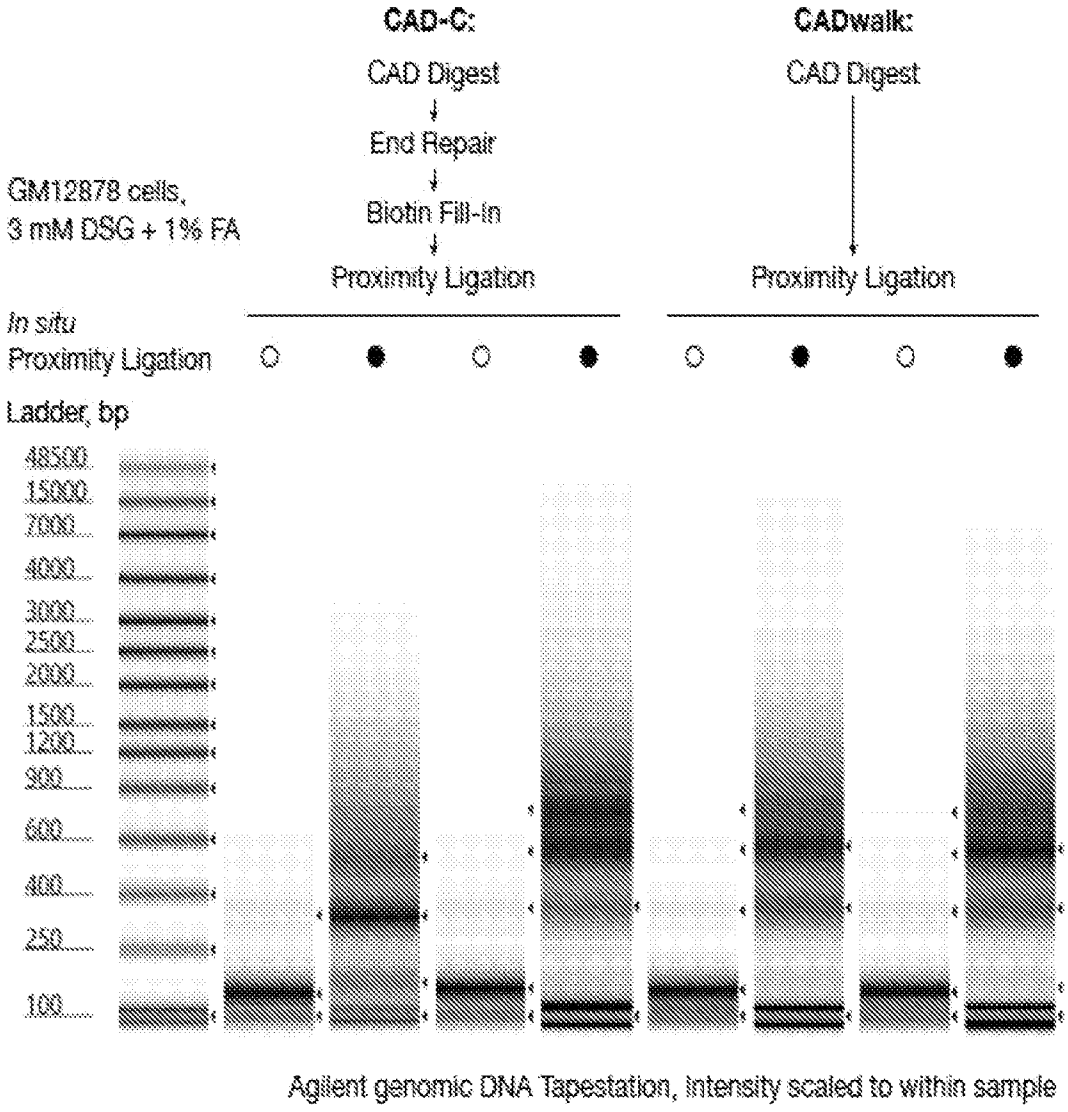

CAD-C:
CAD Digest
End Repair
Biotin Fill-in
Proximity Ligation
CADwalk:
CAD Digest
Proximity Ligation
GM12878 cells,
3 mM DSG + 1% FA
In situ
Proximity Ligation
Ladder, bp
48500
15000
7000
4000
3000
2500
2000
1500
1200
900
600
400
250
100
Agilent genomic DNA Tapestation, intensity scaled to within sample

