From the Department of Medical Biochemistry and Biophysics
Karolinska Institutet, Stockholm, Sweden

NOVEL METHODS TO STUDY GENOMIC FRAGILITY AND STRUCTURAL VARIATION

Reza Mirzazadeh

Karolinska Institutet

Stockholm 2019

Published by Karolinska Institutet.
Printed by Arkitektkopia AB, 2019

ISBN 978-91-7831-615-1

Novel methods to study genomic fragility and structural variation

THESIS FOR DOCTORAL DEGREE (Ph.D.)

Defended at Karolinska Institutet, Biomedicum seminar room 1, Solnavägen 9, Stockholm.

December 6th 2019, at 10:00 a.m.

By

Reza Mirzazadeh

Principal Supervisor:

Assistant Professor Nicola Crosetto
Karolinska Institutet
Department of Medical Biochemistry and Biophysics

Co-supervisor(s):

Associate Professor Theodoros Foukakis
Karolinska Institutet
Department of Oncology-Pathology

Professor Qiang Pan-Hammarström
Karolinska Institutet
Department of Biosciences and Nutrition

Opponent:

Assistant Professor Vicente Pelechano
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Examination Board:

Professor Ulf Landegren
Uppsala University
Department of Immunology, Genetics and Pathology

Professor Mattias Mannervik
Stockholm University
Department of Molecular Biosciences, The Wenner-Gren Institute

Professor Richard Rosenquist Brandell
Karolinska Institutet
Department of Molecular Medicine and Surgery

Dedicated to family, friends, colleagues, my wonderful wife, Nana,
and my amazing daughter Elsa, for their support and love.

ABSTRACT

DNA double-strand breaks (DSBs) are major DNA lesions that when repaired unfaithfully can give rise to loss of genetic information, chromosomal rearrangements such as insertions/deletions (indels) and copy number alterations (CNAs), which in turn lead to genomic instability that is characteristic of almost all cancer types. In this context, it is thought that genomic instability has critical roles in cancer initiation, progression and intra-tumor heterogeneity (ITH). DSBs have also been exploited for genome-editing purposes where using different CRISPR (clustered regularly interspaced short palindromic repeats) systems, one can create DSBs in the target DNA to alter sequences and modify gene function. However, this approach is not without drawbacks, as DSBs can be created at sites other than the intended target (known as off-target effects), which can potentially be mutagenic.

Therefore, given the importance of DSBs in genomic instability, their role in generation of CNAs and genome-editing technologies, it is of great interest to determine genomic locations of DSBs and their frequency along the genome, together with DNA copy number profiling. Thus, the focus of this thesis was to develop molecular tools for detection and quantification of DSBs with single-nucleotide resolution in different model systems, in combination with the development of technologies for DNA copy number profiling, by which we can collectively understand the biology behind DSBs, their links to CNAs in the context of cancer and assess the safety profile of CRISPR systems for therapeutic applications.

In **Paper I**, we developed BLISS (Breaks Labeling *In Situ* and Sequencing) as a quantitative method enabling genome-wide DSB profiling. We showed that BLISS accurately identified both endogenous and drug-induced DSBs genome-wide, even in samples of a few thousand cells and in single tissue sections. Additionally, we demonstrated that BLISS is a powerful tool to measure the off-target activities of Cas9 and Cpf1 CRISPR systems, and indeed we found that Cpf1 was more specific than Cas9.

In **Paper II**, using BLISS-generated DSB data from cell lines, we modeled the contribution of genetic and epigenetic features in shaping the cancer fragility, and made predictions of the frequency of expected breaks across the human genome. We constructed random forest regression models from four DSB datasets and found that the most influential feature in DSB frequency prediction is replication timing across all models. In addition, we noticed that open chromatin at transcriptionally active genes and associated regulatory factors have the largest influence on the frequency of DSBs than transcription per se.

In **Paper III**, we developed CUTseq, which builds on the design of BLISS from **Paper I**, and can be used for gDNA barcoding and amplification to generate multiplexed DNA sequencing libraries for performing reduced representation sequencing of DNA samples extracted from cell lines, FFPE tissue sections or small sub-regions thereof. We demonstrated the applicability of CUTseq for CNA profiling, and showed that CUTseq can reproducibly detect a considerable fraction of high-confidence single nucleotide variants (SNVs) that were also detected by a standard exome capture method. Finally, we demonstrated that CUTseq can be applied for multi-region tumor sequencing to assess ITH of CNA profiles of multiple-small regions of a single FFPE tissue sections of primary and metastatic breast cancer lesions.

LIST OF SCIENTIFIC PAPERS

THESIS PUBLICATIONS

I. Yan, W.X.*, **Mirzazadeh, R.***, Garnerone, S., Scott, D., Schneider, M.W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L., Federova, Y., Zetsche, B., Zhang, F., Bienko, M., Crosetto, N. (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun*. 8, 15058.

II. Ballinger, T. J., Bouwman, B. A. M., **Mirzazadeh, R.**, Garnerone, S., Crosetto, N., & Semple, C. A. (2019). **Modeling double strand break susceptibility to interrogate structural variation in cancer**. *Genome Biology*, 20 (1), 28.

III. Zhang, X., Garnerone, S., Simonetti, M., Harbers, L., Nicoś, M., **Mirzazadeh**, **R**., Venesio, T., Marchiò, C., Sapino, A., Hartman, J., Bienko, M., Crosetto, N. (2019). **CUTseq is a versatile method for preparing multiplexed DNA sequencing libraries from low-input samples**. *Nat. Commun*. 10, 1038.

OTHER PUBLICATIONS

Gelali, E., Girelli, G., Matsumoto, M., Wernersson, E., Custodio, J., Mota, A., Schweitzer, M., Ferenc, K., Li, X., **Mirzazadeh, R.**, Agostini, F., Schell, J.P., Lanner, F., Crosetto, N., Bienko, M. (2019). **iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture**. *Nat. Commun*. 10, 1636.

Wu, C., Simonetti, M., Rossell, C., Mignardi, M., **Mirzazadeh, R.**, Annaratone, L., Marchiò, C., Sapino, A., Bienko, M., Crosetto, N., Nilsson, M. (2018). **RollFISH achieves robust quantification of single-molecule RNA biomarkers in paraffin-embedded tumor tissue samples**. *Commun. Biol*. 1, 209.

Mirzazadeh R., Kallas T., Bienko M., Crosetto N. (2018). **Genome-Wide Profiling of DNA Double-Strand Breaks by the BLESS and BLISS Methods**. In: Muzi-Falconi M., Brown G. (eds) Genome Instability. *Methods Mol Biol*, vol 1672. Humana Press, New York, NY.

* These authors contributed equally to the work.

CONTENTS

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
gDNA	Genomic DNA
DSBs	DNA double-strand breaks
RNA	Ribonucleic acid
CNAs	Copy number alterations
Indels	Insertions/deletions
ITH	Intratumor heterogeneity
CRISPR	Clustered regularly interspaced short palindromic repeats
BLISS	Breaks labeling in situ and sequencing
FFPE	Formalin-fixed paraffin-embedded
SSBs	Single strand breaks
ROS	Reactive oxygen species
CSR	Class-switch recombination
AID	Activation-induced cytidine deaminase
RAG1/2	Recombination activating genes
TOPII	Type II topoisomerases
CFSs	Common fragile sites
ERFs	Early replication fragile sites
crRNA	CRISPR-RNA
tracrRNA	Trans-activating CRISPR RNA
sgRNA	Single-guide RNA
PAM	Protospacer adjacent motif
Cas9	CRISPR associated protein 9
Cpf1	CRISPR from Prevotella and Francisella 1
DDR	DNA damage response
ATM	Ataxia telangiectasia mutated
ATR	Ataxia telangiectasia and Rad3-related protein

γH2A.X	Histone H2A, variant X, phosphorylated at Ser139
NHEJ	Non-homologous end joining
HR	Homologous recombination
DNA-PKcs	DNA-dependent protein kinase catalytic subunit
DSBR	Double-strand break repair
SDSA	Synthesis-dependent strand annealing
LCRs	Low-copy repeats
NGS	Next generation sequencing
ChIPseq	Chromatin Immunoprecipitation Sequencing
dCas9	Catalytic-deactivated Cas9
BLESS	Breaks labeling enriched on streptavidin and sequencing
PCR	Polymerase chain reaction
DSBCapture	Double-strand break capture
End-seq	DSB ends sequencing
ZFN	Zinc-finger-nuclease
qDSB-Seq	Quantitative DSB sequencing
Digenome-seq	Cas9-digested whole genome sequencing
WGS	Whole-genome sequencing
IDLV	Integrase defective lentiviral vector
TALEN	Transcription activator-like effector nuclease
dsODN	double-stranded oligodeoxynucleotide
GUIDEseq	Genome-wide unbiased identification of DSBs enabled by sequencing
LAM-HTGTS	Linear amplification–mediated high-throughput genome-wide translocation sequencing
UMI	Unique molecular identifier
IVT	*in vitro* transcription

TSS	Transcription start site
GO	Gene ontology
spCas9	Streptococcus pyogenes Cas9
TCGA	The Cancer Genome Atlas
ICGC	International Cancer Genome Consortium
ENCODE	Encyclopedia of DNA Elements
COSMIC	Catalogue of somatic mutations in cancer
COAD	Colon adenocarcinomas
MELA	Melanomas
BRCA	Breast cancer
GIST	Gastrointestinal stromal tumors
RNA POL II	RNA polymerase II
MLL	Mixed lineage leukemia
53bp1	Tumor protein p53 binding protein 1
TUNEL	Terminal deoxynucleotidyl transferase dUTP nick end labeling
RADseq	Restriction site Associated DNA Sequencing
scWGS	Single-cell whole-genome sequencing
LCM	Laser-capture microdissection

1 INTRODUCTION

1.1 Role of DNA double-strand breaks (DSBs) in genomic instability

Constant maintenance of the genome sequence information and structural integrity is crucial for faithful transmission of genetic material to daughter cells[1]. However, aberrations are inevitable as the DNA molecule is subject to a multitude of different types of damage on a daily basis[2]. Examples include DNA mismatches, oxidative and hydrolytic cleavage, DNA-protein cross-links, and various forms of breaks such as single-strand breaks (SSBs) and double-strand breaks (DSBs)[1]. The latter constitute the most significant threat to the cell, as if not repaired or miss-repaired they can result in loss of genetic information, cell death or structural and quantitative chromosomal rearrangements, collectively known as structural variations (SVs)[3]. SVs are rearrangements of large DNA segments including insertions/deletions (indels), duplications, inversions and translocations[3], which in turn can lead to genomic instability[4,5]. Genomic instability, which refers to a state of increased mutations and copy number changes, and is characteristic of almost all cancer types, is thought to play a critical role in cancer initiation, progression and also as a driver of intra-tumor heterogeneity (ITH)[6]. Regardless of their causes, elevated levels of DSBs contribute to genomic instability and therefore, properly repairing DSB lesions is indispensable for maintaining the stability and integrity of the genome. In the following sections, I will explain the most common sources of DSB formation, repair systems and links to the generation of copy number variations (CNVs), and lastly the technologies that can be applied to detect DSBs.

1.2 Sources of DSBs

DSBs have been classically associated with exposure to exogenous toxic agents, such as ultraviolet light, X-ray, ionizing radiation and certain chemotherapeutic drugs, such as etoposide[7]. However, the majority of DSBs are thought to form during fundamental physiological processes[7].

Reactive oxygen species (ROS)

ROS, which are generated as by-products of cellular metabolism can oxidize nucleobases of DNA, thus resulting in both SSBs and DSBs[7,8]. In addition, oncogenic activation can lead to increased levels of ROS, which in turn oxidize dNTPs and result in the occurrence of replication stress and the formation of DSBs[9]. Furthermore, accumulation of ROS have been reported to induce lesions and strand breaks on mitochondrial DNA, resulting in mitochondrial DNA degradation[10,11], which in turn causes mitochondrial dysfunction[12].

Antibody diversification and meiotic recombination

In the immune system, during the development of B cells to create antibody diversification, class-switch recombination (CSR) and V(D)J rearrangements are associated with transient induction of DSBs by the activation-induced cytidine deaminase (AID) and recombination activating genes (*RAG1/RAG2*), respectively[13,14]. AID converts cytidine into uridine, and the resulting mismatch is recognized by the cellular repair system that converts it to DSBs and re-joins them[15]. The RAG1/2 complex with its endonuclease capacity can also induce DSBs that are promptly recognized by the repair machinery[16]. In addition, during meiosis, initiation of homologous recombination and proper segregation of sister chromatids is mediated by SPO11 in complex with other proteins to generate transient DSBs, which are subsequently detected by the cellular repair systems[17,18].

DNA replication

In proliferating cells, replication stress (i.e., slow down or stalling of replication forks) is thought to be the main source of DSB formation[19]. Replication stress can happen due to different reasons including: *i*) lack of factors essential for DNA synthesis (e.g. dNTP, DNA polymerases, replication origin firing)[20]; *ii*) decoupling of DNA polymerase-helicase activity; *iii*) obstacles impeding fork progression (e.g. abasic sites, inter/intra strand crosslinks, non-canonical secondary structures like hairpins/quadruplexes and hard to replicate regions such as repetitive telomeric/centromeric sequences)[21,22] and *iv*) collisions with the transcription machinery[20,23].

Transcription process

A growing body of evidence supports the idea that physiologic transcription is also a major source of endogenous DSBs and a potential cause of DSB-related mutations[24]. Transcription can lead to DSBs in several ways. For example, collision of transcription and replication machineries in a head-on or co-directional fashion, can lead to stalling/displacement of RNA polymerase and of the replication fork[25]. In addition, head-on collisions induce the formation of R-loops – special three-stranded nucleic acid structure, composed of DNA-RNA hybrid and the associated non-template single-stranded DNA[26] – leading to fork stalling and DNA breaks[27]. Moreover, non-template single-stranded DNA exposed in R-loops is more vulnerable to break by the formation of noncanonical structures such as hairpins and G-quadruplexes, which increase the chance of replication stress and genomic instability[28]. Transcription activation itself is associated with the formation of transient DSBs inside gene regulatory regions, such as promoters and enhancers, mostly as a result of type II DNA topoisomerases (TOPII), that generate transient DSBs to help resolve torsional stress linked with transcription fork movements and enhancer promoter interactions[29,30].

Intrinsic fragile genomic regions

Certain genomic regions are naturally prone to break. DSBs can occur at higher frequency in regions known as common fragile sites (CFSs) and early replication fragile sites (ERFSs). These two types of fragile hotspots represent distinct classes of fragility, each with features that may contribute to their instability. CFSs occur primarily in late-replicating genomic loci that contain large genes with AT-rich sequences and poor density of replication origins. CFSs largely show cell-type specificity and often overlap with the boundaries of cancer-associated CNAs, particularly large deletions[20]. In contrast, ERFSs occur in early-replicating DNA regions, rich in GC sequences, replication origins and actively transcribed genes[31]. ERFs also overlap with recurrent chromosomal rearrangements and CNAs, especially with the large duplications and translocations found across many cancer types[31].

Programmable DSBs in genome editing

DSBs can be experimentally exploited in such a way that specific sites of the DNA can be targeted for genome editing purposes. Clustered regularly interspaced short palindromic repeats (CRISPR) is a hallmark of the adaptive immune system in many bacteria and most archaea, by which CRISPR-derived RNA in combination with different Cas (CRISPR associated) endonuclease proteins – that act like molecular scissors – induce DSB in the nucleic acids of the invading viral or plasmid DNA, leading to its dysfunctionality[32]. CRISPR is a specialized region of DNA containing short repetitive base sequences separated by stretches of interspersed variable sequences known as spacers. These short spacer sequences are acquired from previous exposure to foreign genetic material that was incorporated into the CRISPR region, which then serve as a memory to enable bacteria to recognize the viruses and defend future attacks[32]. This system can be repurposed when CRISPR components are transferred into other, more complex organisms, such as mammalian cells, and as such the DNA sequence of interest can be targeted and cut for gene manipulation/editing[33].

Since type II CRISPR systems use single-component effector proteins, such as Cas9, the system has been successfully harnessed for genome editing in eukaryotic cells by introducing the three essential components of the system together: Cas9, crRNA (CRISPR RNA), and tracrRNA (Trans-activating CRISPR RNA)[34]. Alternatively, crRNA and tracrRNA have been fused to create a chimeric single-guide RNA (sgRNA), enhancing the simplicity and the possibility to multiplex and edit several regions at the same time[35]. Typically, the first 20 nucleotides of the sgRNA are complementary to the target DNA, followed by the PAM (Protospacer Adjacent Motif, which is used as a recognition handle for Cas proteins to bind to for the cleavage process) sequence, which in the case of the Cas9 system is normally an NGG-rich sequence at the 3' end of protospacer sequence (where N is any nucleotide). Although the development of the CRISPR/Cas9 system for genome editing

was revolutionary, especially in mammalian cells, off-target cleavage events of the system have become a major concern over recent years[36]. Hence, there have been tremendous efforts made into seeking ways to improve cleavage specificity, including sgRNA modifications, engineered versions of Cas9, searching for enzymes with higher specificity, and several other approaches[37]. Recently, another CRISPR-Cas system has been identified and named Cas12a/Cpf1 (CRISPR from Prevotella and Francisella bacterial species). Cas12a has several new features when compared to Cas9[38], which broadens its applications for genome engineering. In this system, only a short single crRNA is required to guide the cleavage process, which makes the system simpler than Cas9. In addition, Cas12a uses a T-rich PAM at the 5' end of the protospacer sequence as a target DNA recognition sequence in order to generate DSBs with 5' overhangs of 4-5 nucleotides at the distal site of the PAM[38]. Cas12a has also been successfully harnessed for human genome editing, but as discussed earlier despite the advantages of CRISPR toolboxes, off-target mutagenesis is still a major concern[36].

1.3 DNA damage response (DDR)

DSBs must be properly repaired regardless of their sources, because of the lethal consequences that arise from their formation, which result in genomic instability and the potential onset of diseases such as cancer. To safeguard genome stability, an efficient network of tightly regulated cellular pathways to sense, signal, and ultimately repair the lesions, referred to as DNA damage response (DDR), exists in different organisms[1]. DNA damage is sensed by proteins Mre11, Rad50 and Nbs1 (MRN complex), which lead to activation of protein kinases that initiate cascades of signaling pathways involved in the DDR. The central regulators that orchestrate signaling in the mammalian DDR are ataxia-telangiectasia-mutated (ATM) and ataxia telangiectasia and Rad3-related (ATR) kinases[39]. ATM is primarily activated upon sensing of DSBs, whereas ATR responds predominantly to SSBs, and to a lesser extent to DSBs and other types of damage, yet there is cross-talks between ATM and ATR[39]. Upon DSB sensing, ATM and ATR become activated and phosphorylate Ser/Thr-Glu motifs in hundreds of proteins, including Ser139 of H2A.X (γH2A.X) histone variant[40]. In turn, γH2A.X recruits repair proteins to the DSB[41], leading to the activation of pathways that eventually slow down or arrest cell cycle progression, providing enough time for repair proteins to resolve the DNA damage or trigger apoptosis if the damage is irreparable[42]. The DDR network is crucial for maintaining genome integrity, hence any mutation in a component of the system poses a major threat to the cell that may lead to the onset of pathological disorders. For example, germline mutations in the ATM gene are linked to the development of an autosomal recessive disorder known as ataxia-telangiectasia

(A-T), a progressive neurodegenerative disorder associated with cerebellar ataxia, immunodeficiency, cell-cycle checkpoint defects, genome instability and cancer predisposition[43]. Mutations in the ATR gene cause Seckel syndrome, which is characterized by intrauterine-growth retardation, dwarfism, microcephaly, and skeletal and brain abnormalities[44].

1.4 DSBs repair pathways

To avoid catastrophic consequences of DSBs, two major repair mechanisms are in place: non-homologous end joining (NHEJ) and homologous recombination (HR), which are separately summarized in the following paragraphs.

Non-homologous end joining

In NHEJ, the two ends of a DSB are quickly re-joined[45]. NHEJ is the major mechanism for DSB repair in human cells[46] and is believed to be the key repair pathway involved in V(D)J and CSR recombination during B and T lymphocyte development[47]. NHEJ is active in several phases of cell cycle, but it is the dominant DSB repair mechanism in G0 and G1. When NHEJ is initiated, the two DSB ends are recognized and bound by a heterodimeric Ku protein complex, containing Ku70 (XRCC6) and Ku80 (XRCC5) polypeptides[48]. Ku is known to have extraordinary affinity for DSB ends in a sequence-independent manner, and acts as a scaffold to which other NHEJ proteins are recruited[48]. The Ku complex recruits the catalytic subunit of DNA-dependent protein kinase (DNA-PKcs) to the two DSB ends, resulting in activation of DNA-PK at the DSB ends[49,50]. DNA-PK binding helps Ku to slide down the DNA duplex and prevent exonucleolytic degradation of the DSB ends[51]. DNA-PK phosphorylates Ser139 on histone variant H2A.X (γH2A.X), which recruits the repair machinery to the DSB sites, and, as such, it serves a marker for DSBs detection[52]. DNA-PK also phosphorylates and changes the function of other NHEJ factors, including Artemis, X-ray cross complementing protein 4 (XRCC4), XRCC4-like factor (XLF), Aprataxin-PNK-like factor (APLF), and DNA ligase IV (LigIV), which accumulate at DSBs to rejoin the broken ends[53]. It had been thought that recruitment of NHEJ repair proteins is a sequential stepwise model, whereas recent studies suggest that the order of repair factors can be flexible based on the complexity of the damage[54]. Given the fact that the type of nuclease, polymerase and ligase that are involved in the NHEJ pathway can act on a wide range of DNA ends with different conformations, NHEJ is a relatively error-prone process, which can lead to the accumulation of small insertions, deletions and translocations at the re-joined ends[55].