FIGURE 18 B

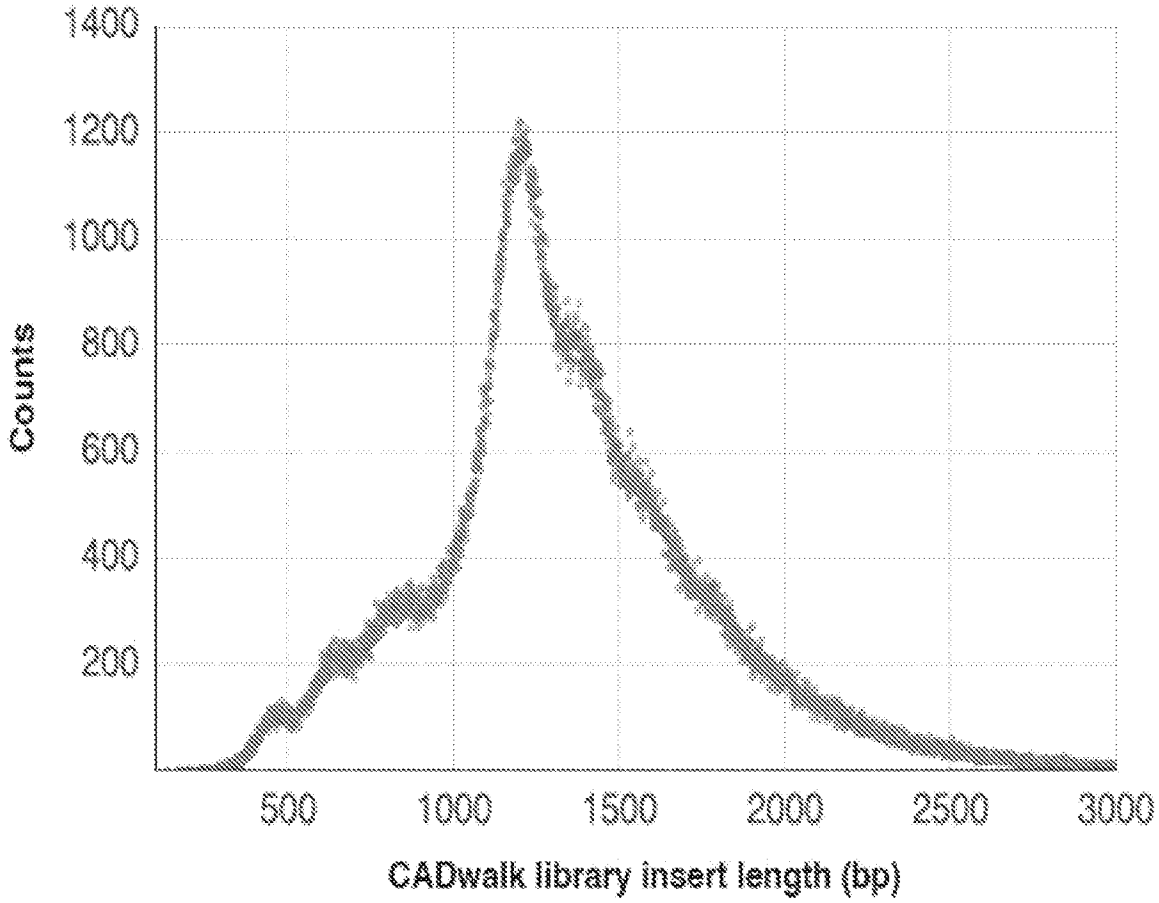

1400
1200
1000
800
600
400
200
Counts
500
1000
1500
2000
2500
3000
CADwalk library insert length (bp)

FIGURE 19

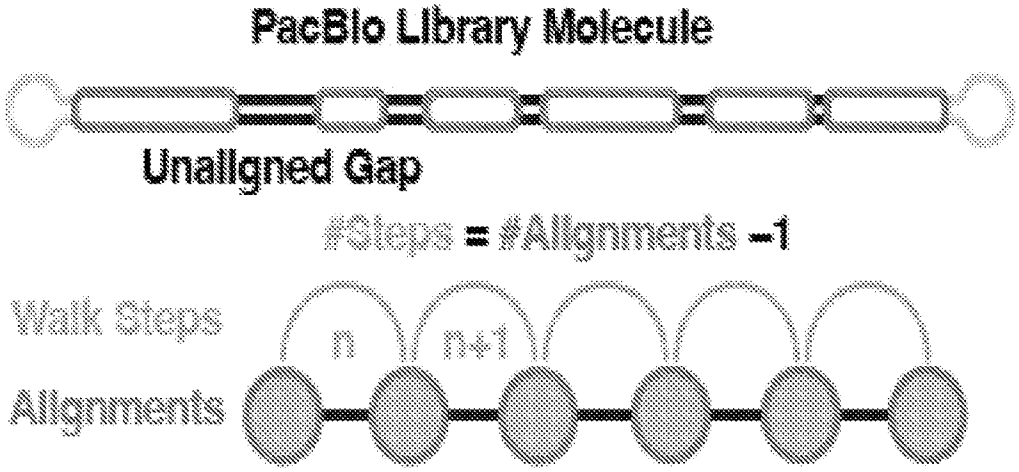

PacBio Library Molecule
Unaligned Gap
#Steps = #Alignments –1
Walk Steps
Alignments
n
n+1