Homologous recombination (HR)

The second most important DSB repair mechanism that enables more precise repair of DSBs is HR. While NHEJ is active in several phases of the cell cycle, HR is active mainly in S and G2 phases, where a homologous template is available[56]. HR is initiated by the binding of the MRN complex to the DSB ends. The 5´ ends of the DSB are processed to create single-stranded DNA 3´-OH overhangs, during a process known as resection. Recombinases Rad51 or Dmc1 have been shown to be the key players that bind to the created ssDNA overhangs to form the presynaptic filament (i.e. a helical filament of recombinase enzyme on ssDNA in preparation for strand invasion)[57]. The presynaptic filament is needed to guide the ssDNA to the homologous duplex DNA and form the synaptic complex that searches for homologous sequences. Once found, the ssDNA 3´end invades the homologous sequence (known as strand-invasion process) in the duplex, creating a DNA joint called a D-loop, which is then extended through DNA synthesis. After strand synthesis, the D-loop is altered into a cross-like structure known as a Holliday junction. From this step onward, the two primary repair pathways have been described for HR: *i)* the double-strand break repair (DSBR) pathway, also known as double Holliday junction model; and *ii*) the synthesis-dependent strand annealing (SDSA) pathway[58]. In DSBR, the 3´-OH overhang that was not involved in strand invasion, invades the homologous duplex DNA to form another Holliday junction. Finally, the junctions are resolved by nicking nucleases, through which, based on how the Holliday junctions are resolved, the decision as to whether chromosomal crossing-over should occur is determined. In the SDSA pathway, there is no crossing-over of recombinant products, and the newly synthesized 3´ end of the invading strand is released from the Holliday junction, and through base-pair complementarity it anneals to the ssDNA on the other break end that was not involved in strand invasion process[58]. Any gap between the newly annealed strands is filled by ligation to finalize the damage repair. The NHEJ and HR pathways are tightly regulated to ensure that undamaged genetic material is transferred to daughter cells. Therefore, deficiencies and mutations in any components of these repair pathways endanger genome integrity, and potentially lead to the onset of pathological disorders, including cancer[59].

1.5 Adverse outcomes of DSB repair and structural changes

Improper repair of DSBs can give rise to SVs (i.e., genomic alterations as small as 50 base pairs reaching up to megabases in length) that involve balanced rearrangements such as translocations and inversions or imbalanced rearrangements (quantitative changes) such as duplications and deletions, also known as copy number variations (CNVs, in germline cells) or copy number alteration (CNAs, in somatic

cells)[60,61,62]. Genomic SV is a major source of variation between human population, underlying human evolution and many diseases from developmental disorders to cancer[63].

Two major mechanisms have been reported to be involved in the formation of SVs: non-allelic homologous recombination (NAHR) and NHEJ[63]. NAHR is a form of HR and occurs between highly similar non-homologous DNA sequences and is thought to be the major source of SVs formation based on the observation that the boundaries of SVs are often found near repetitive sequences, mainly low copy repeats (LCRs)[64]. LCRs are tracts of duplicated DNA sequences of between 1-300 kb in length that have a sequence similarity of more than 95%, so that their misalignment during mitosis or meiosis generates rearrangements such as deletions, duplications and translocations[65]. It was suggested that LCRs act as a substrate for NAHR, by which most of the recurrent SVs arise[64], that share a common size, clustered breakpoints (i.e., breakpoints that fall in the same genomic region), and recur in different unrelated individuals. In the case of occurrence of DSBs in genomic regions lacking extensive homologous sequences that can be used as repair templates for NAHR, ends of DSB are processed by NHEJ instead, which often leads to insertions and deletions at breakpoints. NHEJ is proposed as the cause of non-recurrent SVs that differ in sizes in each individual and show distinct breakpoints[65,66].

It has been reported that SVs may contribute to tumor development by different mechanisms. For example, they can have direct consequences through alteration of the gene dosage (e.g., duplication of oncogenes or deletion of tumor suppressors genes), and creation of oncogenic fusion proteins by translocations. In addition, SVs can also have indirect outcomes, such as the reorganization of the proximity distance of regulatory elements (e.g., active enhancers) and proto-oncogenes, thus leading to cancer-related gene overexpression[67–70].

As discussed above, DSBs can be generated from normal cellular physiological processes (e.g., replication, transcription), and are involved in the formation of SVs/CNAs that threatens genomic stability, and are also exploited by genome-editing technologies. Therefore, precisely mapping the location and frequency of DSBs along the genome is of great interest for better understanding the biology of DSBs, genome fragility, and specificity of genome editing tools. Hence, several methods have been developed for genome-wide detection of these lesions and used in different applications ranging from studying the dynamics of DSBs in DNA damage, to measuring the off-target activity of genome-editing nucleases. Here, we review the most commonly used DSB detection methods.

1.6 Methods for genome-wide profiling of DSBs

1.6.1 Chromatin immunoprecipitation and high-throughput sequencing (ChIP-seq)

In ChIP-seq[71] , using specific antibodies targeting DNA-associated proteins one can study the interaction of a protein of interest with DNA [**Fig. 1**]. ChIP-seq was adopted to detect recruited repair proteins such as γH2A.X at DSB locations, to study fragile sites in yeast and DSB repair processes in mammalian cells as well[72,73]. It has also been exploited for genome-wide profiling of off-target binding activities of CRISPR Cas systems, by detection of DNA repair factors (e.g., MRE11)[74] upon DSB induction or using catalytic deactivated Cas9 (dCas9), which stably binds to DNA once it has recognized its target site [75]. Although, ChIP-seq proved to be capable of detecting DSBs and identifying different off-targets depending on sgRNAs, however, using antibodies targeting the activated DNA repair proteins makes ChIP-seq an indirect method for DSBs detection. In addition, as normally repair factors are not localized to the immediate vicinity of DSB and instead extended to a large chromatin region (up to 1 Mb) around it, ChIP-seq cannot provide nucleotide resolution. Furthermore, applying ChIP-seq for genome-wide off-target profiling of CRISPR systems, using dCas9 may not be ideal, as its deactivation can affect its binding specificity.

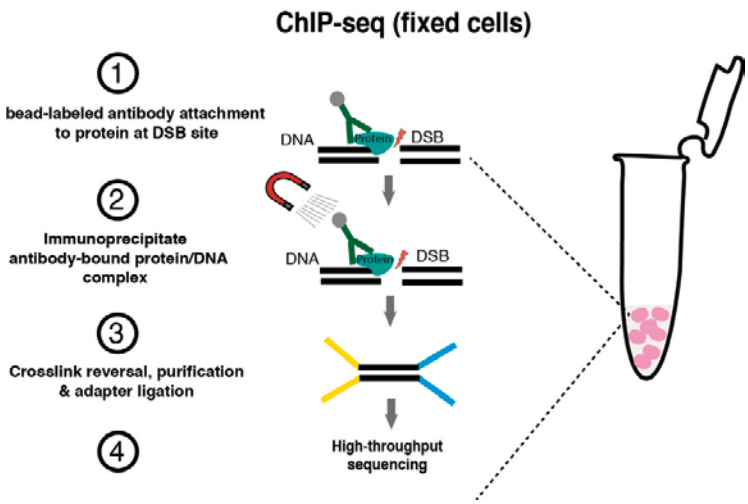

Figure 1. *In ChIP-seq cells are fixed by formaldehyde to cross-link proteins and DNA.* ***1)*** *A bead-labeled antibody targeting protein of interest at DSB site is introduced.* ***2)*** *Using a magnet, the antibody-protein-DNA complex is immunoprecipitated.* ***3)*** *Cross-links are removed, DNA is extracted and sequencing adapters are ligated to the DNA fragments.* ***4)*** *Library is ready to be sequenced to find out which DNA fragment was bound to the target protein.*

To overcome these limitations, several other genome-wide methods for DSBs detection have been developed, which can be divided into two main categories; i) those that can map unrepaired DSBs or exposed DNA ends; and ii) those capable of mapping repaired DSBs.

Unrepaired-DSB mapping methods

1.6.2 Breaks labeling enrichment on streptavidin and next-generation sequencing (BLESS)

BLESS[76] was the first method developed for direct genome-wide DSB detection. In BLESS, cells containing endogenous or induced DSBs are cross-linked with formaldehyde, lysed and briefly incubated with proteinase K in order to purify intact nuclei. Next, DSBs are blunted, 5′-phosphorylated and the DSB ends are *in situ* labeled by a short hairpin-like proximal biotinylated adapter and enriched on streptavidin beads. The captured DSBs are ligated to another hairpin-like distal adapter, polymerase chain reaction (PCR) amplified using primers binding to proximal and distal adapters and finally subjected to high-throughput sequencing [**Fig. 2**].

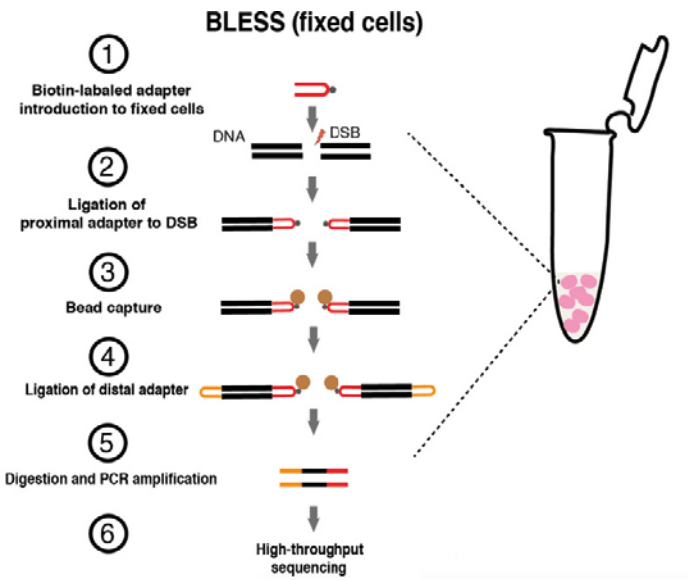

Figure 2. *BLESS detects DSBs in fixed materials.* ***1)*** *Cells are fixed by formaldehyde; intact nuclei are isolated, DSBs are in situ blunted and biotin-labeled oligonucleotide (adapter) is introduced.* ***2)*** *Adapter is ligated to DSB ends.* ***3)*** *Labeled DSB fragments are enriched by streptavidin beads.* ***4)*** *Another adapter (distal) is ligated to the free extremity of labeled DSBs.* ***5)*** *Adapter loops are released by I-SceI endonuclease and fragments are PCR amplified.* ***6)*** *Samples are subjected to high-throughput sequencing.*

The *in situ* nature of the method with no need of introducing oligonucleotide into the living cell, lack of dependency to the NHEJ pathway, and capability of detecting DSBs in tissue samples treated by *in vivo* delivery of Cas9 in mice[77], are some of the BLESS features. However, as the method works on fixed material, it can only detect DSBs that are not repaired at the time of fixation. Cell fixation per se may also be a source of inducing artificial DNA damage and breaks[78] although, a recent study reported that DSB signals obtained from both non-fixed and fixed samples were similar with very low noise levels[79]. In addition, having a PCR step for DSB amplification can also introduce biases. Furthermore, BLESS needs more than 10^6 cells as starting material and the protocol proved to be labor intensive. To overcome these limitations, we therefore aimed to improve this method, which is the core part of this doctoral thesis.

1.6.3 Double-strand break capture (DSBCapture)

DSBCapture[80] in principle follows BLESS workflow but with the introduction of A-tailing (adding a non-templated nucleotide to the 3' end of blunted DSBs) to increase the efficiency of ligation of BLESS adapters to DSB ends, and also addition of the Illumina RA5 adapter sequence directly into the BLESS adapter [**Fig. 3**]. By these changes, the authors reported an increased ligation efficiency that led to higher sensitivity of DSB detection compared to BLESS. DSBCapture was able to detect DSBs induced by EcoRV in HeLa cells, AsiSI system in U2OS cells – i.e., an engineered cell line, known as DIvA that relies on the expression of the AsiSI restriction enzyme, which upon 4-hydroxytamoxifen treatment induces about a hundred DSBs, where the method detected 74 out 100 most γH2AX-enriched sites. In addition, DSBCapture was used for endogenous DSBs detection in normal human epidermal keratinocyte cells, where 4.5-fold more DSBs compared to BLESS were found. Furthermore, the authors reported that DSBs were enriched at regulatory and G-rich regions. Despite the fact that DSBCapture improved BLESS detection sensitivity, however, similar to BLESS, capturing only unrepaired DSBs, the need of high number of starting materials and also having a labor-intensive protocol are some of its limitations.

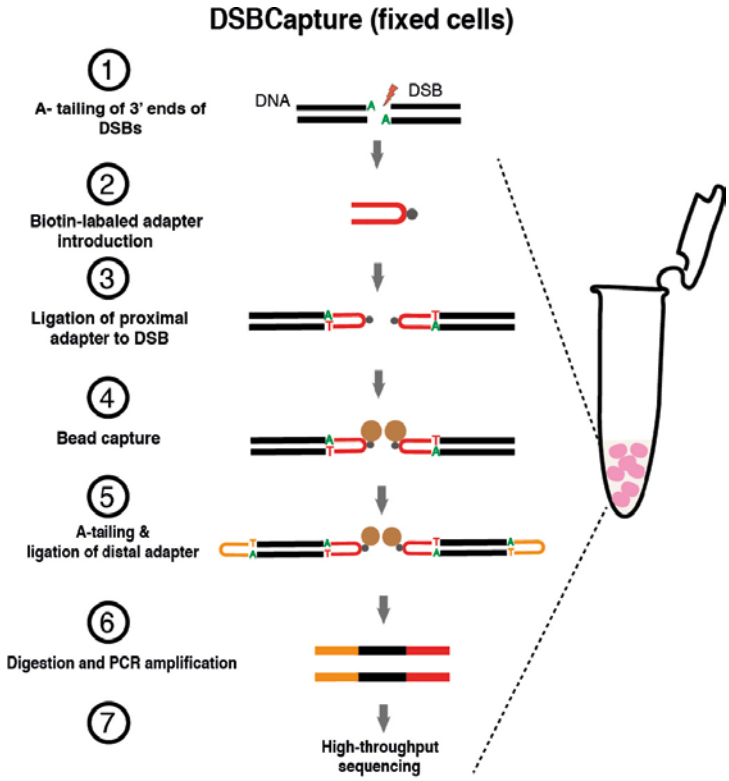

Figure 3. *DSBCapture is based on the BLESS workflow for detection of DSBs, but with the introduction of an A-tailing step. Cells are fixed by formaldehyde, intact nuclei are isolated, DSBs are in situ blunted and* ***1)*** *3' end of DSB fragments are A-tailed.* ***2)*** *Biotin-labeled oligonucleotide (adapter) is introduced.* ***3)*** *Adapter is ligated to DSB ends.* ***4)*** *Labeled DSB fragments are enriched by streptavidin beads.* ***5)*** *Another adapter (distal) is ligated to free extremity of labeled DSBs.* ***6)*** *Adapter loops are released by I-SceI endonuclease and fragments are PCR amplified.* ***7)*** *Samples are subjected to high-throughput sequencing.*

1.6.4 End Sequencing (End-seq)

END-seq[81] also follows the design of BLESS and DSBCapture, but with the difference that cells are embedded in low-melting agarose plugs, lysed, and DSBs are ligated without any fixation process. END-seq also introduced an A-tailing step – adding a non-templated nucleotide to the 3' end of blunted DSBs – to increase the ligation efficiency of DSB-tagging adapter, in addition to incorporation of the Illumina RA5 adapter sequence into the adapter [**Fig. 4**]. Using this method, the authors were able to monitor DSBs that were induced by AsiSI restriction enzyme, zinc-finger-nuclease (ZFN) and RAG endonuclease. END-seq in principle can also be used for specificity profiling of CRISPR systems. In comparison to BLESS, a greater number of DSBs were detected by End-seq at individual cleavage sites.

In addition, the authors estimated the sensitivity of the method as to be able to detect one DSB in 10,000 cells, while the sensitivity of BLESS and DSBCapture was not reported. Despite the fact that in this method cells are not fixed, however a long incubation time of embedded cells in agarose plugs before DSB ligation (i.e., 1 hour at 50 °C, then for 7 hours at 37 °C, followed by storage at 4 °C) may itself be a source of artificial DNA damage and breaks[81]. Similar to BLESS and DSBCapture, the need for high-input samples, capturing only exposed DNA ends that are not repaired, being labor intensive and challenging on tissue samples are some limitations of END-seq.

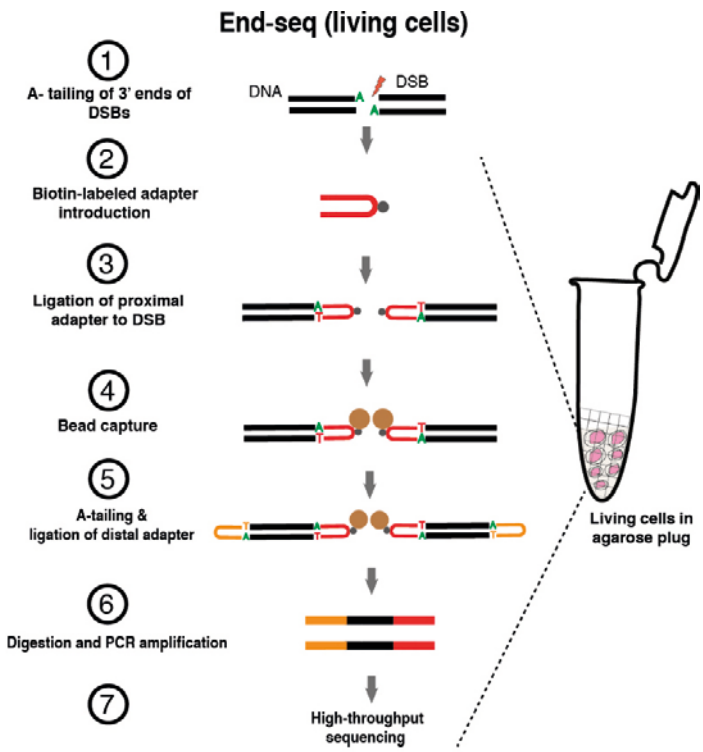

Figure 4. *End-seq is also based on the BLESS workflow for detection of DSBs, but is performed in living cells.* ***1)*** *Living cells are embedded in an agarose plug, lysed, DSBs are in situ blunted and 3' end of DSB fragments are A-tailed.* ***2)*** *Biotin-labeled oligonucleotide (adapter) is introduced.* ***3)*** *Adapter is ligated to DSB ends using a ligase enzyme.* ***4)*** *Labeled DSB fragments are enriched by streptavidin beads.* ***5)*** *Another adapter (distal) is ligated to free extremity of labeled DSBs.* ***6)*** *Adapter loops are released by I-SceI endonuclease and fragments are PCR amplified.* ***7)*** *Samples are subjected to high-throughput sequencing.*

1.6.5 iBLESS and qDSB-seq

Recently, these two new DSB detection methods aimed to further improve the original BLESS and End-seq protocol. iBLESS[79] was used in yeast models and tried to avoid fixation of cells by embedding them in agarose beads (they claimed that the efficiency of reagents diffusion is more than agarose plugs that was used in End-seq). In parallel, authors also compared different fixation procedures and showed that DSB signals obtained from both non-fixed and fixed samples were similar with very low noise levels. In addition, to improve the sensitivity of the method, the authors explored a variety of experimental parameters (e.g., duration of fixation and proteinase K conditions), and reported that intensive proteinase K treatment for overnight at 50 °C with gentle fixation (2% formaldehyde for 5 minutes) is crucial for reducing background signal. iBLESS showed to be able to capture DSBs induced by BamHI restriction enzyme (1,620 sites out of 1,667 were detected), and several other enzymes that create different kinds of dsDNA ends (i.e., NotI, AsiSI, SrfI and I-SceI). iBLESS reported the sensitivity of detecting a single DSB per 100,000 cells, which is higher than End-seq (i.e., 1:10,000). On the downside, the use of agarose beads requires a large amount of material (2x10 ^{9} cells), which is a major obstacle[79].

qDSB-seq[82] was essentially developed to bring accurate quantification power to any sequencing-based DSB labeling method, though so far it has only been applied for BLESS (in DIvA cells) and iBLESS (yeast cells). In qDSB, a “spike-in” strategy was introduced to BLESS/iBLESS protocols through inducing low-frequency DSBs at known loci using site-specific restriction enzymes, which allows for the calibration of BLESS/iBLESS data, and to calculate the absolute frequency of DSBs per cell. The quantification of qDSB-seq is based on the proportion of reads originating from induced and studied DSBs. The method was used for quantification of Top1-dependent DSBs at natural replication fork barriers, DSBs induced by radiomimetic drug and replication stress. qDSB-seq showed to be accurate, stable, robust and claimed to be potentially applicable in combination with any genome-wide DSB labeling method.