FIGURE 20

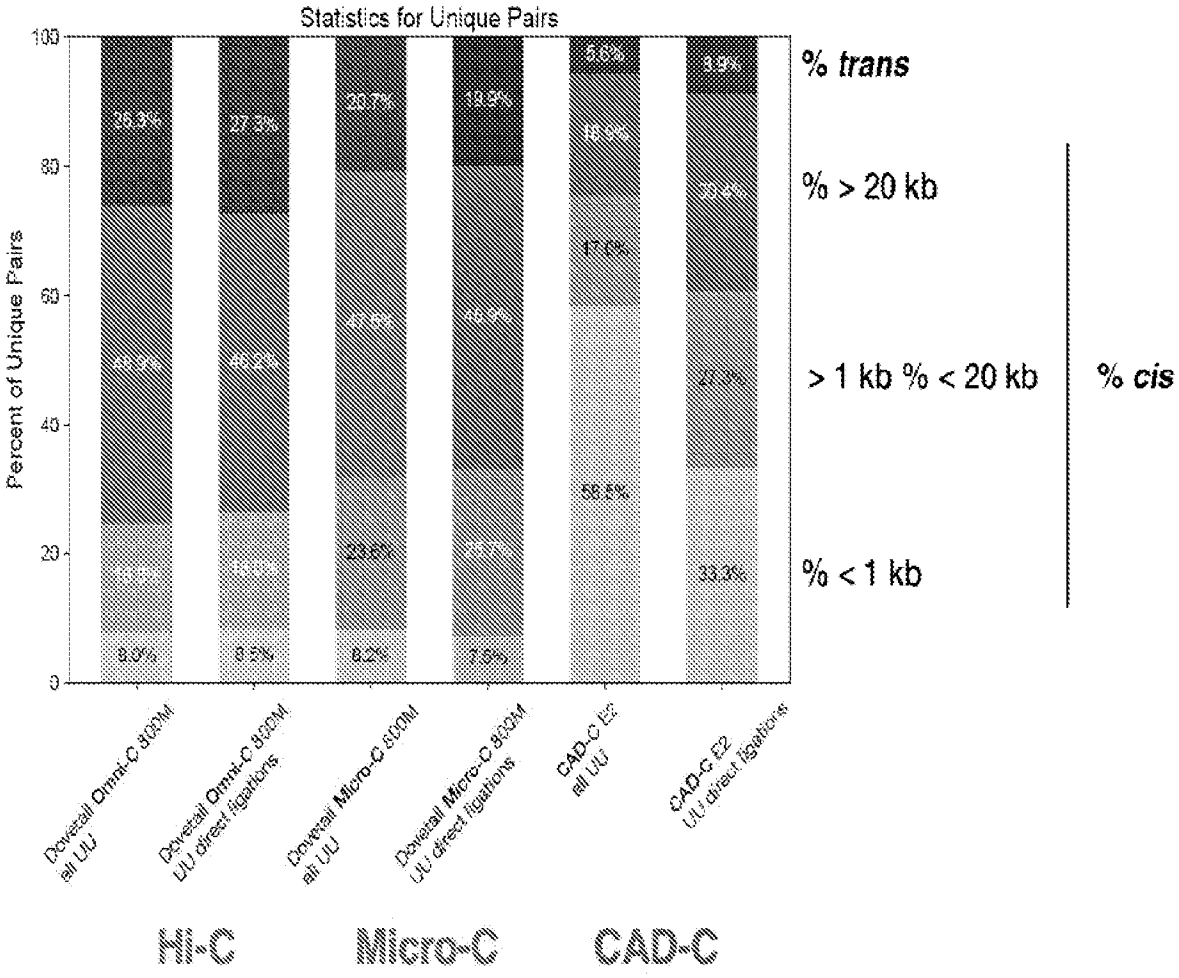

Statistics for Unique Pairs
Percent of Unique Pairs
100
80
60
40
20
0
35.3%
27.2%
20.7%
14.2%
5.6%
3.9%
49.4%
46.2%
47.6%
48.2%
17.6%
52.8%
15.2%
18.2%
23.6%
23.2%
58.5%
33.3%
8.0%
8.5%
8.2%
7.8%
Dovetail Omni-C 300M all UU
Dovetail Omni-C 300M UU direct ligations
Dovetail Micro-C 600M all UU
Dovetail Micro-C 600M UU direct ligations
CAD-C E2 all UU
CAD-C E2 UU direct ligations
% trans
% > 20 kb
> 1 kb % < 20 kb
% < 1 kb
% cis
Hi-C
Micro-C
CAD-C