1.6.6 In vitro Cas9-digested whole genome sequencing (Digenome-seq)

Digenome-seq was originally developed for profiling off-target activity of CRISPR systems [**Fig. 5**]. Its strategy is to identify the insertions and deletions (indels) made by *in vitro* Cas9/Cpf1 cleavage of purified DNA, followed by whole-genome sequencing (WGS)[83,84]. *In vitro* digestion on cell-free DNA provides the advantage that the method is not limited to cell-based factors such as chromatin. In addition, the lack of need for PCR amplification before WGS and the ability of multiplexing for off-target detection of several sgRNA at the same time shows the progression of Digenome-seq. One major limitation of the method is the high cost of WGS. In addition, as the technology lacks the enrichment strategy for cleavage sites, all of the non-cleaved sequences are also sequenced, which causes increased background noise reads. Furthermore, *in vitro* digestion does not consider the presence of chromatin and the 3D structure of nucleus that may affect the cleavage, repair and, specificity of RNA-guided nucleases.

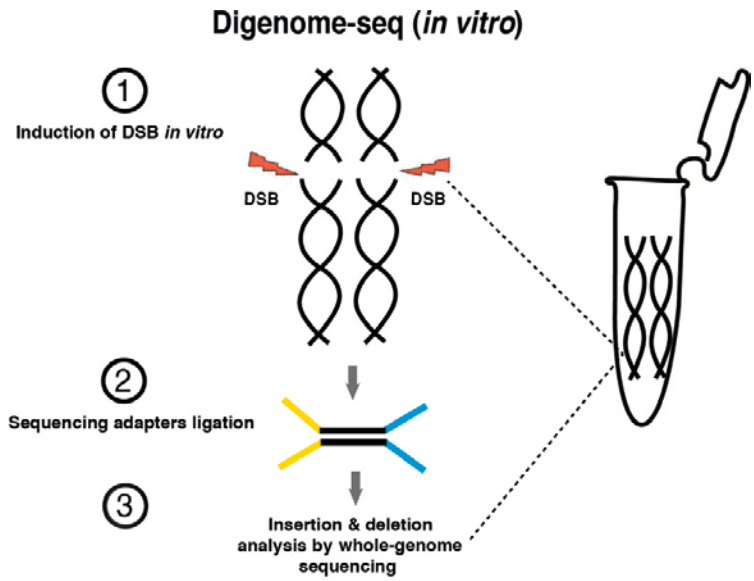

Figure 5. *Digenome-seq is an in vitro assay to find on/off-target sites of nuclease-induced DSBs.* ***1)*** *Purified genomic DNA is subjected to nucleases to create DSBs.* ***2)*** *Sequencing adapters are ligated.* ***3)*** *Whole-genome sequencing is performed for analysis of insertions and deletions made by nucleases.*

1.6.7 CIRCLE-seq

This method was developed to improve the high background noise of Digenome-seq. In CIRCLE-seq[85], gDNA is sheared and circularized though intramolecular ligation. Then Cas9 cleavage site within the circularized DNA molecule is cut by Cas9, making it linearized and creating new DNA ends for ligation of sequencing adapter **[Fig. 6].** CIRCLE-seq proved to substantially reduce the background noise by enrichment strategy for sequencing cleaved gDNA and requirement of low-sequencing depth (only 4-5 million reads).

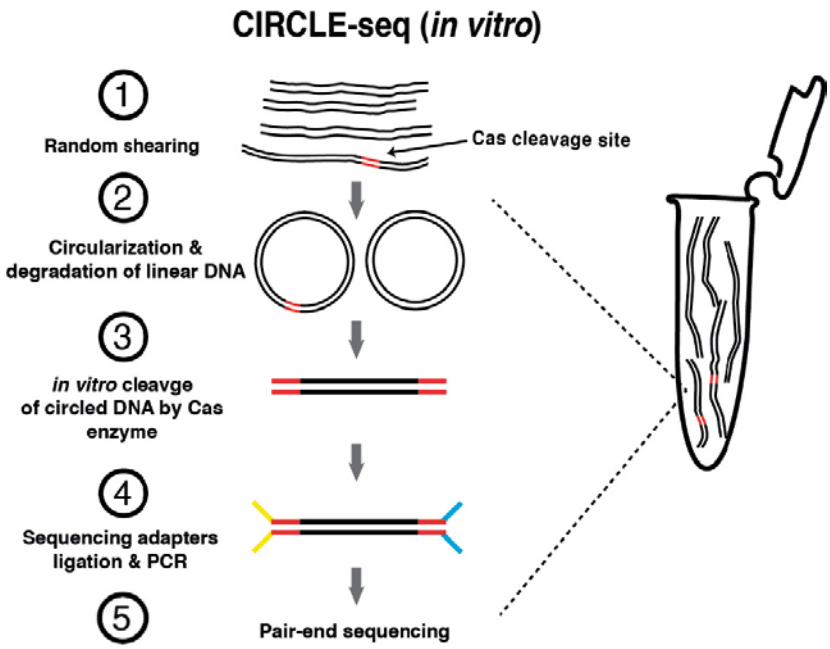

Figure 6. *CIRCLE-seq workflow:* ***1)*** *purified gDNA is randomly sheared to average length of 300 bp.* ***2)*** *Fragmented DNA is circularized by intramolecular ligation and the remaining undesired linear DNA molecules are degraded away by exonuclease treatment.* ***3)*** *Circularized DNA molecule containing Cas cleavage site (red) is linearized by in vitro Cas treatment, which creates newly cleaved DNA ends.* ***4)*** *The ends are ligated to sequencing adapter, PCR amplified and* ***5)*** *subjected to pair-end sequencing.*

1.6.8 Integrase defective lentiviral vector (IDLV)

This *in vivo* method is amongst the earliest based on tagging the DSBs generated by specific nucleuses with double-stranded integrase defective vector, which gets trapped into the broken DNA through the NHEJ-repair pathway. Labeled DSBs are then enriched by LAM-PCR (Linear amplification–mediated PCR) using primers specific to long terminal repeats (LTRs) of vector and finally subjected to high-throughput sequencing [**Fig. 7**]. Using this method, several reports showed the ability of IDLV for detection of zinc finger nucleases (ZFN), transcription activator-like effector nuclease (TALEN) and CRISPR/Cas9 nucleases off-target sites [86,87]. As a method that can *in vivo* label DSBs through integration of viral vector, makes it a good choice for monitoring the DSBs generation and repair processes that occur in living cells. However, dependency on NHEJ repair pathway, transfection efficiency, possible integration of IDLV at varying distanc from actual break sites, low detection frequency and high costs are some of the IDLV limitations.

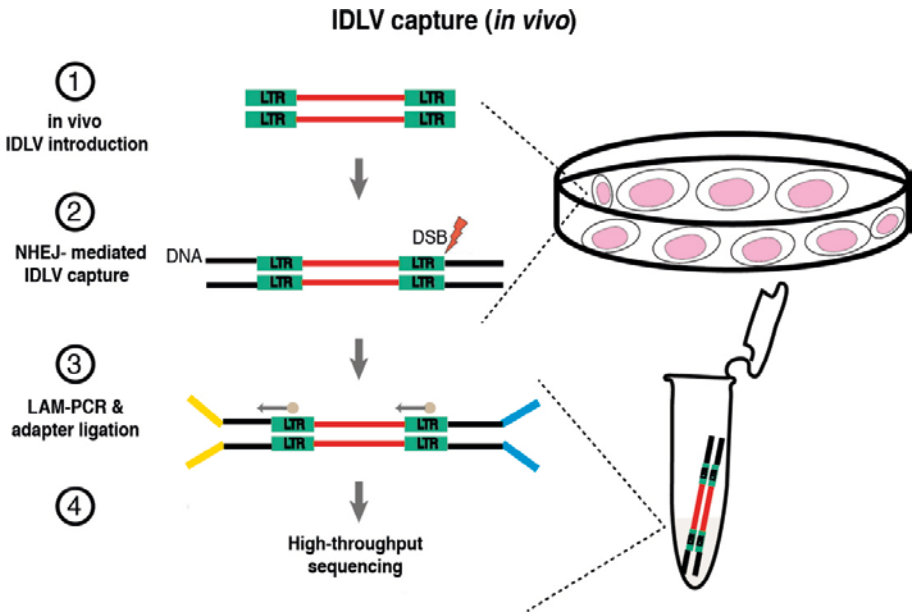

Figure 7. *IDLV capture is based on in vivo integration of IDLV vector into DSBs sites.* ***1)*** *IDLV is introduced into living cells.* ***2)*** *It is then captured into DSB sites through NHEJ.* ***3)*** *Labeled DSB is enriched by LAM-PCR using primers specific to LTRs.* ***4)*** *The sample is subjected to high-throughput sequencing.*

1.6.9 Genome-wide unbiased identification of DSBs enabled by sequencing (GUIDE-seq)

In this *in vivo* technology that was also developed for detection of off-targets of CRISPR systems, a small end-protected and blunted double-stranded oligodeoxynucleotide (dsODN) is incorporated *in vivo* through NHEJ-mediated pathway to DSBs that are generated by Cas9. The tagged DSBs are PCR amplified and subjected to high-throughput sequencing[88][**Fig. 8**]. Guide-seq proved to be a very sensitive method by detecting off-targets occurring at the frequency of 0.1% or lower in a cell population. Although, this method has been used frequently for Cas9 off-target detection, transfection efficiency of dsODN into the cells, dependency to the NHEJ- mediated repair process to incorporate the dsODN, and being challenging on primary cells and tissue samples are some of GUIDE-seq limitations.

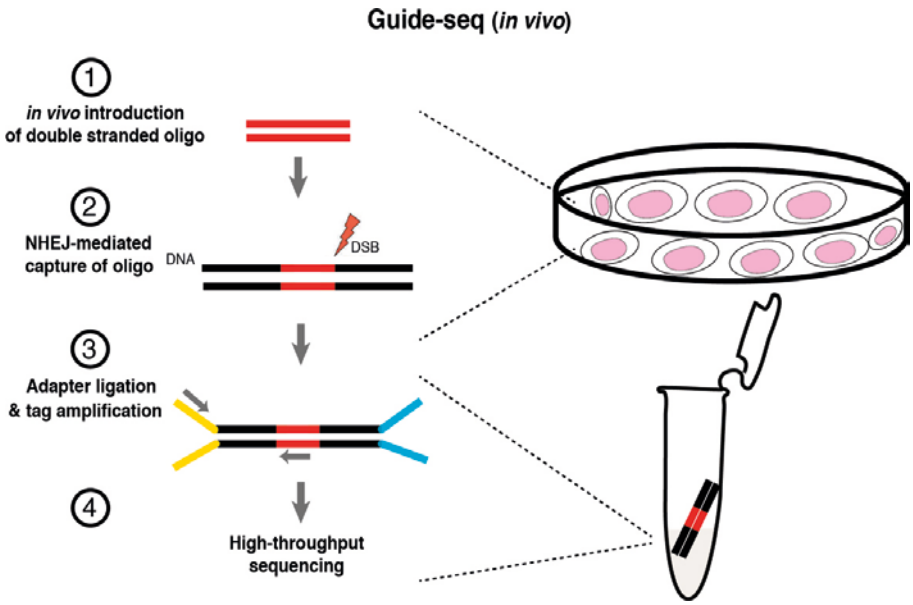

Figure 8. *Guide-seq is an in vivo based assay for on/off target detection of DSBs formed by nucleases.* ***1)*** *In living cells, DSB is induced by nuclease of interest and then double-stranded oligodeoxynucleotide (dsODN) is introduced to living cells.* ***2)*** *dsODN is incorporated to DSB site through NHEJ-mediated capture.* ***3)*** *Labeled DSB is amplified using dsODN specific primer and sequencing adapters are ligated.* ***4)*** *Prepared library is subjected to high-throughput sequencing.*

1.6.10 Linear amplification–mediated high-throughput genome-wide translocation sequencing (LAM-HTGTS)

HTGTS method was initially invented for the detection of translocation events upon DSB induction[89]. The strategy is based on *in vivo* induction of DSB in a known genomic region by nucleases, and uses this DSB as a "bait" that translocate to another DSB from another genomic region as a "prey". The translocated junction is then PCR-amplified and exposed to high-throughput sequencing. Using this method, the authors were able to show that DSBs have a tendency to translocate with regions that are actively transcribed[89]. Recently, LAM-HTGTS [**Fig. 9**], an improved version of HTGTS by introducing linear amplification strategy has been applied for CRISPR/Cas9 off-target detection through chromosomal translocation between the on-target (bait) and off-target (prey) DSBs. The translocated junction is then linearly amplified by known primer to bait sequence, and a library is prepared for high-throughput sequencing[90]. LAM-HTGTS method confirmed the off-target sites detected by Guide-seq, yet with additional sites identified unique to LAM-HTGTS. However, as the occurrence of translocation events induced by Cas9 are scarce, a large input sample is required for their detection. In addition, the influence of 3D genome organization biases the translocation frequencies to occur on regions with close nuclear proximity, by which the frequency of DSBs can be underestimated.

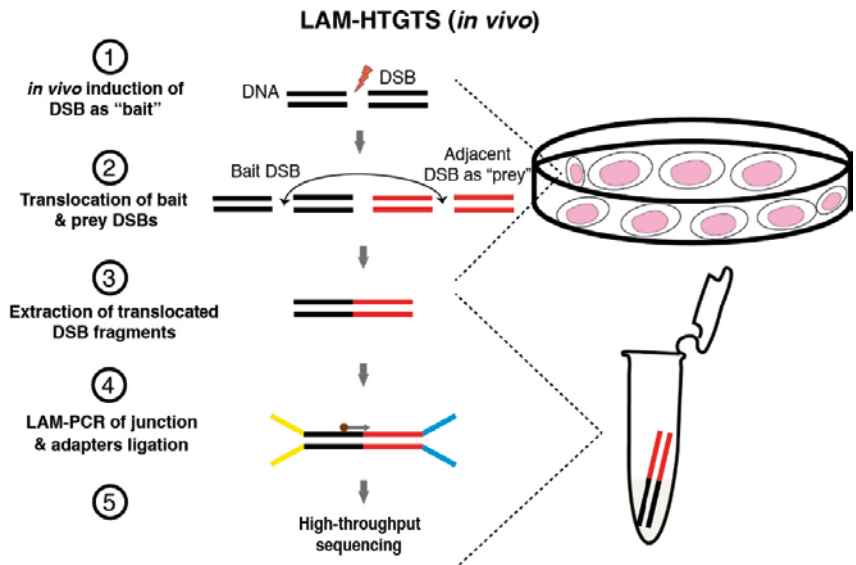

Figure 9. *LAM-HTGTS is an in vivo assay for the detection of DSBs through translocation events of known DSB (bait) with other unknown DSBs (prey).* ***1)*** *in vivo induction of DSB (bait) by nucleases.* ***2)*** *Translocation of bait DSB with other DSB (prey) in its vicinity.* ***3)*** *Translocated DSB fragments are extracted into a tube.* ***4)*** *Linear amplification PCR (LAM-PCR) of translocated junction using biotinylated primer targeting bait DSB and ligation of sequencing adapters.* ***5)*** *Sample is then subjected to high-throughput sequencing.*

2 DOCTORAL THESIS

2.1 Aims of the study

As genomic instability is a hallmark of cancers[91,92] and given the importance of DSBs in driving SVs/CNAs in cancer and their role in genome-editing technologies, it is of great interest to have methods that can profile DSBs and CNAs with high accuracy and genome-wide scale. **The overall aim of this thesis was to develop molecular tools for genome-wide detection and quantification of DSBs and CNAs in cells and tissue samples, by which we can collectively better understand the biology behind DSBs, assess the safety profile of CRISPR systems and investigate how DSBs relate to CNAs.**

The specific aims are as follows:

Paper I

- To establish a novel genome-wide technology to precisely map the location and frequency of DSBs in low-input samples, and assess its specificity and sensitivity.
    - o Develop Breaks Labeling *In Situ* and Sequencing (BLISS) to profile DSBs in low-input samples
    - o Verify the versatility of BLISS in a variety of cell models, tissue sections and biopsies
    - o Demonstrate BLISS sensitivity and specificity for detection of induced-DSBs in known genomic locations using CRISPR systems
- To generate a database of endogenous DSBs occurring in different cancer cell types and compare them to CNA breakpoints in tumor

Paper II

- o Apply BLISS technology to widely used normal and cancer cell lines that are well characterized by different assays such as DNase-seq, Chip-seq, Repli-seq, RNA-seq for genomic and epigenomic features (e.g. open chromatin, POL2B & CTCF binding, histone modifications, replication timing and gene expression) that have been implicated in genome fragility in the Encyclopedia of DNA Elements (ENCODE) to construct models that can quantitatively measure how these genomic and epigenomic features influence DSB susceptibility
- o Construct DSB models based on BLISS data from cell lines, compare the frequency of DSBs and CNA breakpoints in tumor sequencing repositories in order to ask how the generated DSB models relate to the pattern of CNAs in different cancers

Paper II

- To establish a novel technology for reduced representation genome sequencing to characterize CNAs and their heterogeneity in tumor samples
    - o Develop CUTseq to perform copy number profiling in low-input samples
    - o Verify the versatility of CUTseq in a variety of cell models and FFPE (Formalin-Fixed Paraffin-Embedded) tissues samples
    - o Assess DNA CNAs and intratumor heterogeneity of CNAs in primary and metastatic breast cancer lesions using CUTseq

The abovementioned aims were addressed in the papers that constitute this thesis and the approaches to reach them required multidisciplinary work combining cell and molecular biology assays, sequencing and microscopy methods, and bioinformatic tools.

2.2 Key methodologies for BLISS and CUTseq

2.2.1 Cells and tissues

For BLISS:

KBM7 cells were obtained from Oscar Fernandez-Capetillo (SciLifeLab, Stockholm, Sweden) and cultured in Iscove's modified Dulbecco's medium (IMDM, Life Technologies, cat. no. 10829018) supplemented with 10% fetal bovine serum (FBS, Gibco, cat. no. F2442). U2OS cells were obtained from Prof. Mats Nilsson (SciLifeLab, Stockholm, Sweden) and cultured in Dulbecco's modified Eagle's medium (DMEM, Life Technologies, cat. no. D0819) supplemented with 10% FBS. Mouse embryonic stem cells (mESCs) were obtained from Dr. Simon Elsaesser (SciLifeLab, Stockholm, Sweden) and cultured in in minimal essential medium (MEM, Sigma, cat. no. M2279), supplemented with 20% FBS, 1% GlutaMAX (Gibco, cat. no. 35050061), 1% nonessential amino acids (Gibco, cat. no. 11140035), 1% sodium pyruvate (Gibco, cat. no. 11360070), and 0.2% β-mercaptoethanol, in the presence of leukemia inhibitory factor (Sigma cat. no. L5158-5UG) corresponding to 1,000 U ml^{-1}. HEK 293T cells were bought from ATCC and cultured in in DMEM supplemented with 10% FBS. Liver biopsies were obtained from wild-type 6-weeks old C57/BL6 male mice sacrificed following the guidelines in the MIT protocol #0414-027-17 'Modeling and Treating Genetic Disease Using Targeted Genome Engineering' (IACUC AWA #A3125-01, IACUC #0411-040-14, approval date 5/16/2013).

For CUTseq:

The cell lines were purchased from ATCC: IMR90 (cat. no. CCL-186) cells were cultured in MEM (Gibco, cat. no. 10370021) supplemented with 10% non-heat-inactivated fetal bovine serum (Gibco, cat. no. 16000044), 2 mM L-glutamine (Sigma, cat. no. 59202C) and 1% non-essential amino acids (Gibco, cat. no. 11140035); A549 (cat. no. CCL-185) cells were cultured in RPMI 1640 (Sigma, cat. no. R8758) supplemented with 10% heat-inactivated FBS (Sigma, cat. no. F9665). MCF7 (cat. no. HTB-22), HeLa (cat. no. CCL-2) and Caov3 (cat. no. HTB-75) and BT474 (cat. no. HTB-20) were cultured in DMEM (Sigma, cat. no. D6429) supplemented with 10% heat-inactivated FBS (Sigma, cat. no. F9665). SKBR3 (cat. no. HTB-30) cells were cultured in McCoy's 5A (Sigma, cat. no. M9309) supplemented with 10% heat-inactivated FBS (Sigma, cat. no. F9665). Cells were maintained in a humidified at 37 °C containing 5% CO_2. All cell lines were tested to be mycoplasma free using MycoAlert™ Mycoplasma Detection Kit (Lonza, cat. no. LT07-118).