FIGURE 21

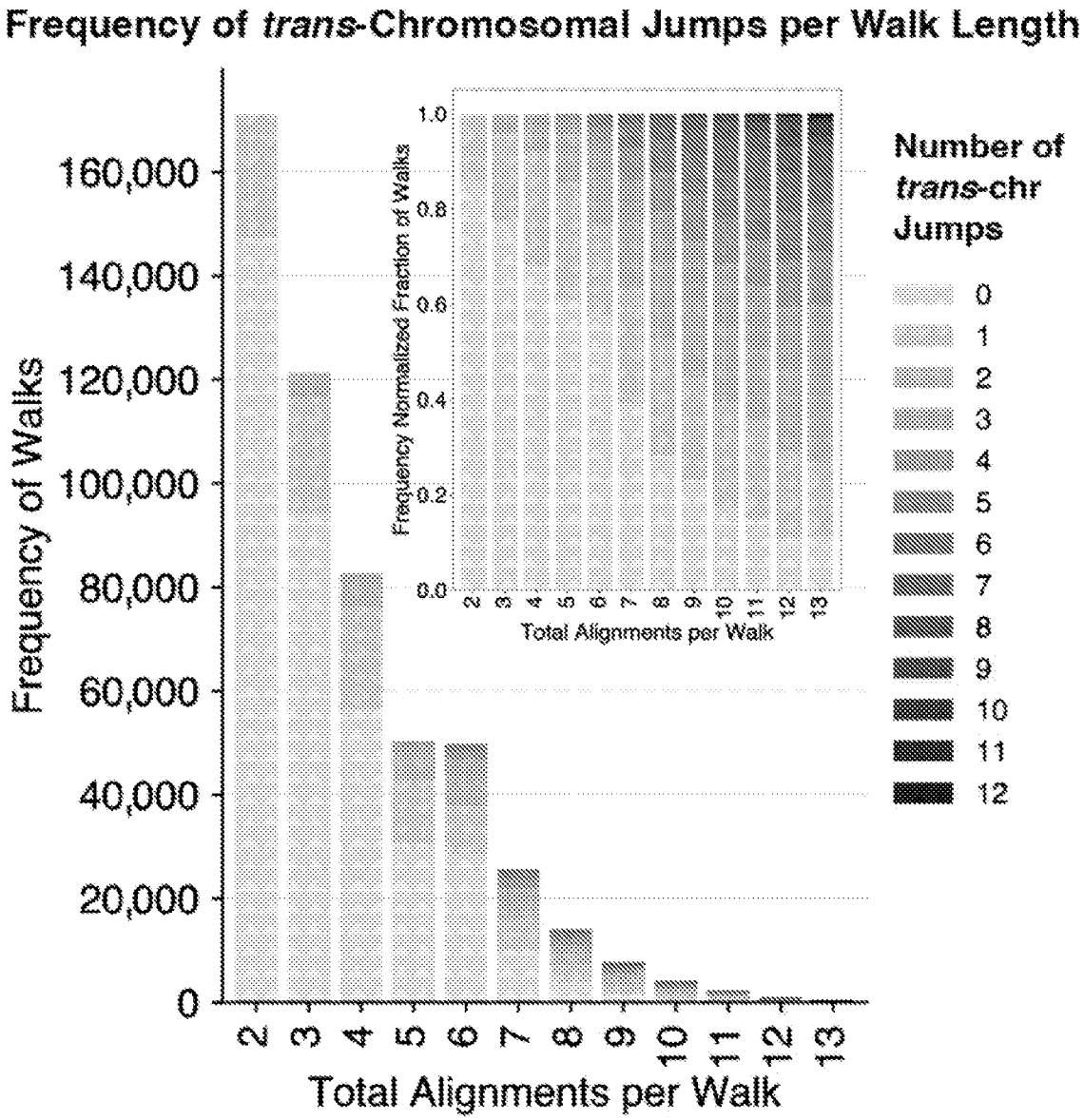

Frequency of *trans*-Chromosomal Jumps per Walk Length
Frequency of Walks
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
2 3 4 5 6 7 8 9 10 11 12 13
Total Alignments per Walk
Frequency Normalized Fraction of Walks
1.0
0.8
0.6
0.4
0.2
0.0
2 3 4 5 6 7 8 9 10 11 12 13
Total Alignments per Walk
Number of *trans*-chr Jumps
0
1
2
3
4
5
6
7
8
9
10
11
12