FFPE tissues from Turin (*TRN* samples) of 31 tumor specimens from different origin – gastrointestinal stromal tumor (GIST), colon adenocarcinoma (COAD), breast invasive carcinoma (BRCA), and melanoma (MELA) – were collected at the Pathology Unit of IRCC Candiolo, Italy (ethical permission "Profiling" # 001-IRCC-00IIS-10). FFPE tissue sections from Karolinska Institutet (*KI* samples) were collected from 14 female breast cancer patients – one section of 4 μm thickness per lesion from primary and distant metastases – with ethical permission from Karolinska Institutet no. 2013/1273-31/4 with amendment 2013/1739-32.

2.2.2 Sample preparation for BLISS and CUTseq

For BLISS

Cell lines were either grown directly onto 13 mm coverslips (VWR, cat. no. 631-0148) or spotted onto poly-L-lysine (Sigma, catalogue number P8920-100ML) pre-coated coverslips. For CRISPR experiments, HEK293T cells were grown in a 24-well plate pre-coated with poly-D-lysine (Merck Millipore, catalogue number A003E), and finally cells were fixed in paraformaldehyde 4%. For BLISS in tissue we used two approaches: 1) Tissue cryopreservation and sectioning, where freshly obtained mouse liver biopsies were fixed in 4% paraformaldehyde for 1h at 25 °C, then immersed in sucrose gradient (15% overnight and then 30% until the tissue sank) before embedding in optimal cutting temperature medium (OCT). Then 30 μm-thick sections were mounted onto microscope slides, air-dried for 60 min at room temperature and stored at 4 °C until processing by BLISS. 2) Obtaining nuclei from the biopsies, where freshly extracted mouse liver biopsies were cut into small pieces, transferred into a DNA LoBind tube (Sigma, cat. no. Z666548) containing nucleus isolation buffer (NaCl 146 mM, Tris-HCl 10 mM, $CaCl_2$ 1 mM,

$MgCl_2$ 21 mM, bovine serum albumin 0.05%, Nonidet P-40 0.2% pH 7.8) and gently rotated for 15–40 min until the tissue fragments became transparent. The nuclei were centrifuged for 5 min at 500 g, resuspended in 500 µl of 1 × PBS, and 100 µl of it was dispensed onto a poly-L-lysine (Sigma, catalogue number P8920-100ML) pre-coated 13 mm coverslip. Nuclei were let to sediment for 10 min at room temperature, followed by gentle dispensing of 100 µl paraformaldehyde 8% on top of them, which made the final concentration of 4% fixative. Nuclei were washed twice in 1 × PBS and stored at 4 °C until further processing.

For CUTseq:

All FFPE tissue sections were deparaffinized by xylene (Honeywell, cat. no. 534056), followed by immersion in ethanol gradient and hematoxylin-eosin was used to stain the sections. The gDNA from different types of samples was extracted as follow:

Cell lines: obtained pellets after trypsinization of cells were washed twice in 1xPBS (Ambion, cat. no. AM9625) and lysed using a buffer containing 10 mM Tris-HCl, 100 mM NaCl, 50 mM EDTA, 1% SDS, 19 mg/ml Proteinase K (NEB, cat. no. P8107S), pH 7.5, incubated overnight at 55 °C on a thermomixer, shaking at 800 rpm. gDNA was purified using common phenol-chloroform extraction protocol, quantified using Qubit 2.0 Fluorimeter and High Sensitivity DNA Kit (Agilent, cat. no. 5067-4626).

TRN tissue sections: gDNA from five 10 µm-thick sections with more than 50% tumor cells were extracted. We obtained 200 ng gDNA after manual dissection using the QIAamp DNA FFPE Tissue Kit (Qiagen, cat. no. 56404). Extracted gDNA was quantified using Qubit 2.0 Fluorimeter and High Sensitivity DNA Kit (Agilent, cat. no. 5067-4626).

KI tissue sections: in order to extract gDNA from multiple small regions within a single FFPE section, we used PinPoint Slide DNA Isolation System™ (ZymoResearch, cat. no. D3001), which acts as a glue that is air-dried, peeled off and placed into a DNA LoBind tube (Sigma, cat. no. Z666548). The remaining parts of the tissue section were also collected as a whole using the same PinPoint system into a separate DNA LoBind tube (Sigma, cat. no. Z666548). The tissue is then lysed in the same buffer that we used for cell lines, purified using common phenol-chloroform extraction protocol and quantified using Qubit 2.0 Fluorimeter and High Sensitivity DNA Kit (Agilent, cat. no. 5067-4626). We note that gDNA extraction with silica-based kits is also perfectly compatible for cell lines and tissue sections.

2.2.3 Workflow of BLISS and CUTseq

For BLISS

A detailed step by step protocol has been published in "Nature Protocol Exchange" (https://protocolexchange.researchsquare.com/article/nprot-5597). Briefly, fixed samples on a solid-surface were permeabilized, incubated for 1 hour in a blunting reaction mix (NEB, cat. no. E1201L) at room temperature, followed by *in situ* ligation of double-stranded oligo to DSB ends in a T4 DNA ligase reaction mix (NEB, cat. no. M0202M) for 16–18 hours at 16 °C. Next, gDNA is extracted, incubated in proteinase K (NEB, cat. No. P8107S) for at least 5 hours at 55 °C on a thermomixer, followed by purification. Purified gDNA was sonicated and *in vitro* transcribed using T7 RNA polymerase (ThermoFisher, cat. no. AM1334) for 14 hours at 37 °C. Finally, RNA product was used for library preparation by a modified Illumina TruSeq Small RNA Library Prep Kit (RS-200-0012).

For CUTseq

A detailed step by step protocol has been published in "Nature Protocol Exchange" (https://protocolexchange.researchsquare.com/article/d0ef0512-37b2-461b-9687-eeec11f167e1). Briefly, extracted gDNA is digested with the restriction enzyme HindIII or NlaIII (NEB, cat. no. R3104L or R0125L) for 16–18 hours at 37 °C, followed by ligation of CUTseq double-stranded oligo (compatible with staggered ends) in a T4 DNA ligase reaction mix (NEB, cat. no. M0202M) for 16–18 hours at 16 °C. Ligated gDNA was cleaned up, sonicated and *in vitro* transcribed using T7 RNA polymerase (ThermoFisher, cat. no. AM1334) for 14 hours at 37 °C. Finally, RNA product was used for library preparation by a modified Illumina TruSeq Small RNA Library Prep Kit (RS-200-0012).

2.2.4 Cas or Cpf1 expression constructs and transfections (For BLISS)

We used plasmids containing the spCas9 and sgRNA cassette targeting *EMX1* locus (5′-GAGTCCGAGCAGAAGAAGAAgGG-3′) and *VEGFA* gene locus (5′-GGTGAGTGAGTGTGTGCGTG tGG-3′). The same expression vector was used to clone AsCpf1 and LbCpf1 together with their cognate sgRNA for direct comparison of these nucleases. For transfection, 24-well plates were coated using poly-D-lysine (Merck Millipore, catalogue number A003E) and cells were seeded at a density of ~125,000 per well and grown for 16-18h to reach 60-70% confluency. Once cells were ready, a mix of 2 µl of Lipofectamine 2000 (Life Technologies, catalogue number 11668019) and 500 ng of Cas9/Cpf1 plasmids in 100 µl of OptiMEM (Gibco, catalogue number 31985062) was used for each well.

2.2.5 Immunofluorescence, Hematoxylin-eosin staining, imaging and automated cell counting

For BLISS:

Immunostaining of DSB marker γH2A.X was performed using a mouse anti-phospho-histone H2A.X (ser139) primary antibody (Millipore, catalogue number 05-636) diluted 1:1000 in blocking buffer – 3% BSA with 0.1% Tween-20 – and a goat anti-mouse IgG (H+L) conjugated with Alexa Fluor 647 (Thermo, catalogue number A-21235) secondary antibody diluted 1:1000 in blocking buffer. Stained γH2A.X foci were imaged every 0.4 μm using Z stack module to cover entire volume of nuclei by a × 40 oil objective and an LSM 780 confocal microscope (Zeiss). To count γH2A.X foci we used custom-made scripts implemented in MATLAB.

For CUTseq:

All tissue sections for this study were deparaffinized by xylene (Honeywell, cat. no. 534056) followed by immersion in ethanol gradient. Hematoxylin-eosin was used to stain 35 FFPE breast cancer sections that were used for multi-region tumor sequencing. Each tissue section was scanned by Eclipse Ti inverted wide-field fluorescence microscope (Nikon, Japan) using in phase contrast mode with 10X objective. In order to count the number of cells captured for gDNA extraction from the tissue region, 16 independent different breast cancer FFPE tissue sections were stained by 1ng/ul Hoechst 33342 (ThermoFisher, cat. no. 62249) in 1x PBS, for 15 min at 30 °C. 11 cm region of sections was scanned using Eclipse Ti inverted wide-field fluorescence microscope (Nikon, Japan) with 40X objective. Automatic nuclei segmentation was performed using Ilastik[93] open-source pixel classifier software, by training the software on a single tissue scan. Cells were counted in five 1.71.5 mm regions of each tissue sections that overlap with tumor dense areas annotated in the same section by a certified pathologist.

2.2.6 BLISS and CUTseq adapters

We purchased the oligonucleotides as standard desalted oligos from Integrated DNA Technologies (IDT). UMIs were generated by random incorporation of the four standard dNTPs using the ‘Machine mixing’ option. Oligos were diluted to 10 μM and sense oligo phosphorylated for 1 h at 37 °C with T4 Polynucleotide Kinase (NEB, cat. no. M0201), after which an equimolar amount of anti-sense oligo was added. PCR thermocycler was used to anneal both oligos by incubating for 5 min at 95 °C, followed by gradual cooling down to 25 °C over a period of 45 min (1.55 °C min−1).

2.2.7 Sequencing and data pre-processing

For BLISS and CUTseq:

The sequencing was performed on Illumina NextSeq 500 platform using NextSeq 500/550 High Output Kit v2 chemistry for single-end (1x76) or paired-end (2x150) sequencing. Based on index sequences of pooled libraries, raw sequencing reads were demultiplexed by Illumina's BaseSpace and FastQ files were generation. A custom-built pipeline was used to scan for the reads that contain the proper prefixes (i.e., 8nt UMI and 8nt sample barcode) with allowance of up to two mismatches in UMI and up to one mismatch in the barcode. Then the prefixes were clipped off and stored, reads were aligned to (GRCh37/hg19 for human, NCBI37/mm9 for mouse) reference genome with BWA-MEM (version 0.7.17-r1188)[94]. The reads with the mapping quality scores of ≥ 30 were kept, and a further filtering step based on UMI sequences to identify and remove PCR duplicates was applied by searching for proximal reads (at most 30 bp apart in the reference genome with at most two mismatches allowed in the UMI sequence). Eventually, for downstream analysis a BED file containing a list of genomic locations associated with a number of unique UMIs was generated.

2.3 Summary of research papers

2.3.1 BLISS is a method to profile natural and artificially induced DSBs

Given the importance of DSBs in pathological disorders and their role in revolutionary genome-editing technologies such as CRISPR systems, precisely mapping the location and frequency of DSBs along the genome is of great interest for better understanding the biology of DSBs, genome instability and specificity of genome editing technologies. Although several technologies have been developed to map DSBs, each comes with some limitations as discussed in the introduction chapter. Therefore, to overcome the limitations related to existing technologies, in **Paper I** we developed a quantitative, genome-wide method termed Breaks Labeling *in Situ* and Sequencing (BLISS)[95] [**Fig. 10**] to profile the genomic landscape of DSBs. In BLISS, cells or tissue sections are deposited onto a solid surface (e.g. coverslips or microscope slides) and fixed in 4% PFA, which allows the subsequent reactions to be performed *in situ*, thus minimizing the introduction of artificial DNA breaks and sample loss. DSBs are blunted in situ using T4 DNA polymerase, and then a double-stranded DNA oligonucleotide adapter containing the T7 promoter sequence, followed by RA5 Illumina sequencing adapter, a short random stretch of 8–12 nucleotides that acts as unique molecular identifier (UMI), and finally a barcode sequence that enables sample multiplexing, is ligated to the blunted DSB ends. After DSB ligation, genomic DNA (gDNA) is extracted, sonicated to create fragments between 300-800 bp, and then the sequence juxtaposed to the

labeled DSB is linearly amplified by T7-mediated in vitro transcription (IVT). The RNA product of IVT is then used to prepare sequencing library to be loaded onto Illumina sequencing platforms. BLISS adapter design features some important advantages; linear amplification mediated by T7 approach reduces PCR amplification biases as previously described[96] and it makes the assay more sensitive as only tagged DSB sites are amplified; the UMI is used to filter out PCR duplicates[97] and enables distinguishing and quantifying breaks occurring at the same nucleotide position in different alleles or cells; and the barcode is employed to label different samples for multiplexing and cost reduction.

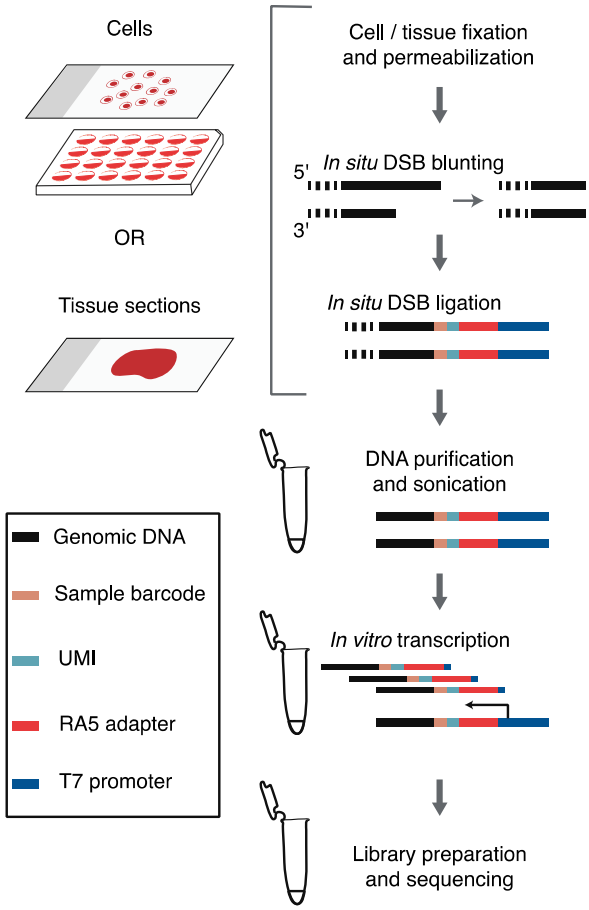

Figure 10. *BLISS is an in situ method for detection of DSBs in fixed low-input materials. Cell or tissue samples are fixed on a solid surface and a double-stranded synthetic oligo linker (see legend) is ligated to blunted DSB ends. Tagged DSBs are linearly amplified by in vitro transcription using the T7 promoter sequence that is integrated in the BLISS linker. The second sequencing adapter is ligated and the sample is PCR amplified. Finally, a BLISS library is subjected to high-throughput sequencing.*

First, we demonstrated the versatility of BLISS in different sample types, from cell lines to tissue sections and nuclei suspension obtained from biopsies. Using BLISS, we were able to quantify and map endogenous and drug-induced DSBs (for example by DNA topoisomerase II inhibitor etoposide) in low-input samples. We measured endogenous DSBs in low input samples (approx. <5000) of three KBM7 cell replicates sequenced at saturation and demonstrated that BLISS was accurate and quantitative enough to be able to estimate 80–100 DSBs per cell, which was in line with the number of γH2A.X foci quantified by microscopy in the same cell line [**Fig. 11 A-C**]. The number of DSBs per cell was estimated by counting the number of unique reads – meaning a correct BLISS barcode, followed by a unique UMI, and a read that mapped to a unique genomic location – divided to the number of cells or genome-copies equivalent.

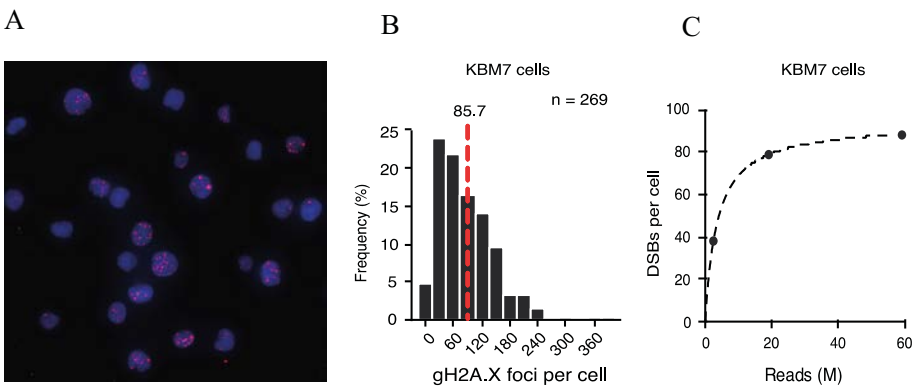

Figure 11. (A) *immunofluorescent staining of γH2A.X. Blue; DAPI, Red; γH2A.X signals.* ***(B)*** *Distribution of the number of γH2A.X foci per cells. n, number of cells analyzed. The number near the dashed red line equals the mean number of γH2A.X foci per cell.* ***(C)*** *Estimated number of DSBs per cell in three replicates of KBM7 cells sequenced at increasing sequencing depth. Dashed line, hyperbolic interpolation.*

To show the applicability of BLISS in primary cells and tissue, which would help investigation of DNA damage and repair processes in animal models and clinical samples, we profiled endogenous DSBs in mouse embryonic stem cells (mESCs) [**Fig. 12 A-B**], mouse liver biopsy [**Fig. 12 C-D**] and mouse tissue sections [**Fig. 12 E-F**], which revealed strong enrichment of DSBs both in the neighborhood of transcription start sites (TSS) and along the gene body of highly expressed genes, which corroborates with previous findings[80] that transcriptional-associated processes can induce DSBs that in part govern gene regulation[80,98,99]. Furthermore, the top hit genes with highest enrichment of DSB levels in liver samples revealed to be involved in liver-specific metabolic processes found by Gene Ontology (GO) analysis, indicating that BLISS is capable of capturing endogenous DSBs related to tissue-specific processes along with its versatility.

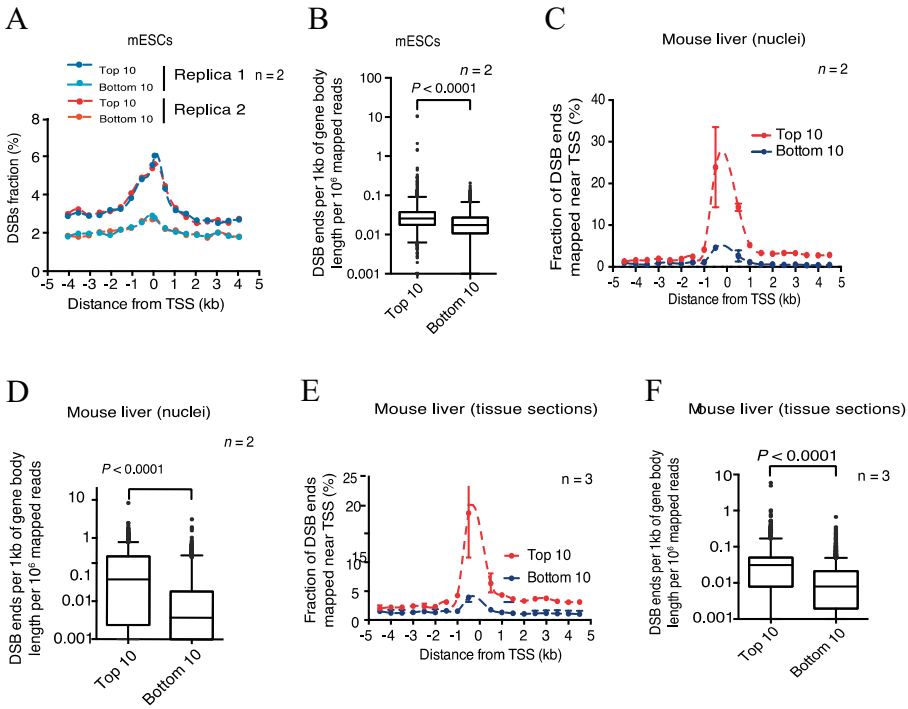

Figure 12. ***(A)*** *Percentage of sequenced DSB ends mapped ±4.5 kb around the TSS of the top 10% and bottom 10% expressed genes in two biological replicates of mouse embryonic stem cells. Dashed lines, spline interpolation.* ***(B)*** *Number of sequenced DSB ends mapped per kilobase inside the gene body of the top 10% and bottom 10% expressed genes in mouse embryonic stem cells. n, number of biological replicates. Whiskers, 2.5–97.5 percentile range. P, Mann-Whitney test.* ***(C)*** *Percentage of sequenced DSB ends mapped ±4.5 kb around the TSS of the top 10% and bottom 10% expressed genes in mouse liver nuclei. n, number of biological replicates. Dots, mean value. Whiskers, range. Dashed lines, spline interpolation.* ***(D)*** *Number of sequenced DSB ends mapped per kilobase inside the gene body of the top 10% and bottom 10% expressed genes in mouse liver nuclei. n, number of biological replicates. Whiskers, 2.5–97.5 percentile range. P, Mann-Whitney test.* ***(E)*** *BLISS on mouse liver tissue section showing sequenced DSB ends of the top 10% (red) and bottom 10% (blue) of expressed genes mapped in ±1 kb interval around the TSS. n, number of biological replicates. Dots, mean value. Whiskers, min-max range. Dashed lines, spline interpolation.* ***(F)*** *Number of sequenced DSB ends mapped per kilobase inside the gene body of the top 10% and bottom 10% expressed genes in mouse liver tissue section. n, number of biological replicates. Whiskers, 2.5–97.5 percentile range. P, Mann–Whitney test.*

Moreover, BLISS in U2OS cells treated with the topoisomerase II inhibitor etoposide enabled us to quantify the number of unique DSB ends in an increasing dose-dependent manner similar to microscopy-based γH2A.X measurements [**Fig. 13 A-B**]. We observed that upon etoposide treatment, DSBs are accumulated at recurrent genomic locations that are significantly enriched in the neighborhood of TSS [**Fig. 13 C**]. This finding was also in line with previous reports that etoposide has prominent effects around TSS – likely because of the role of topoisomerase II in relieving the torsional stress associated with replication/transcription fork progression and enhancer-promoter interactions[29,100,101] – and further demonstrated the ability of BLISS to quantitatively measure DSBs in different sample types and conditions.

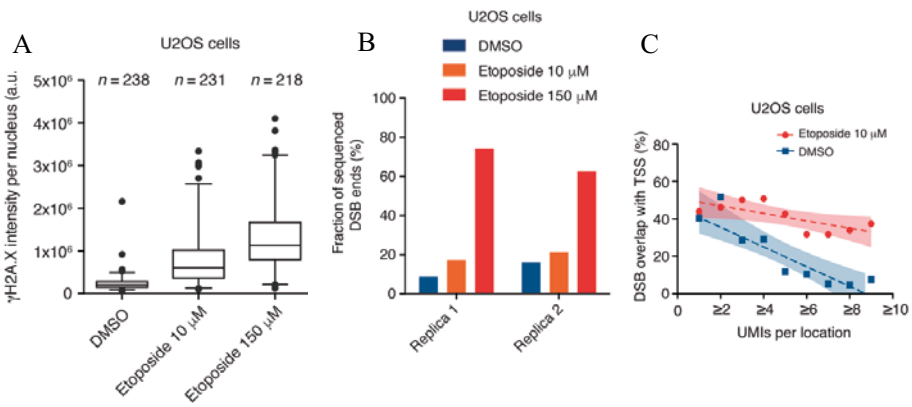

Figure 13. *(**A**) Nuclear intensity measurements of γH2A.X in untreated versus etoposide-treated U2OS cells. n, number of analyzed cells. Whiskers extend from 2.5 to 97.5 percentiles. (**B**) Percentage of sequenced DSBs in control versus etoposide-treated U2OS cells. Two biological replicates were analyzed. For each condition, the same amount of genomic DNA was loaded into a single IVT reaction, and a single sequencing library was prepared for each replicate. (**C**) Fraction of DSB locations mapped around the transcription start sites (TSS) in control versus etoposide-treated U2OS cells as a function of the minimum number of UMIs per DSB location. Dashed lines, linear interpolation. Color shades, 95% confidence intervals.*

To broaden the application of the BLISS method and assess its sensitivity at the same time, we next aimed to characterize the genome-wide on/off-target activities of two CRISPR-associated RNA-guided endonucleases, Cas9 and Cpf1 (Cas9-BLISS & Cpf1-BLISS), and also benchmark BLISS with other available gold standard CRISPR off-target screening methods, such as GUIDE-seq, Digenome-seq and BLESS (reviewed in introduction chapter). Cas9 and Cpf1 act as molecular scissors and their cutting mechanisms generate site-specific blunt and staggered DSBs respectively. However, off-target cleavage activities of these CRISPR nucleases represents a major concern as they can lead to unwanted mutations and cancer risk[102,103], and since these genome editing technologies are already used in several clinical trials, their specificity needs to be thoroughly assessed.

We developed a workflow as described in the method chapter to screen the cleavage activities of aforementioned nucleases [**Fig. 14 A**]. We first started to characterize the specificity of Streptococcus pyogenes Cas9 (spCas9) and two sgRNAs targeting EMX1 and VEGFA genes that showed high off-target frequencies in the human genome and have already been assessed by all the other three available methods. In brief, Cas9-BLISS was able to identify on-target DSB sites as expected, along with discovering of many off-target sites that were successfully validated by targeted next generation sequencing (NGS), including many sites previously identified by other methods [**Fig. 14 B-D]**. Although these comparisons showed that all methods agree on the top identified off-target sites, they differ in the number of weaker off-target sites, especially in the *VEGFA* gene.

Eventually, we aimed to compare the DNA-targeting specificity of the recently described CRISPR-associated endonuclease Cpf1[38] with Cas9 using BLISS technology. We evaluated Cpf1 from Acidaminococcus sp. (AsCpf1) and Lachnospiraceae bacterium (LbCpf1), and since the Cpf1 system recognizes a T-rich protospacer-adjacent motif (PAM) 5′ of target sites, compared to the Cas9 system where a 3′ G-rich (NGG) PAM is utilized, we selected six Cpf1 targets across four different genes for off-target evaluation using BLISS and targeted NGS, by which four targets have NGG PAMs on the 3′-end to enable a dual targeting for simultaneous comparison between these CRISPR systems. Comparing BLISS results with guides used for AsCpf1, LbCpf1, SpCas9 and eSpCas9, we consistently found fewer *bona fide* off-target sites for the two Cpf1 orthologues, suggesting that indeed Cpf1 has a higher level of specificity than Cas9, and that is in line with other recent findings[83] [**Fig. 15**].

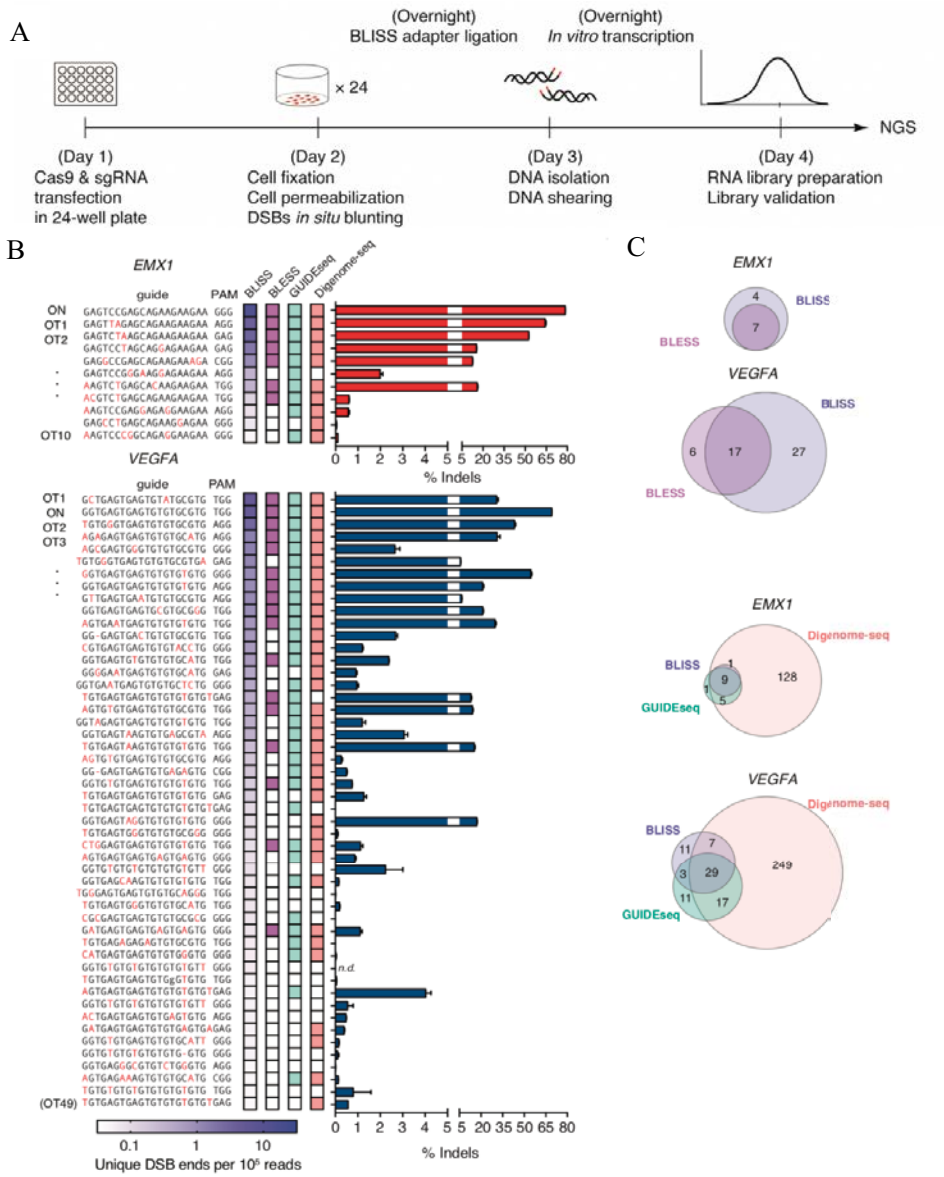

Figure 14. *(**A**) BLISS workflow for CRISPR off-target detection. (**B**) Shows on and off target sites identified by BLISS, BLESS, GUIDEseq and Digenome-seq. BLISS targets were ranked in descending order based on the number of unique DSBs aligned to the target per 10^5 unique BLISS reads. Colored boxes in the three other benchmarked methods, BLESS, GUIDEseq and Digenome-seq columns indicate when the BLISS target was previously found by these methods. Each individual site was validated by targeted deep sequencing and the percentage of reads containing an insertion or deletion (indel) is shown. (n=3, error bars show s.e.m.). ON, means on-target. OT, off-target. (**C**) Overlaps between on and off-target sites identified by BLISS versus BLESS, GUIDEseq and Digenome-seq.*

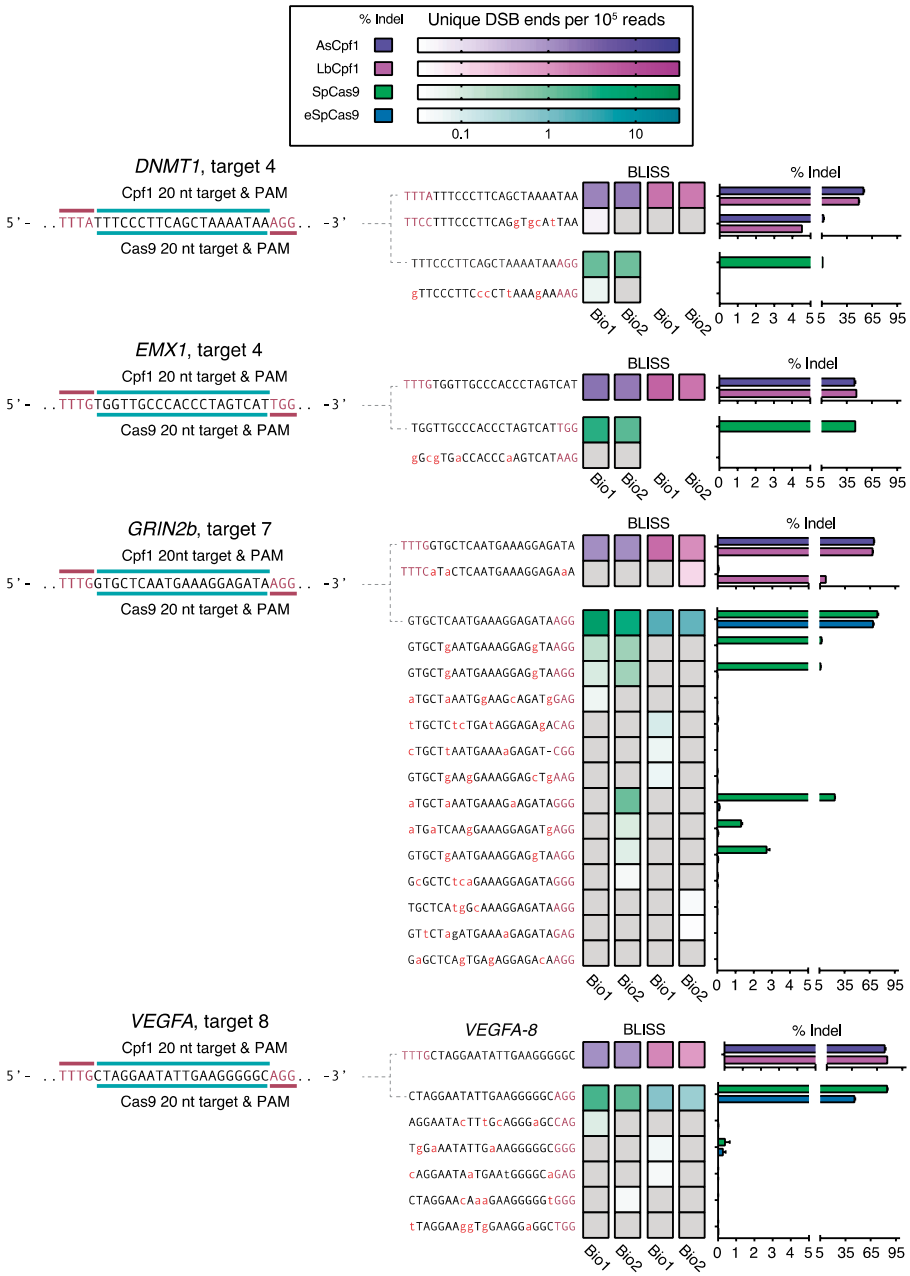

Figure 15. *Comparison of BLISS results with guides used for AsCpf1, LbCpf1, SpCas9, and eSpCas9. Individual sites identified across two bio-replicates of BLISS were validated by targeted deep sequencing and the percentage of reads containing an insertion or deletion (indel) is shown. (n=3, error bars show S.E.M.). Dotted lines indicate the matching on target sequences for the different enzymes.*

We observed that, although all the four methods are generally in agreement on the top identified off-targets, they differ in the number of weaker off-target sites, which indicates that they all can be used as complementary/alternative tools based on their limitations. These results further demonstrated the specificity, sensitivity and broad applicability of BLISS. Altogether, we developed a quantitative and highly versatile method that can offer several unique features such as: 1) direct *in situ* DSBs labeling on solid surfaces to avoid introduction of artificial breaks, providing the applicability to low-input samples and tissue sections, as well as easy scalability by performing all in situ reactions and washes on a solid surface – if input sample is not limiting, the reactions can also be done in solution in-tube which provides additional versatility; 2) lack of dependency on repair pathways (e.g. NHEJ) for DSB labeling; 3) by employing UMI for quantification of DSBs and removing PCR duplicates the method is able to discriminate the DSB events that occur at the same genomic locations but in different alleles or cells; 4) the methods is multiplexable and cost-effective using barcodes before pooling samples; and 5); and provides a genome-wide quantitative view of DSB landscapes. We demonstrated that BLISS is a quantitative method to detect endogenous and drug-induced DSBs, which is sensitive enough to assess the DNA-targeting specificity of CRISPR-associated RNA-guided DNA endonucleases. Therefore, we believe that BLISS is a powerful and versatile method for genome-wide DSB profiling to advance the study of endogenous and artificially induced DSBs in different sample types and conditions.

2.3.2 BLISS-generated DSBs data can be used to model genome fragility

It is thought that improper repair of DSBs can lead to structural variations (SVs)[104], and that each tumor contains a unique constellation of single nucleotide mutations and structural variants, of which only a tiny fraction of these mutations is cable of driving tumorigenesis[105]. Although there is evidence that SVs can drive the tumor progression, it has been difficult to determine among the many thousands of detected SVs, which ones might be advantageous for the tumor.

In a collaborative project, we aimed to computationally build a set of DSBs models to predict the frequency of expected breakage across the human genome. We speculated that cancer genomes harbor vulnerable regions (i.e., hotspots) where DSBs occur at higher frequency, while others have low frequency of genomic rearrangements (i.e., cold-spots) due to their essential integrity. One approach to find such hot and cold loci is to compare the genome-wide DSBs and observed rearrangement breakpoints frequencies in large tumor cohorts, such as TCGA (The Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium). These public sources contain a collection of large structural variations from many patients and different cohorts. Since BLISS data were not available for the

samples in these repositories, we applied a machine learning approach – random forest – on four DSB datasets produced by different DSB mapping platforms that cover three cell lines; NHEK (keratinocyte cells, DSB data generated by BLESS and DSBCapture), K562 (erythroleukemia cells, DSB data generated by BLISS), MCF7 (breast cancer cells, DSB data generated by BLISS), and applied a random forest regression model to link DSB and SV profiles.

We selected these cell types as they have been extensively profiled by the Encyclopedia of DNA Elements (ENCODE) for a variety of chromatin-binding factors and genomic features (i.e., open chromatin assayed by DNase-seq, POL2B binding, CTCF binding and five histone modifications assayed by ChIP-seq, replication timing assayed by Repli-seq, expression assayed by RNA-seq). We build the random forest regression models at 50 kb resolution in order to generate quantitative measures of the relative importance of a variety of these matched features to model DSB susceptibility [**Fig. 16**].

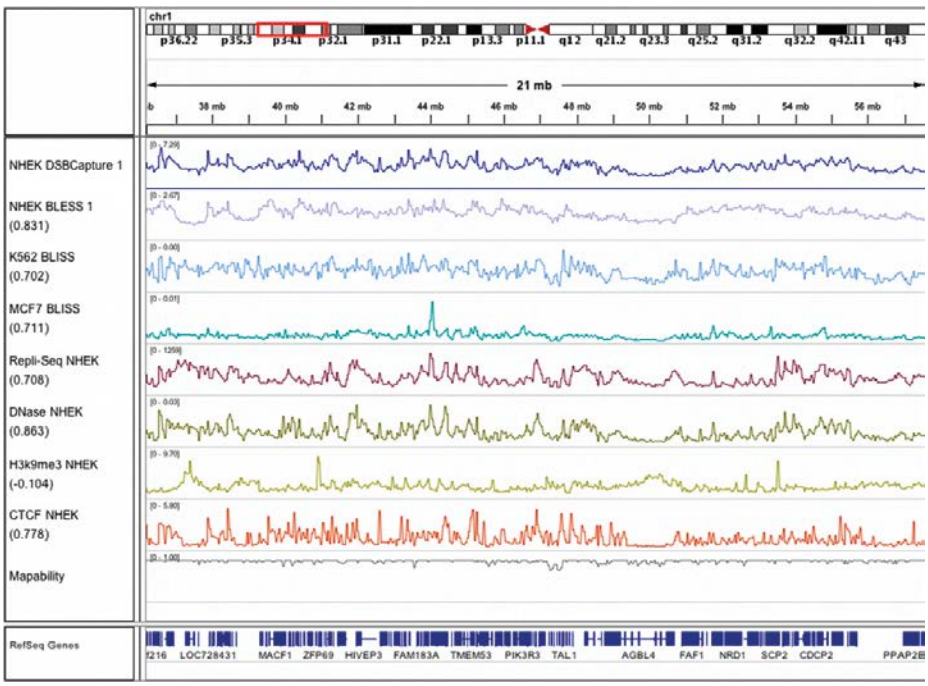

Figure 16. Genomic *tracks at 50 kb resolution across a representative region of chromosome 1, which display similar patterns of DSB frequency profiles of NHEK cells (generated by DSBCapture and BLESS methods), K562 and MCF7 cells (generated by BLISS) and a variety of chromatin and sequence features.*

Using DSB frequency models we found strong and significant correlations between predicted and observed DSB frequency in all of the four DSB datasets (Pearson's coefficients 0.83-0.92) [**Fig. 17 A**]. In addition, we found that 11 selected genomic features suffice for the construction of highly predictive accuracy models. Our analysis revealed that the most influential feature in DSB frequency prediction is replication timing across all models [**Fig. 17 B**]. We observed that early replication correlates with high DSB regions, which was in line with previous report[106]. We also found that histone marks H3K36me3, usually associated with active genes was enriched at high DSB regions and H3K9me3, generally linked to gene-poor heterochromatin, at low DSB regions [**Fig. 17 C**], which was also in agreement with previous reports that, in cancer, SVs disproportionally accumulate within the early replicating, gene-rich portions of the genome, and are rather depleted in late replicating gene-poor regions [107–109].