FIGURE 22

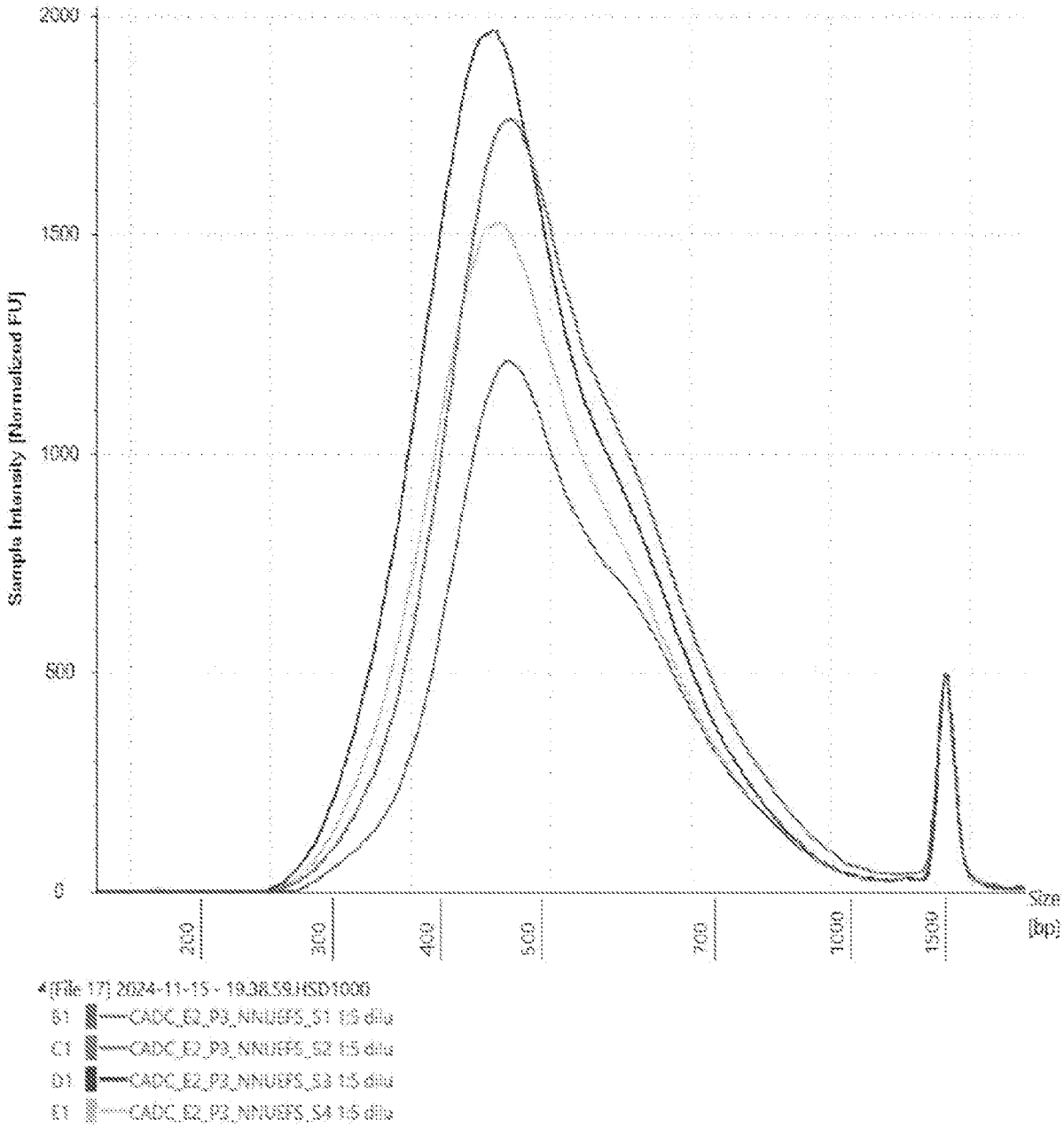

2000
1500
1000
500
0
Sample Intensity [Normalized FU]
200
300
400
500
700
1000
1500
Size [bp]
[File 17] 2024-11-15 - 19.38.59.HSD1000
B1 CADC_E2_P3_NNUSF5_S1 1:5 dilu
C1 CADC_E2_P3_NNUSF5_S2 1:5 dilu
D1 CADC_E2_P3_NNUSF5_S3 1:5 dilu
E1 CADC_E2_P3_NNUSF5_S4 1:5 dilu

FIGURE 23

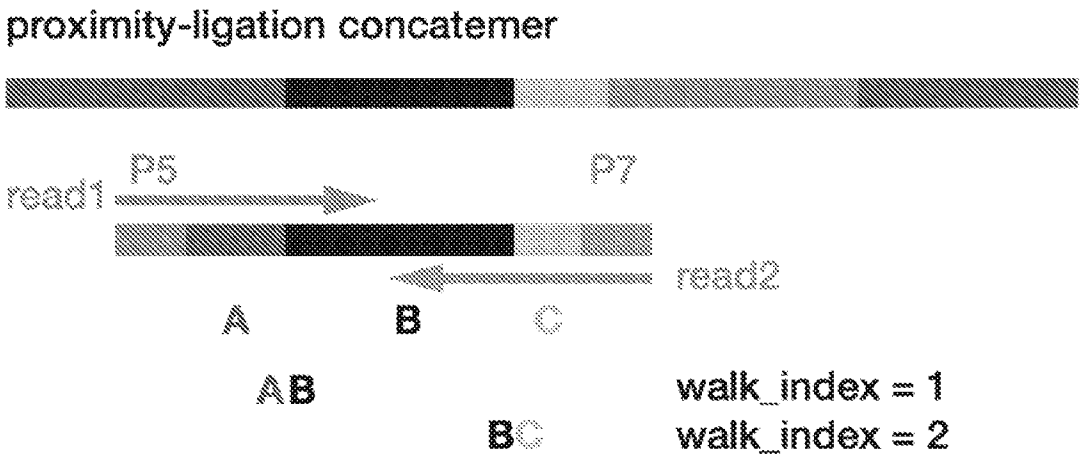

proximity-ligation concatemer
read1
P5
P7
read2
A
B
C
AB
walk_index = 1
BC
walk_index = 2