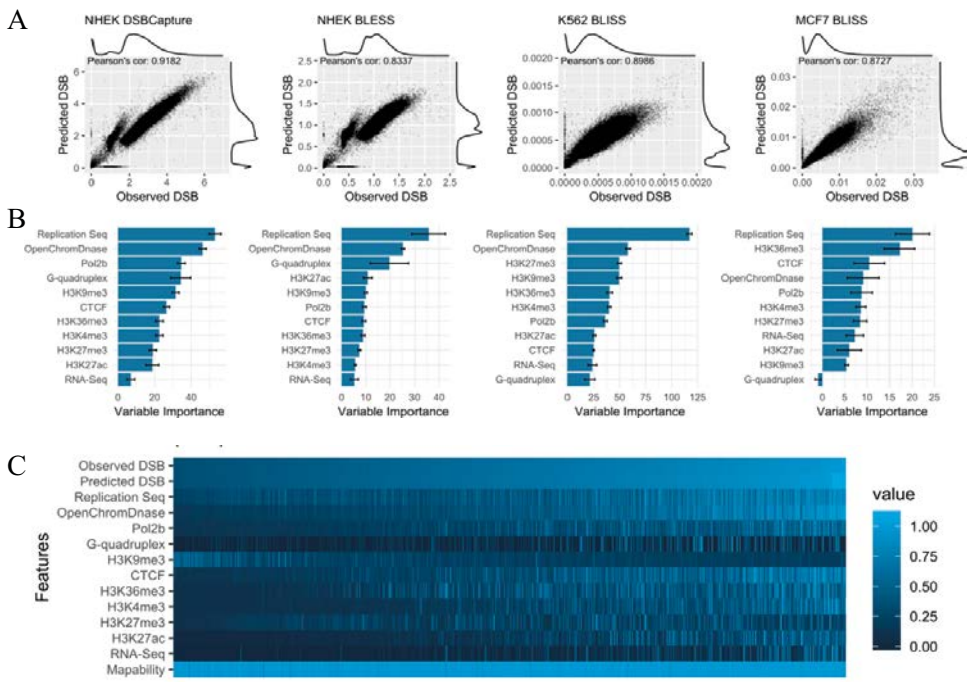

Figure 17. *DSB frequency models generated from chromatin and sequence features. (**A**) Eleven genomic features at 50 kb resolution used for random forest regression model predictions and compared to DSB frequencies datasets of four cell lines. (**B**) Particular predictive features are ranked based on their importance for the models. (**C**) Heatmap showing modeling accuracy and the polarity of genomic features, in which, columns are ordered by observed frequency of DSBs (top row) and features used to create the model ordered based on average variable importance shown in rows (the third to second to the last row).*

DNAse-seq experiments, which are used to identify regions of open chromatin, ranked as the second most important future across three models, and as the fourth in the MCF7 model. In addition, we found that RNA-seq is not a strong predictor for DSB susceptibility, despite the fact that DNAse-seq peaks are often enriched at the promoters of active genes. This observation suggests that information on open chromatin at transcriptionally active genes and associated regulatory factors (e.g., DNAseq, H3K4me4 and POL2B binding), instead of their transcription levels, is the main predictor of the frequency of DSBs.

The recurrent patterns of genomic features indicate that many factors have a similar effect on frequency of DSB in each of the investigated cell types, and therefore, a model trained in one cell type can generalize well to another cell type. For instance, a model trained in NHEK cells can be used to predict frequency of DSBs in K562 cells with its associated genomic features as input. This approach enabled us to construct predictive DSB frequency profiles for cell lines that lack actual high-resolution DSB data. In addition, we found a moderate correlation between the frequency of DSBs across cell types, which suggests that a considerable proportion of DSB susceptibility of the genome is cell type-specific.

Once we confirmed the accuracy of these models, we asked how the DSB predictions for the three cell types (as they are often used as models for cancer) that we have DSB maps for relate to the patterns of SV breakpoints observed in tumor sequencing studies of squamous cell carcinomas (relevant to NHEK cells), blood cancers (relevant to K562 cells), and breast tumors (relevant to MCF7 cells) in TCGA and ICGC public repositories. The data in these repositories were analyzed as pancancer datasets, gathering all cancer types together, but also as three cancer type subgroups. Similar to what we performed for DSB datasets, the number of tumor SV breakpoints was determined at 50 kb resolution (i.e., in a 50 kb bin, a single DSB was counted if either or both ends of a SV overlapped the region) for each of the TCGA and ICGC SV datasets. Overall, by looking at the ICGC data, we found low-correlations between the number of SV breakpoints and DSB predictions. However, when we restricted our analysis to enriched SV breakpoint (ESB) regions (i.e., 50 kb bins with SV breakpoint counts in the top 5% genome-wide) we found increased correlation with the DSB model predictions [**Fig. 18]**. A significant increase in the correlation with DSB model predictions in NHEK and MCF7 was seen for pancancer, carcinoma, blood, and breast tumor enriched ESBs. In addition, in K562 DSB model predictions, significant increases were observed for all cancer subsets, except blood ESBs [**Fig. 18 A-C]**. Seeing the significant increase in DSB model predictions for carcinoma ESBs, indicates that in these cancer types, DSB susceptibility – as captured by predictive models – possibly shapes the landscape of SVs. However, from TCGA data, none of the subgroups

showed agreement, apart from the blood cancer ESBs, which might be due to low resolution of TCGA that makes it not very suitable for accurate detection of breakpoints. However, focusing on ICGC-annotated SV across all tumor types, overall, significant elevations for ESBs covering all SV classes except insertions were seen by our models [**Fig. 18 D-F**]. In relation to that, as insertions may occur through different mechanisms – for example by transposable element activity rather than by DNA damage and repair processes – they may be less influenced by DSB susceptibility.

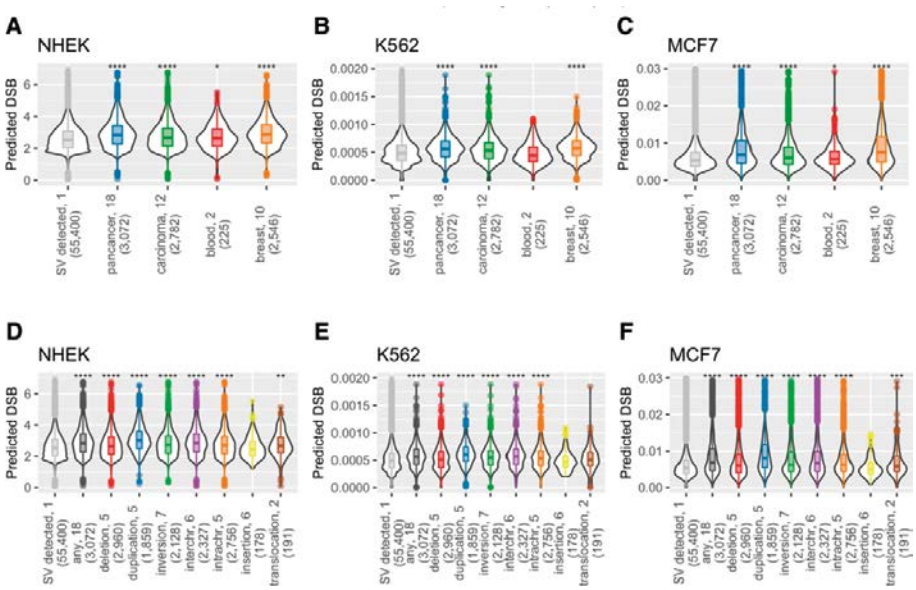

Figure 18. *Violin plots illustrating regions enriched for cancer SV breakpoints (ESBs). (**A-C**) The top 5% SV breakpoint frequencies regions from ICGC database are shown with their predicted DSB values for three cell types (NHEK, K562, MCF7) that were used as models. Cohorts in ICGC are shown as pancancer and divided into three cancer categories: carcinoma, blood, and breast cancers. (**D-F**) The counts of SV breakpoints from ICGC separated by the type of SV, and regions with SV breakpoint frequencies in the top 5% are shown with their predicted DSB values for three cell types. The numbers after the labels in x-axis indicate cutoffs for SV breakpoint counts of the top 5% ESBs, and the numbers in parenthesis show how many 50 kb regions meet the cutoff. Significant higher values in DSB predictions for the ESBs relative to non-ESBs for each category are indicated by stars (Wilcox ranked sum test).*

Next, we aimed to identify hot and cold spots for structural variant breakpoints in tumors. To this end, we developed a new metric, namely the *d*-score, to measure the difference between the SV breakpoints log *p* value in tumors and the predicted DSB log *p* value of a given DSB model in a 50 kb window of the genome. Then in the ICGC tumor cohort, we classified regions of interest based on the *d*-score metric as: i) regions with significantly more SV breakpoints than expected (CancHpredL; cancer high, predicted low or SV hotspots), ii) regions that have lower SV breakpoints than expected (CancLpredH cancer low, predicted high or SV coldspots), iii) regions that have both high SV breakpoint frequencies and high predicted DSB values (cancHpredH), which correspond to genomic regions with remarkably high frequencies of SV on the background of high susceptibility to DSB, and lastly iv) regions with high SV breakpoint frequencies but close to zero predicted DSB rates (cancHpredL2; in principle they are SV hotspots but we found that due to their association with low mappability, they are most likely repetitive, heterochromatic and artifact-enriched regions). We then used circular permutation test to assess the significance of a range of functional annotation enrichments (i.e., consisting of two putative cancer gene sets; 260 genes from the Cancer5000 dataset, and 561 genes from the COSMIC database). In addition, a set of 15,415 super enhancers, common fragile sites, and chromatin states from ENCODE were included in these four classes of regions. We found that, in cancHpredL class of hotspot regions significant enrichments in both gene sets were observed, although not in RefSeq genes, which suggests SVs may underlie the frequent alteration of these genes in cancer. In addition, it turned out that cancHpredL hot spot regions are significantly depleted for active chromatin regions (e.g., active promoters, enhancers and insulators), possibly due to the fact that these regions do not have low predicted DSB. In another class of regions, CancHpredH (high susceptibility region) enrichment was observed for RefSeq genes, in both cancer gene sets, active promoters, strong enhancers, and insulators [**Fig. 19]**. Likewise, the coldspots CancLpredH class occupy gene-rich regions, active promoters and strong enhancers regions, indicating that perhaps purifying selection controls the integrity of some genes and distal regulatory regions.

We finally looked into functional annotation of regions of interest. We examined two classes of regions; the cancHpredL class (ten 50 kb regions with the highest *d*-scores) to uncover genes that might be re-assigned as oncogenic because of having higher SV breakpoints frequency than expected in cancer, and cancLpredH class (ten 50 kb regions with the lowest *d*-scores) that we predicted to be under purifying selection, for signals of potential functionality. To perform this analysis, the NHEK cell DSB model predictions were paired with carcinoma SV breakpoints from ICGC. We found that, in the regions with highest *d*-scores, nine out of ten examined regions overlapped with a gene, and four of them overlapped with

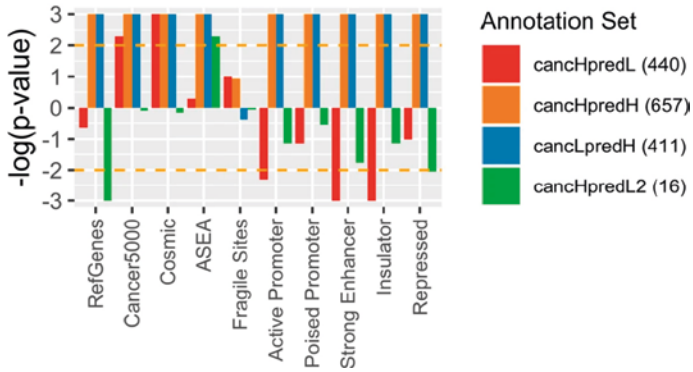

Figure 19. *Regions were categorized into four classes based on d-score (a measure of the deviation of the observed breakpoint frequencies from the predicted or expected DSB frequencies); high (cancHpredL) and low (cancLpredH), cancHpredH (regions with d-scores near zero), and cancHpredL2 (low mappability regions). Using circular permutation, each category was tested for enrichment of various annotations. The yellow dotted line indicates significance ($p < 0.01$). Numbers in parentheses show how many 50 kb regions found in each category, out of 61,903 in total.*

COSMIC genes; *CHEK2*, *CDKNI2A* (known tumor suppressors), *TMPRSS2* and *ERG* (often involved in translocation and in the formation of fusion oncogenes). We also found that two adjacent 50 kb regions of chr17q12 overlapped *GRB7*, which its protein product interacts with a well-known proto-oncogen and *IKZF3*, which is involved in B lymphocyte regulation, differentiation and chromatin remodeling. In addition, a known fragile site, FRA17A, corresponds to this region as well. Regarding regions with the lowest *d*-score, seven out of ten overlapped with a known gene and two oncogenes. For example, the oncogene *CDC27* is a highly conserved gene that interacts with mitotic check points proteins and is thought to be necessary for cell survival. We also found a non-coding RNA (*LOC654342*) in the centromeric region on chr2, which overlaps with H3K27ac, an active enhancer mark, that may be acting as a regulatory component.

These findings show another aspect of BLISS technology as a genome-wide DSB mapping method that can be used to generate a comprehensive list of candidate hot and cold spots regions in different tumor types to further illuminate the molecular basis of genome fragility, and possibly identify novel diagnostic markers or targets for treatments.

2.3.3 CUTeq is a cost-effective method for DNA copy number alterations profiling

Cancer is a complex and heterogeneous disease condition which is strongly associated with genomic instability, such as copy number alterations (CNAs) – duplications and deletions of genomic sequence[110,111]. There are several mechanisms that underlie the generation of somatic CNAs, including cellular repair systems NAHR and NHEJ, which are activated by the presence of DSBs[64]. Conceptually, chromosomal instability like CNAs can promote the production of genetically distinct populations of cells, thus in case of cancer, they can increase intra-tumor heterogeneity (ITH) – i.e., subpopulations of cells within a single tumor can exhibit distinct genomic profiles – that may confer a selective growth advantage such as enhanced cell proliferation, metastatic behavior and chemotherapeutic drug resistance to a certain subpopulation of cells[112]. Thus, the ability to accurately identify and evaluate CNAs and the causative genes are of great clinical interest, as they may represent valuable diagnostic and prognostic biomarkers with implications for cancer treatment decisions. Sometimes, CNAs, which may affect only a small number of cells, can be undetectable, especially if the molecular analysis is performed on a larger mixed pool of normal and variant tumor cells[113]. Understanding the spatial complexity of ITH from a CNAs perspective requires the development of high-throughput methods that can preserve topographical information about the tissue context to capture and analyze multi-regions of a single tumor section. With this goal in mind we aimed to establish a method to construct highly multiplexed DNA libraries for reduced representation genome sequencing of multiple samples in parallel that also enables integrating microscopy and sequencing tools to explore spatial ITH of CNAs in primary tumors, metastases and as well as cell lines that we termed CUTseq. Since the focus of our lab was to develop methods to study genomic instability, during the course of BLISS development in **paper I**, we redesigned the BLISS protocol to exploit it for CNAs profiling in **paper III**. Basically, CUTseq is similar to BLISS, with the difference that DSBs are artificially induced by restriction digestion on purified genomic DNA prior to ligation in solution.

Essentially, CUTseq just needs gDNA that is extracted from any types of samples, ranging from non-fixed, fixed cell lines or even FFPE tissue sections. For cell lines, gDNA can simply be extracted according to desired protocol, and in case of tissue sections, multiple small regions of a single tumor tissue section on a microscopy slide is marked with PinPoint DNA Isolation System (ZymoResearch), which acts as a glue on tissue and can then be peeled off using a small needle and placed into different tubes. Genomic DNA is isolated and subjected to *in vitro* digestion using either a 4-base cutter (NlaIII) or a 6-base cutter (HindIII) that were chosen among a list of commercially available restriction enzymes that generate staggered DSB

ends and are methylation-insensitive, followed by ligation of modified BLISS linker that is compatible with staggered ends. As linker contains sample barcode, each small region collected from a single tissue section can be tagged with different barcodes, which provides an advantage of pooling them together after ligation step and proceed with only single-pooled sample. The sample is then linearly amplified using the power of T7 promoter sequence that is incorporated to linker (similar to BLISS) and finally RNA product is used to generate sequencing library[114] [**Fig. 20**].

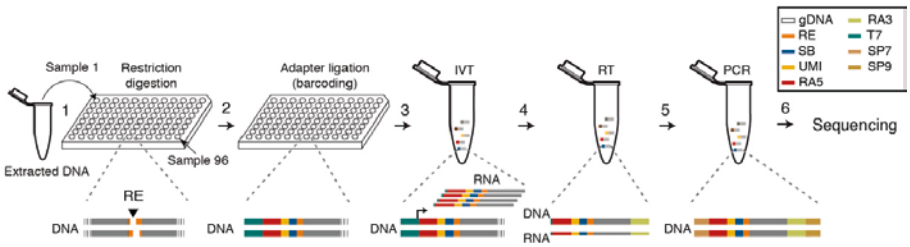

Figure 20. *The workflow of CUTseq for multiplexed library preparation. (1) Extracted and purified gDNA samples are loaded into a multi-well plate (this can be done either manually or with a robotic device) and digested with a restriction enzyme (RE) that generates sticky ends. (2) DNA cut sites are ligated to a double-stranded oligo adapter containing complementary sticky ends, a sample barcode (SB), a unique molecular identifier (UMI), and the T7 promoter (T7). (3) Labeled gDNA is amplified by in vitro transcription (IVT), (4) followed by reverse transcription, (5) amplified by PCR for incorporation of sequencing adapters and (6) eventually is sequenced. RA5, RA3, SP7 and SP9: Illumina's sequencing adapters.*

Multiplexing feature of CUTseq is greatly beneficial for instance in tumor multi-region sequencing application. Currently, in this approach DNA is either extracted from multi regions of the same tumor mass or from multiple tumor sites of the patient, and for each region a single library that is differently indexed is generated and pooled together in a same sequencing run[115]. Tumor multi-region sequencing approach has been used to profile ITH and reconstruct tumor evolution in different cancer types[115], however, one big limitation is the size of the examined regions, which must be large enough to extract sufficient amount of DNA for preparing single sequencing library per selected region, and therefore preventing examining a larger number of small regions. In addition, constructing a single library for every examined region comes with a high cost, which limits the applicability of tumor multi-region sequencing approach as routine cancer diagnostics. Therefore, to overcome the aforementioned limitations, we took advantage of CUTseq features, including its multiplexity and versatility which enable us to construct multiplexed DNA libraries from different sample types for both genome and exome sequencing to call for CNAs and single-nucleotide variants. First, we

demonstrated the reproducibility of CUTseq in five cancer cell lines for CNAs profiling at increasing resolutions ranging from 1 Mb up to 30 kb. In segmented CNA profiles at all resolutions we found high correlation between matched HindIII and NlaIII samples [**Fig. 21 A-B**]. We also found that each cell lines showed a unique pattern of CNAs profiles at different genomic locations regardless of tested enzymes. To confirm CUTseq specificity, we aimed to detect a clinically relevant *ERBB2/HER2* oncogene amplification on chromosome 17, which is reported in BT474 and SKBR3 cells, but not in MCF7 cells[116,117]. Indeed, we observed that CUTseq was able to reproducibly detect cell type-specific amplification of *ERBB2/HER2* locus, using both HindIII and NlaIII restriction enzymes [**Fig. 21 C**].

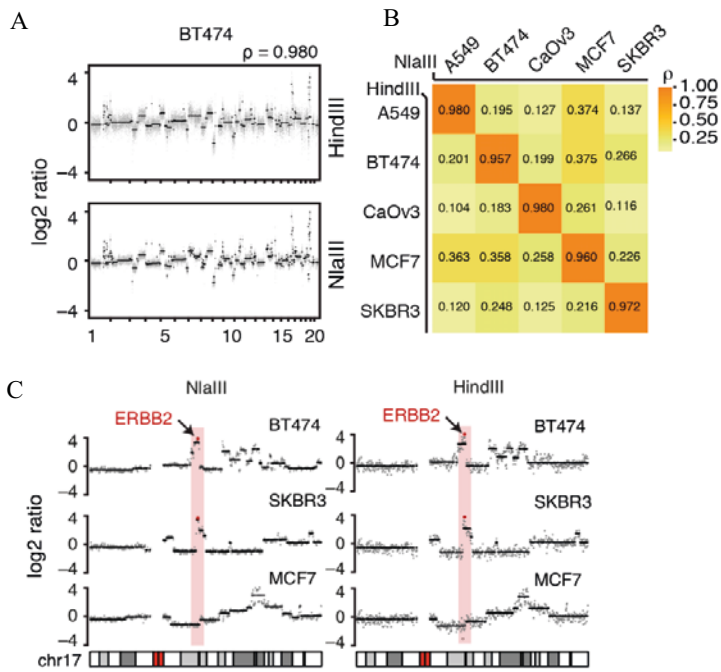

Figure 21. *(**A**) Example of CNA profiles (shown at 100 kb resolution) obtained from BT474 cells using either HindIII or NlaIII. p, Pearson's correlation between the two profiles. (**B**) Correlations (Pearson's p) between the CNA profiles (shown at 100 kb resolution) of five different cancer cell lines gDNA digested with either HindIII (rows) or NlaIII (columns). (**C**) CNA profiles along chr17 (NlaIII, shown at 100 kb resolution) in two HER2-positive (SKBR3 and BT474) and one HER2-negative cell line (MCF7). The locus containing the ERBB2/HER2 gene is highlighted in red. The ideograms of Chr17 are shown at the bottom (red: centromeric regions). In all the CNA profiles, grey dots show individual genomic windows, while black lines represent segmented genomic intervals after circular binary segmentation. The numbers below each box represent chromosomes from chr1 (leftmost) to chr22 (rightmost). In all the cases, TRN refers to the ID of Turin samples.*

Next, we aimed to assess the reproducibly of CUTseq on gDNA retrieved from FFPE tumor specimens. We generated two CUTseq library replicates from five FFPE tumor samples, including two colon adenocarcinomas (COAD) and three melanomas (MELA). We found that CNA profiles between replicates are highly similar at different resolutions [**Fig 22. A**]. In relation to this, we observed that between corresponding replicates, the fraction of genome that was detected by CUTseq as either amplified or deleted was highly correlated [**Fig 22. B**] and also, we reproducibly found that the distribution of amplified/deleted regions became shorter in length as we increased the resolution [**Fig 22. C**].