FIGURE 24

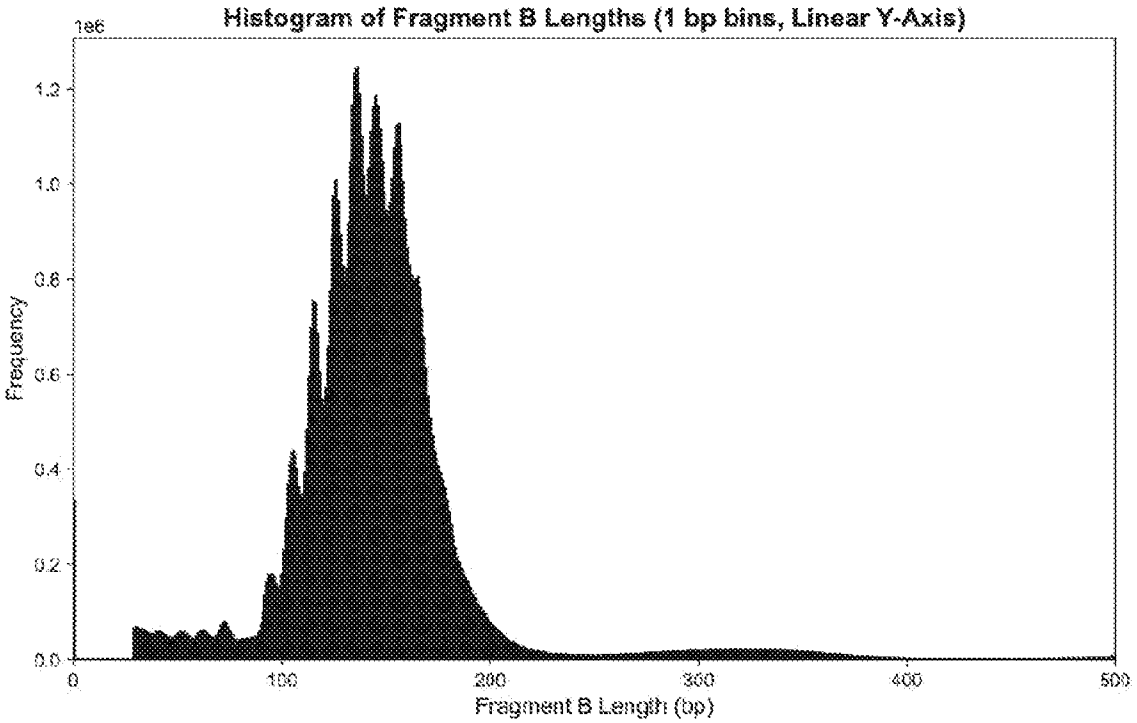
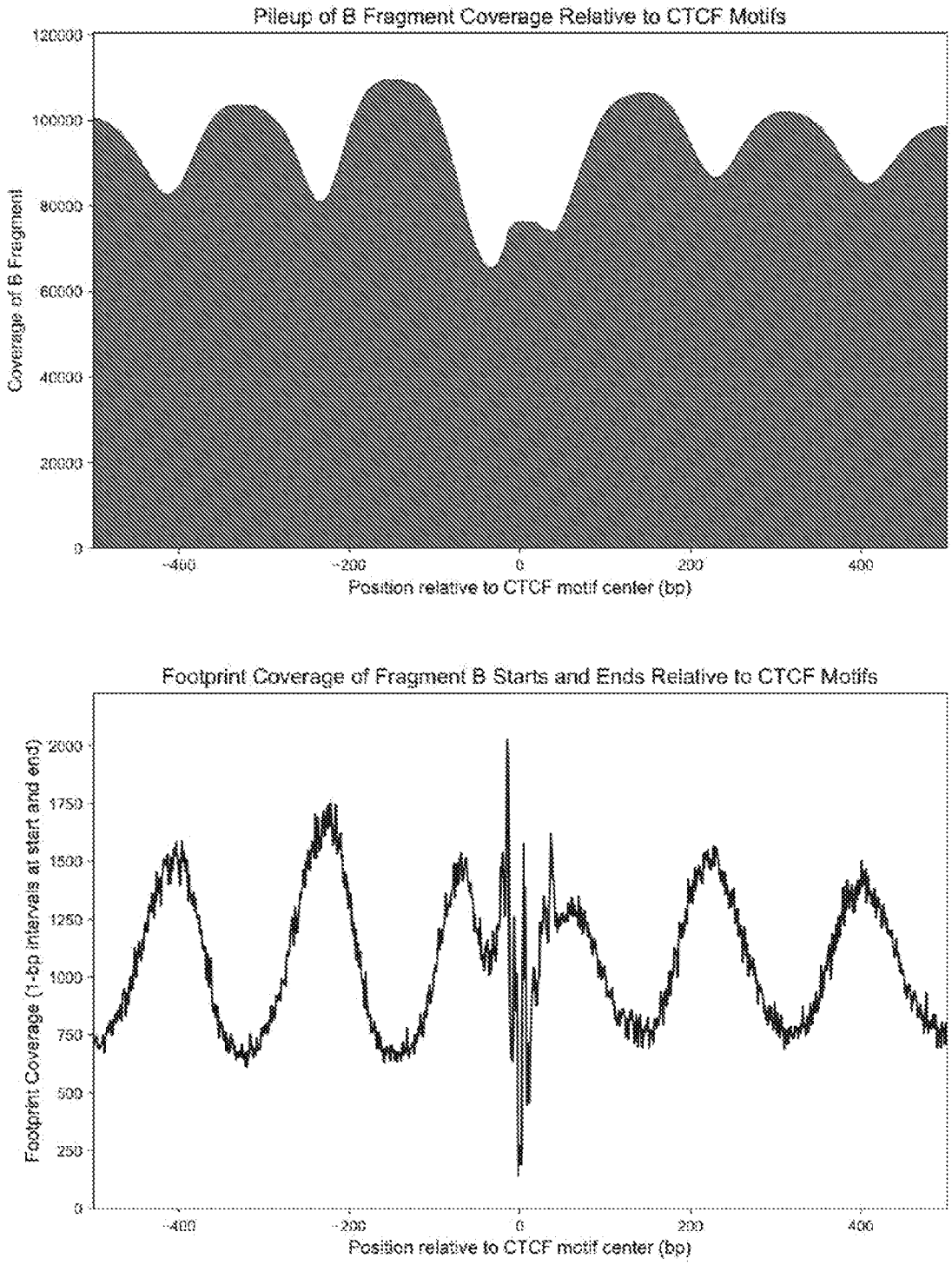