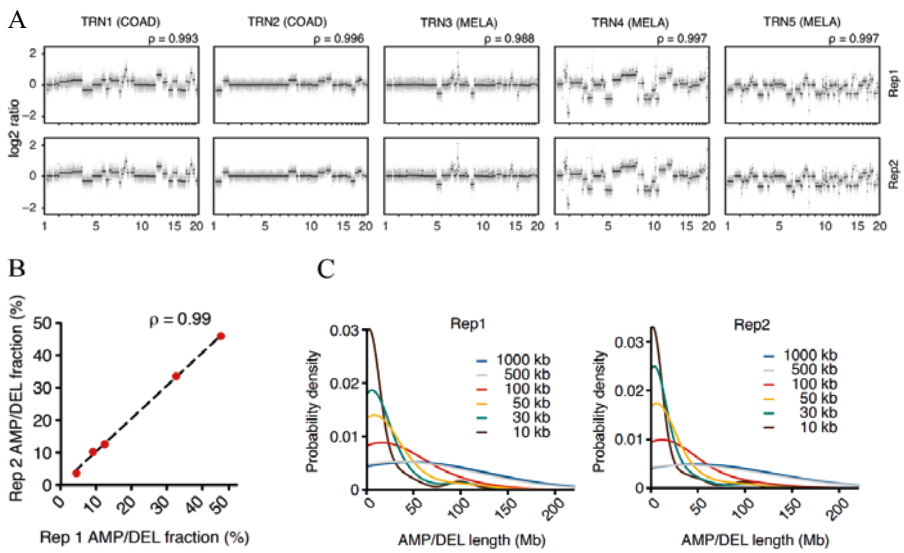

Figure 22. *(**A**) CNA profiles (NlaIII, 100 kb resolution) in five matched replicate (Rep) libraries generated from gDNA extracted from FFPE tumor samples. COAD, colon adenocarcinoma. MELA, melanoma. p, Pearson's correlation between matched replicates. (**B**) Correlation between the fraction of the genome (shown at 100 kb resolution) either amplified or deleted in each of the replicates (Rep) shown in (A). Each dot indicates one pair of replicates. Dashed line: linear regression. (**C**) Distributions of the length of segmented genomic intervals called as amplified (AMP) or deleted (DEL) in the Rep1 and Rep2 samples shown in (A), at different resolutions. In all the CNA profiles, grey dots show individual genomic windows, while black lines represent segmented genomic intervals after circular binary segmentation. The numbers below each box represent chromosomes from chr1 (leftmost) to chr22 (rightmost). In all the cases, TRN refers to the ID of Turin samples.*

In addition, it turned out that overall CNA profiles were reproducible even at 10 kb resolution when we zoomed-in on individual chromosomes and new features including focal amplifications, deletions and more resolved complex patterns of alterations that at lower resolution could not be appreciated were revealed [**Fig. 23**].

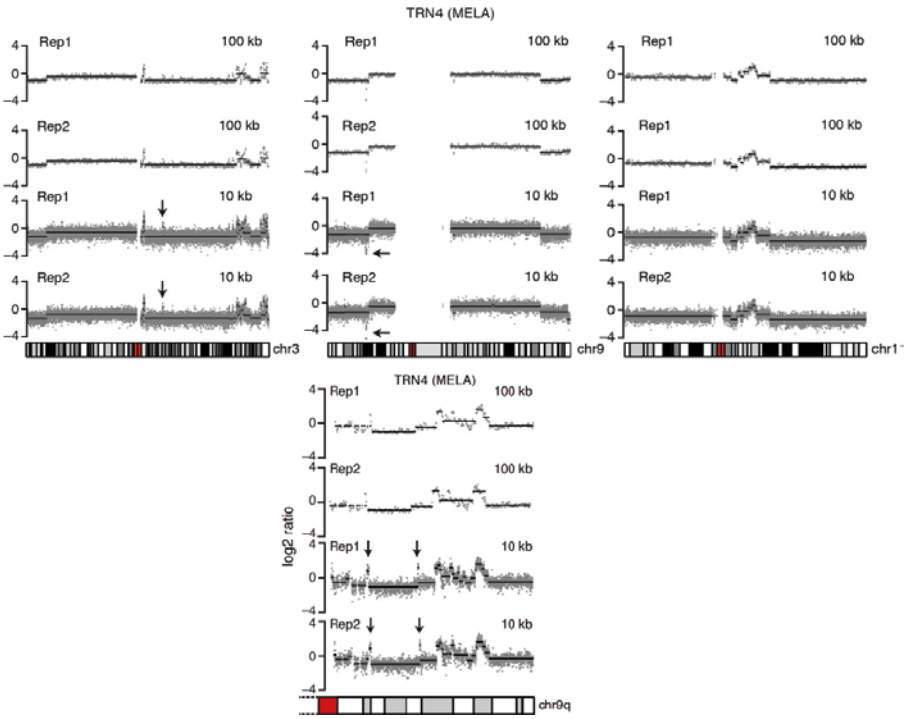

Figure 23. *Examples of CNA profiles of three selected chromosomes, at two different resolutions, for the TRN4 replicates shown in Fig. 22 A. Focal alterations that are detected only at 10 kb resolution, in both replicates are indicated by arrows. Red: centromeric region. In all the CNA profiles, grey dots show individual genomic windows, while black lines represent segmented genomic intervals after circular binary segmentation. In all the cases, TRN refers to the ID of Turin samples.*

We then assessed the sensitivity of CUTseq at sub-nanogram gDNA input, which is not achievable for most of commercially available kits. We extracted gDNA from a single FFPE breast cancer (BRCA) tissue section and created a multiplexed library with decreasing amount of CUTseq barcoded gDNA input (ranging from 1, 0.5, 0.25, 0.125 ng) into the same IVT reaction. In relation to this, different PCR cycles were also tested to exclude biases introduced by PCR cycles. We found that segmented CNA profiles at various resolutions and PCR cycles remained extremely stable with high correlation between each other, even for the 0.125 ng lowest gDNA input [**Fig 24. A-B**]. These findings demonstrated that CUTseq is a sensitive and reproducible method for robust CNAs profiling from picogram amount of gDNA extracted from FFPE specimens, at different resolutions.

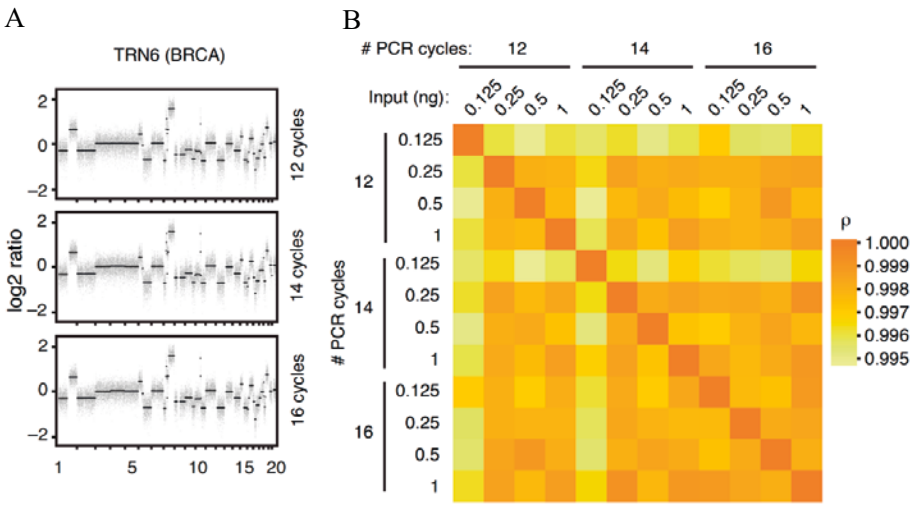

Figure 24. *(**A**) Genome-wide CNA profiles (NlaIII, shown at 100 kb resolution) obtained using 120 pg of gDNA extracted from one FFPE breast cancer (BRCA) sample and three different numbers of PCR cycles performed during the final steps of CUTseq library preparation. (**B**) Correlations (Pearson's ρ) between all the genome-wide CNA profiles (shown at 100 kb resolution) obtained from different amounts of gDNA extracted from the sample shown in (A). In all the CNA profiles, grey dots show individual genomic windows, while black lines represent segmented genomic intervals after circular binary segmentation. The numbers below each box represent chromosomes from chr1 (leftmost) to chr22 (rightmost). In all the cases, TRN refers to the ID of Turin samples.*

Next we aimed to show the compatibility of CUTseq for NGS DNA library preparation. We benchmarked CUTseq with NEBNext® Ultra™ II, which is used as a standard commercial DNA library preparation kit. For this comparison, we used 10 FFPE-derived gDNA samples of four different tumor types, consisting of three breast adenocarcinomas (BRCA), three colon adenocarcinomas (COAD), two gastrointestinal stromal tumors (GIST), and two melanomas (MELA). We generated two libraries per samples, one made by CUTseq and the other using NEBNext® Ultra™ II. We found strong correlations between CUTseq and NEBNext CNA profiles independent of various resolutions [**Fig. 25 A**]. In line with this, we also found high correlation between matched samples in fraction of the genome that was detected as either amplified or deleted [**Fig. 25 B**]. These findings further demonstrated the ability of CUTseq as a sensitive and reliable method for CNA profiling in FFPE samples.

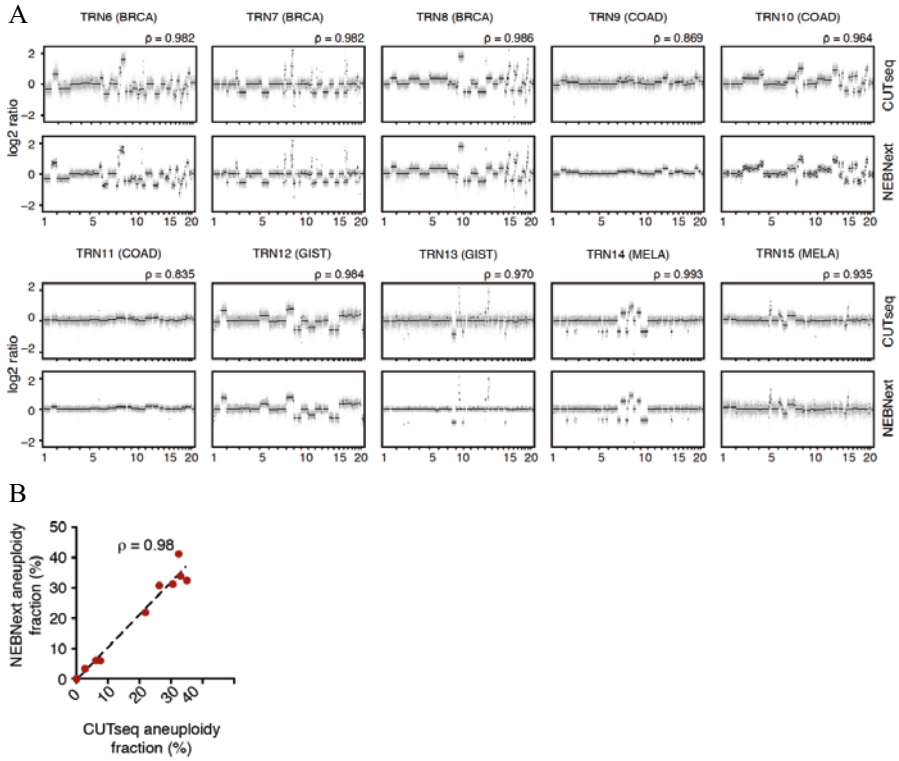

Figure 25. *(**A**) Genome-wide CNA profiles (NlaIII, shown at 100 kb resolution) obtained by CUTseq and NEBNext using gDNA extracted from ten different FPPE tumor sample. BRCA, breast cancer. COAD, colon adenocarcinoma. GIST, gastrointestinal stromal tumor. MELA, melanoma. p, Pearson's correlation between matched CUTseq and NEBNext CNA profiles. (**B**) Correlation between the fraction of the genome (100 kb resolution) either amplified or deleted in each of the paired CUTseq and NEBNext samples shown in (A). Each dot represents one pair of replicates. Dashed line: linear regression. In all the CNA profiles, grey dots show individual genomic windows, while black lines represent segmented genomic intervals after circular binary segmentation. The numbers below each box represent chromosomes from chr1 (leftmost) to chr22 (rightmost). In all the cases, TRN refers to the ID of Turin samples.*

We next aimed to assess the compatibility of CUTseq libraries for exome capture. We generated two replicates CUTseq libraries from gDNA of SKBR3 cells and performed exome capture using SureSelect kit from Agilent Technologies. As a control, two libraries using the same gDNA as input were used to generate two replicates libraries by a commercial kit from Agilent, which then were used for exome capture by SureSelect kit. Single-nucleotide variants (SNVs) calling of CUTseq and Agilent Technology revealed that genomic distribution and type of high-confidence SNVs were very similar between replicates of both methods [**Fig. 26 A**]. In both replicates of CUTseq, 72.3% of all the high-confidence SNVs

identified were detected [**Fig. 26 B**], while between the two methods 37.8% of all the SNVs were shared [**Fig. 26 B**], with CUTseq showing lower mean coverage per SNV, due to its nature as a reduced representation sequencing method [**Fig. 26 C**]. In relation to this, we also obtained similar results using two different FFPE-derived gDNA tumor samples [**Fig. 26 D**]. These results all together demonstrated that CUTseq libraries are compatible with standard exome capture and therefore can be used as a method for reduced representation exome sequencing.

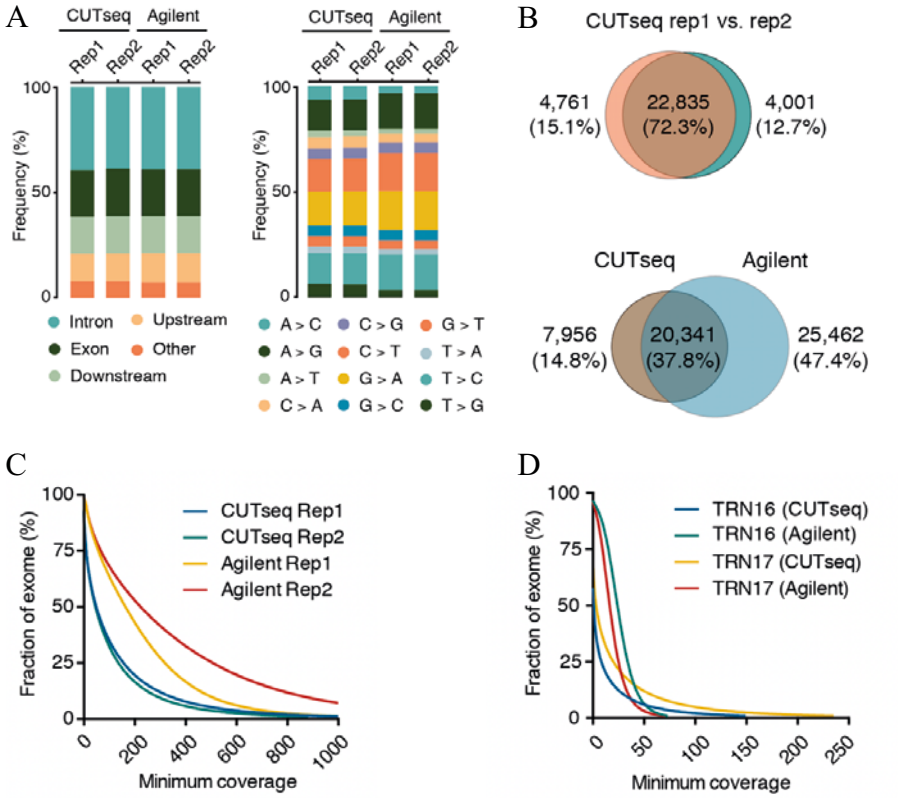

Figure 26. *(**A, left**) Shows all the SNVs called in two replicate (Rep) exome capture experiments from SKBR3 cells-extracted gDNA and using either CUTseq or a commercial kit for preparation of the library (Agilent), in multiple different annotated genomic regions that are partitioned. The legends, up- and downstream indicate 5 kb windows before and after the start of protein-coding genes, respectively. (**A, right**) Same as in (left plot), but different substitution types are shown. (**B, top**) High-confidence SNVs (at least 50 × coverage) overlaps that called in the two CUTseq replicates shown in (A). (**B, bottom**) Overlap between all the high-confidence SNVs identified by CUTseq vs. Agilent, after merging matched replicates shown in (A). The percentages indicate the total number of SNVs in the union of the two sets. (**C**) Exome coverage for the same libraries shown in (A). (**D**) Same as in (c), but represents the libraries prepared from two distinct FFPE breast adenocarcinoma (BRCA) samples.*

To assess the versatility and high multiplexity of CUTseq, we developed a semi-automated workflow based on contactless liquid-dispensing robot, which allows working with nanoliter reaction volumes and enables preparation of ready to sequence libraries in about 8 hours **[Fig. 27 A]**. As a proof-of-principle, 5 ng gDNA of HeLa cells were dispensed in each wells of 96-well plate, digested, differentially ligated to 96 different CUTseq barcodes and then pooled together into one single tube for IVT reaction and library preparation. The library was sequenced shallowly yielding 88 out of 96 replicates (91.7%) with at least 100K usable reads [**Fig. 27 B**]. We also notice that the sequencing error rate was very low (median: 1.62%; interquartile range: 1.58%–1.68%), indicating that even with quick digestion and ligation in nanoliter volume, CUTseq is very precise [**Fig. 27 C**]. CNAs analysis of all the 88 replicates with at least 100K usable reads showed highly similar profiles with strong correlation between each other [**Fig. 27 D**]. In addition, genome fraction that was either amplified or deleted was uniform across replicates [**Fig. 27 E]**. These results indicated that CUTseq is highly multiplexed and cost-effective – total cost is considerably low for preparation of libraries for large number of samples compared to commercial kits – for preparing libraries of large number of samples, in addition to its precision, reproducibility and short turnaround time.

Lastly, since one of the CUTseq feature is its applicability in multi-region tumor sequencing, we set to assess intratumor heterogeneity of CNAs and aimed to profile 35 archival FFPE samples from 14 patients (age of specimens: 9–27 yrs), consisting of primary tumors and one or more matched metastases previously profiled by whole exome sequencing. Each tumor sections were first stained with hematoxylin-eosin for imaging and morphological assessment before extracting gDNA from a large region (labeled as L with the size of ~7 mm^2) that was confirmed by a trained pathologist to contain tumor cells [**Fig. 28 A]**. Each L region gDNA was split into half and labeled as L1 and L2 as technical replicates. In two cases we also extracted gDNA from several smaller regions – labeled as S with the size of ~3 mm^2 [**Fig. 28 A**]. In addition, remaining material after collecting L and S regions where used to extract gDNA from the full tissue sections (labeled as F) [**Fig. 28 A**]. In total, 133 different regions were retrieved, gDNA collected from each region was separately barcoded, pooled into four libraries (i.e., L1, L2, S and F samples) and sequenced enough to obtain at least 200K reads per region to generate CNAs profiles of 100 kb resolution. CUTseq revealed that CNAs profiles of matched L1 and L2 replicates are very similar and aneuploid genome fractions showed high correlation (Pearson's p 0.98) [**Fig. 28 B-C**]. In addition, hierarchical clustering revealed that CNAs profiles of matched L1 and L2 replicates always cluster together from different samples and patients [**Fig. 28 D**], further indicating applicability and reproducibility of CUTseq in different sample types and formats.