Histogram of Fragment B Lengths (1 bp bins, Linear Y-Axis)
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Frequency
0
100
200
300
400
500
Fragment B Length (bp)

FIGURE 25


Pileup of B Fragment Coverage Relative to CTCF Motifs
Coverage of B Fragment
120000
100000
80000
60000
40000
20000
0
-400
-200
0
200
400
Position relative to CTCF motif center (bp)
Footprint Coverage of Fragment B Starts and Ends Relative to CTCF Motifs
Footprint Coverage (1-bp intervals at start and end)
2000
1750
1500
1250
1000
750
500
250
0
-400
-200
0
200
400
Position relative to CTCF motif center (bp)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2024/059554

A. CLASSIFICATION OF SUBJECT MATTER

IPC: ***C12N 9/22*** (2025.01); ***C12Q 1/6806*** (2025.01); ***C12N 15/10*** (2025.01); ***C12N 15/55*** (2025.01); ***C12N 15/63*** (2025.01); *C12Q 1/6874* (2025.01); *C40B 40/06* (2025.01)

CPC: ***C12N 9/22; C12Q 1/6806; C12N 15/1093; C12N 15/52; C12N 15/63;*** *C40B 40/06; C12Q 1/6874*

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2022/0205017 A1 (ARIMA GENOMICS INC.) 30 June 2022 (30.06.2022) entire document	21, 22, 30-32
A	US 2023/0265499 A1 (FACTORIAL DIAGNOSTICS INC.) 24 August 2023 (24.08.2023) entire document	1-6, 21, 22, 30-32, 34
X	XIAO et al. "Engineered apoptotic nucleases for chromatin research", Nucleic Acids Research; 10 July, 2007; Vol. 35(13) e93, Pgs. 1-7, [retrieved on 1.24.2025]. Retrieved from the internet: <URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC1935020/pdf/gkm486.pdf> entire document	1-6, 34
Y	entire document	21, 22, 30-32
A	MORGAN et al. "Engineered cellular gene-replacement platform for selective and inducible proteolytic profiling", Proc Natl Acad Sci USA; 7 July, 2015, Vol. 112(27), Pgs. 8344-8349, [retrieved on 1.24.2025]. Retrieved from the internet: <URL: https://www.pnas.org/doi/pdf/10.1073/pnas.1504141112 entire document	1-6, 21, 22, 30-32, 34

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 January 2025 (28.01.2025)	**06 February 2025 (06.02.2025)**

Name and mailing address of the ISA/US	Authorized officer
COMMISSIONER FOR PATENTS **MAIL STOP PCT, ATTN: ISA/US** **P.O. Box 1450** **Alexandria, VA 22313-1450** **UNITED STATES OF AMERICA**	**TAINA MATOS**
Facsimile No. **571-273-8300**	Telephone No. **571-272-4300**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2024/059554

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
   a. ☑ forming part of the international application as filed.
   b. ☐ furnished subsequent to the international filing date for the purposes of international search (Rule 13*ter*.1(a)),
      ☐ accompanied by a statement to the effect that the sequence listing does not go beyond the disclosure in the international application as filed.

2. ☐ With regard to any nucleotide and/or amino acid sequence disclosed in the international application, this report has been established to the extent that a meaningful search could be carried out without a WIPO Standard ST.26 compliant sequence listing.

3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2024/059554

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☑ Claims Nos.: **7-20, 23-29, 33, 35-40**
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).