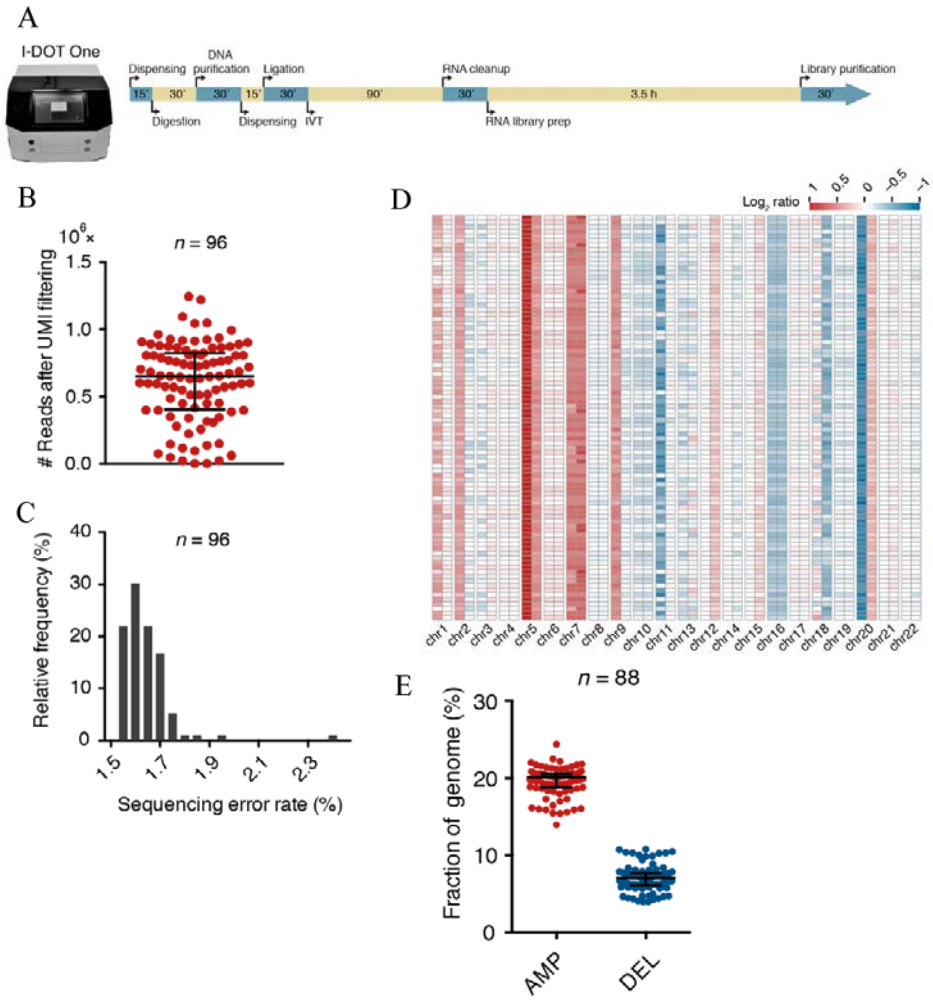

Figure 27. *(**A**) Low-volume non-contact dispensing device, I-DOT One MC, that was used in this study, and timeline for high-throughput CUTseq library preparation are depicted. IVT, in vitro transcription. The total workflow for a single person to prepare 1–2 libraries takes ~8 hours, each containing up to 96 samples. The step for dispensing samples can be either performed manually or using a liquid handling device such as I-DOT One. (**B**) Number of usable reads (after alignment and PCR duplicates removal) per sample (in this example 5 ng gDNA extracted from HeLa cells) in one multiplexed CUTseq library prepared from 96 replicate samples (n) , using I-DOT One. (**C**) The sequencing error rates distribution in the 96 replicates (n) shown in (b). (**D**) Genome-wide CNA profiles (shown at 1Mb resolution, averaged at arm level for visualization) for 88 replicates shown in (b) that yielded at least 300K usable reads. The remaining 8 samples were not included since the number of usable reads was insufficient to perform reliable CNAs calling. (**E**) Fractions of the genome either amplified (AMP) or deleted (AMP) in the 88 replicates (n) shown in (D). Each dot represents one sample. Error bars indicate the median and interquartile range.*

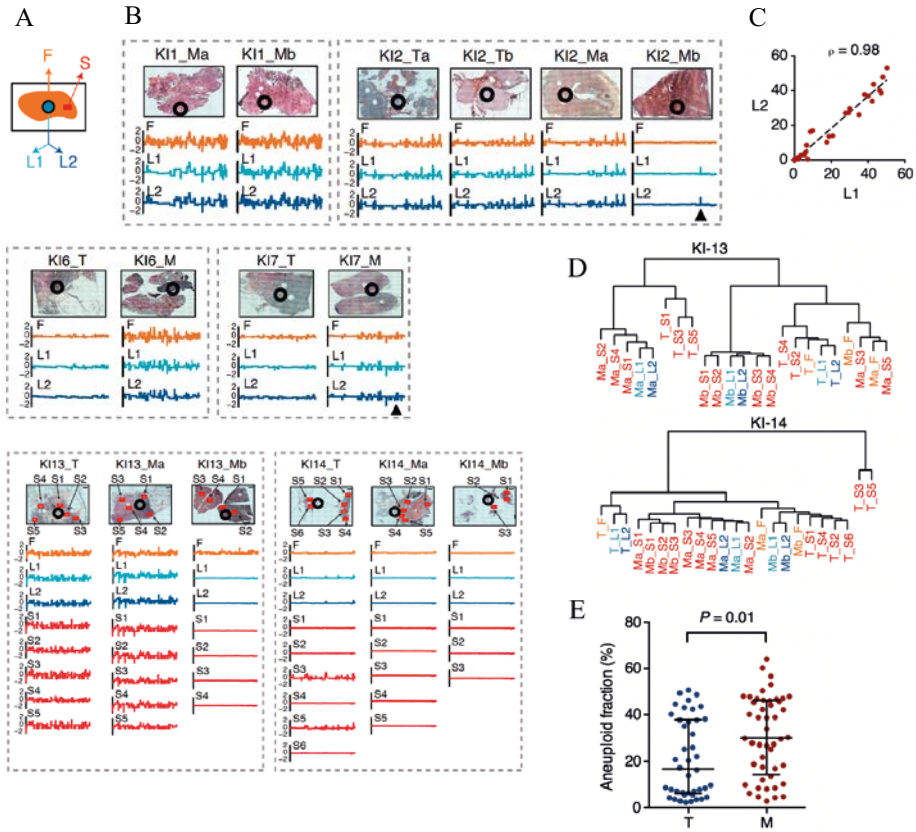

Figure 28. Multi-region DNA copy number profiling in FFPE breast cancer tissue sections. *(**A**) Schematic representation of the tumor-rich regions, within individual FFPE breast cancer tissue sections, from which gDNA was extracted. S, small regions of ~3 mm^2. L, large regions of ~7 mm^2. For each L region, the extracted gDNA was split in two technical replicates, L1 and L2. F, all remaining tissue in the section. To capture the tissue from each region the PinPoint Slide DNA Isolation System™ was used. (**B**) Low-resolution (10X magnification) images of 35 tissue sections from primary (T) and metastatic (M) breast cancers from 14 different patients (note that not all the images shown here), stained with hematoxylin-eosin, and corresponding genome-wide CNA profiles, at 100 kb resolution, for F, L, and S (only for KI patients no. 13 and 14) regions. Black circles show the position, in each section, corresponding to the L region from which L1 and L2 replicates were obtained. Black arrowheads represent amplification of chr14q24 encompassing the RAD51B gene in patient KI2, and of chr17q12 encompassing the HER2 gene patient KI7. In all the CNA profiles, chr1 is on the left and chr22 on the right. (**C**) Pearson's correlation (ρ) between the aneuploid genome fractions across all L1-L2 replicates shown in (B). (**D**) Hierarchical clustering of the CNA profiles of all the F, L, and R regions, for KI patient no. 13 and KI patient no. 14. (**E**) Fractions of the genome either amplified or deleted in the regions with at least 2% of the genome either amplified or deleted (n), separately for primary (T) and metastatic (M) lesions. Each dot shows one region (n). Error bars represent the median and interquartile range. P, Mann-Whitney test, two-tailed.*

Although we found that L regions typically cluster together with corresponding F regions, notably we observed that in patient KI-2 metastasis-b sample [**Fig. 28 B, black arrowhead**] of L region there is a ~900 kb amplification on chr14q24, encompassing the *RAD51B* gene, which was undetected in full section (F). Furthermore, two S regions of primary tumor from patient KI-14 showed numerous CNAs profiles than other S, L and F regions [**Fig. 28 B and D**]. Moreover, from CNAs profiles and hierarchical clustering trees we observed that metastatic regions of the same tumor typically clustered together and apart from the regions of the corresponding primary lesion. Metastatic regions also showed a significantly higher load of amplifications and deletions compared to corresponding primary tumor regions (P-value = 0.006, Mann-Whitney test, two-tailed) [**Fig. 28 D-E,** note that not all the data shown here]. These results are in line with recent finding that breast cancer distant metastases show a different mutational landscape, although phylogenetically related, in comparison to primary tumor, due to accumulation of genomic instability and evolution of tumor[118].

These findings highlighted the power of CUTseq for high resolution multi-region sequencing to detect sub-clonal CNAs which can be undetectable when working with gDNA samples retrieved from larger tissue samples.

Eventually, to find out how many of cancer genes are affected by CUTseq detected CNAs in different tumor regions, we explored the COSMIC[119] database for 712 cancer-associated reported genes, and we found that 241 genes (33.8%) were amplified, and 261 genes (36.6%) were deleted in one or more tumor sites, regions, or patients in our cohort. The top-three genes that found to be amplified were *MYC*, *NDRG1*, *RAD21*, while three-most frequently deleted genes were *KMTA*, *PAFAH1B2*, *POU2AF1* [**Fig. 29 A**]. In addition, performing hierarchical clustering analysis showed that there are at least two major groups; the first one that includes both amplifications and deletions in a large subset of COSMIC genes, and another group harboring predominantly amplifications in a smaller subset of genes, including *MYC*, *ERBB2*, *CCND1*, *MDM2*, *PIK3CA* [**Fig. 29 B**] that are reported to be recurrently affected by CNAs in breast cancers[120].

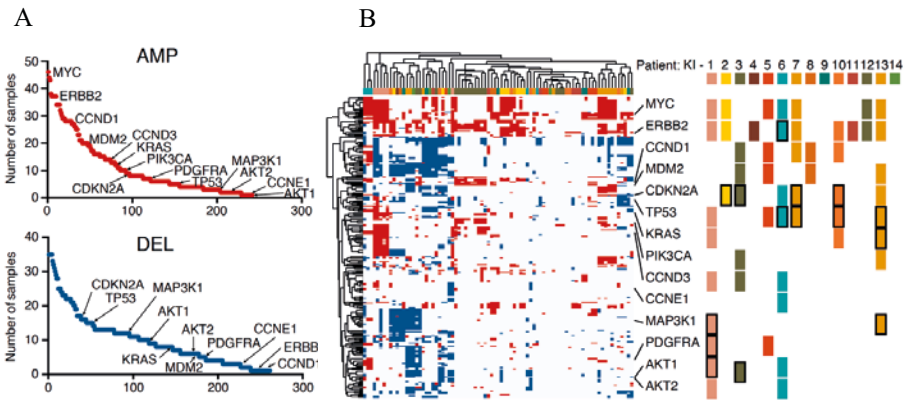

Figure 29. *(**A, top plot**) Shows the ranking of the 712 cancer-associated genes in COSMIC based on the number of samples (133 samples shown in Fig. 28 B), in which they were found amplified (AMP). The gene names shown on the plot refer to a subset of 31 COSMIC genes that were identified to be frequently amplified or deleted in 560 breast cancers. (**A, bottom plot**) Same as (A, top plot), but for genes deleted (DEL) in the samples (133 samples shown in Fig. 28 B). (**B**) Hierarchical clustering of the 712 COSMIC genes (rows) based on their amplification (red) or deletion (blue) status in each of the 133 samples shown in (Fig. 28 B) (columns). Gene names on the right represent 14 out of 31 genes that were previously identified to be either amplified or deleted in breast cancer. For each gene, the rectangles on the right indicate whether it is amplified (no boundary) or deleted (black boundary) in at least one sample (F, L, R) in the KI patient depicted with the same color in the first row above. In all cases shown in the figure, KI refers to the ID of samples from Karolinska Institute.*

Among frequently amplified genes, 7 out of 14 patients (50%) had *MYC* and 9 out of 14 patients (64%) had *ERBB2* amplifications, whereas, 4 out of 14 patients (28.6%) showed deletion of classical tumor-suppressor *TP53* gene among the frequently deleted genes. Furthermore, CUTseq was able to detect HER2 amplification in patient primary tumor samples that were also annotated as HER2-postive by immunohistochemistry (KI-2, 4, 9, 10, 11, 12 samples). Overall, our results demonstrated that CUTseq is a robust, versatile and sensitive method that is applicable in low-input materials from different sample types ranging from cell lines to clinically relevant tumor specimens to investigate CNAs at high spatial resolution and gain further insights into intratumor heterogeneity.

2.4 Discussion and conclusions

This thesis reports the development of two novel genome-wide methods: BLISS for *in situ* profiling the genomic landscape of DSBs (in **Paper I**), and its application to model genome fragility (**Paper II**); and CUTseq for profiling DNA copy number alterations (**Paper III**).

2.4.1 BLISS maps the landscape of DSBs

DSBs play crucial roles in the fundamental physiological processes of the cell (e.g., DNA replication and transcription). However, when not properly repaired, these represent one of the most dangerous types of DNA lesions, and are primary source of chromosomal rearrangements such as CNAs, which are closely related with genomic instability – a hallmark of cancers[91,92]. Dangerous as they are, site-specific DSBs can also be deliberately introduced by programmable nucleases for genome-editing purposes[33]. Therefore, over the past few years there has been a significant interest in determining the DSBs genomic locations and frequencies. Presently, there exist a number of genome-wide technologies that can be applied to investigate various aspects of DSBs, ranging from studying the dynamics of DSBs in DNA damage, to measuring the off-target activity of genome-editing nucleases. For example, as described in the introduction chapter, several genome-wide methods such as GUIDE-seq[88], HTGTS[89,86] and BLESS[76] have been developed for DSB mapping and also specificity evaluation of CRISPR-Cas systems, each with its own strengths and weaknesses. To overcome some of the limitations of these technologies (described further in the introduction chapter), we introduced a novel method into the DSB-mapping technology toolbox. In **Paper I** we presented BLISS, as a versatile, sensitive and quantitative method for direct genome-wide detection of endogenous and induced-DSBs in low-input sample of cells and tissue specimens. At the time this thesis was written, there were a handful of genome-wide DSB mapping methods available, which indicates the importance of accurately identifying these lesions and also reflects how complex and diverse is the field of DSB identification.

With the currently available DSB mapping technologies that reveal different aspects of DSBs, assay choice can be based on specific research question and the types of samples. Those methods based on capture of DSBs through repair pathways, such as GUIDEseq[88], IDLV[87] and HTGTS[90] reported to have high sensitivity, but their applicability beyond model cell lines is limited because: *i*) transfection procedure may induce cellular stress response, which can be a source of DSBs formation; *ii*) applicability of these methods are challenging in transfection-sensitive cells and *in vivo* settings ; *iii*) in the case of GUIDEseq for CRISPR off-target detection purpose, if off-target cleavage is detrimental for the cell (e.g., if the exogenously-transfected DNA insert in an essential gene), it would cause the off-target to be masked due to underrepresentation within the population of cells; i*v*) in addition, studying endogenous DSBs related to replication and transcription may be challenging due to the need for transfection and the time-range required for the introduction of exogenous DNA – for example, when repair and accuracy of endogenous DSBs are high, those repair-dependent DSB tagging methods may miss a substantial amount of DSB events.

Despite the fact that *in vitro* DSB tagging technologies such as BLISS, DSBCapture and BLESS are more versatile, less disruptive, and do not rely on repair pathways due to labeling DSBs after fixation they do however have some limitations as well; for example, they provide only a snapshot of the DSBs that were not repaired at the time of fixation and therefore these methods do not reveal past breakage. In addition, their sensitivity can be affected by chromatin accessibility for blunting DSB ends and adapter ligation.

One of these *in vitro* methods, Digenome-seq[83], tried to overcome the chromatin accessibility problem by purifying the genomic DNA, digesting it with a high-concentration of the nuclease of interest, and subsequently performing whole-genome sequencing. Apart from the fact that this technique can only be used for nuclease off-target detection and not for endogenous DSB mapping (as the input material is purified gDNA), the problem with such a technology is that without any enrichment strategy of the cleaved sites, many sequencing reads derived from unmodified genome are wasted, which results in limited read depth and sensitivity. In relation to this, another subsequent method, CIRCLE-seq was developed to overcome lacking enrichment for cleavage sites in Digenome-seq and eliminate the high background of random reads. This method is based on random shearing of gDNA and creation of intramolecular circularization, followed by removal of uncircularized (linear) DNA before *in vitro* Cas9 cleavage. This strategy reported to have higher sensitivity for off-target detection over Digenome-seq. However, this *in vitro* assay also misses the impacts of chromatin accessibility or other cellular factors on cleavage activity. In addition, it requires a large amount of gDNA for circularization (~25 µg), which based on the availability of cellular source of gDNA can be a limiting factor.

It must be noted that, all of the available DSB profiling technologies are population-based assays, which typically require a high amount of starting material (e.g., a few million cells), and therefore they mostly capture mostly the loci with recurrent DSBs and miss the low-abundance events, which end up as background. However, so far only the BLISS method has proven to be applicable in low-input samples, and is the first method that introduced the quantitative DSB mapping using UMI. In line with this, BLISS implementation for single-cell DSB profiling can shed light on cell-to-cell variation and provides a better understanding of the hetero-geneity of genome fragility in cells (particularly cancer cells) and repair outcomes, albeit, its feasibility and data analysis remain to be explored.

Admittedly, BLISS technology has also its own limitations. Firstly, if the aim is to investigate rare DSB or translocation events, having only few thousands of cells may not be enough to capture those rare events, however, in follow up studies[69,121] we showed that BLISS can also be applied in cell suspension format in millions of cells, in order to not to be limited with only low-input samples. Secondly, quantita-

tive application of BLISS using UMI may be challenging and complicated, since the total number of UMIs in the sample must be counted in order to estimate the total number of breaks. This is especially challenging when working with new samples, a large number of cells and also samples with abundant DSBs, and as such, one might be required to perform a series of sequencing at increasing depth (which is costly) in order to estimate the true number of DSBs in the sample. However, to solve this problem, a few approaches can be considered: i) a cheap and reliable solution we proposed was to use mathematical modelling methods, which can be used to determine the complexity of sequencing libraries based on data from very shallow sequencing runs. This approach can be helpful understand how deeply one should sequence, or predict how beneficial additional sequencing can be[122]. Another solution, could be the integration of a newly developed method for DSB quantification (qDSBseq)[82], which claims to bring an accurate quantification power to any sequencing-based DSB labelling method with BLISS. Its strategy is based on introducing spike-in DSBs by low-frequency cutting restriction enzymes and use them to for the calibration of BLISS data, which ultimately enables calculation of absolute DSB frequencies per cell. However, a side-by-side comparison with BLISS remains to be performed in order to see if it truly improves the quantification power. Thirdly, the application of BLISS in FFPE tissue samples can be challenging, since fixation and long tissue storage procedures often cause nucleic acid degradation and fragmentation of the DNA, which may result in the induction of artificial DSBs[123].

In terms of off-targets of CRISPR-Cas nucleases there are important aspects to consider, especially when moving into gene editing in human: *i*) despite the fact that genome-editing technologies are increasingly becoming more specific with lower-off target rates (e.g., by engineering and development of high-fidelity versions of Cas systems, such as eSpCas9[124] and SpCas9-HF1[125] together with discovery of CRISPR systems with naturally higher specificity feature such as Cas12a)[126], the question is whether the development of increasingly sensitive assays is also needed to detect the very rare off-target events. In line with this, fundamentally, achieving ever-greater sensitivity for off-target detection is limited by the number of input cells used in DSB detection assays. In relation to this, for instance, an off-target event with occurrence probability of one in million is likely missed when we analyze less than 500,000 patient cells edited in a culture dish *ex vivo*, because this off-target break may not occur in such a small number of cells. On the contrary, it is more likely to occur when we analyze one billion cells in an *in vivo* setting. Thus, it is important to note that although available technologies for off-target detection (e.g., *in silico*, *in vitro*/*in vivo*) can identify high risk loci, it will still be crucial to keep in mind certain aspects such as number of cells analyzed in order to extrapolate in vitro results to in vivo; *ii*) as the human genetic variation can alter the landscape of off-target, it can also create unique specific

off-target sites, and since detecting rare off-targets with possible negative outcome is of our interest, in ideal scenario, all gene-therapy recipients should have their complete genome sequenced for *in silico* off-target prediction, in addition to having enough collected genomic DNA for personalized *in vitro* reaction; *iii*) a recent report showed that off-targets are not the only issue of concern in genome-editing, in fact, unintended on-target changes such as large chromosomal deletions, insertions and inversions that extend over many kilobases is another new dilemma[127]. Such unintended chromosomal rearrangements by which crucial genes might end up being altered can potentially lead to serious health consequences. Importantly, these rearrangements might remain undetected because for the on/off-target analysis usually a short-read NGS technology, such as Illumina is used for genotyping CRISPR/Cas-induced mutations. Therefore, long-read sequencing (such as PacBio) and long-range PCR genotyping can be applied to detect these unintended rearrangements[127].

Overall, it is indeed difficult to choose the best assay to capture most of either endogenous or induced-DSBs, however, when resources permit, the ideal strategy would be to combine repair-dependent DSB capture methods (e.g., IDLV, HTGTS, GUIDE-seq) and *in vitro* assays (e.g., BLISS, DSBCapture, BLESS) to generate a complementary picture of DSB events at a certain time point and also over a period of time. However, if the samples to be DSB-profiled are from clinical and precious origins, generally the cell numbers are limited and transfection is not feasible, and therefore *in vitro* based assays compatible with low-input material, such as BLISS, are favorable.

Extended applications of BLISS technology

With the help of available genome-wide DSB mapping technologies we have now gained further insights into the distribution of DSBs along the genome. One of the observations that has been revealed thus far is that transcriptionally active loci and TSS are particularly prone to break. Using BLISS technology, investigating the intricate relationship of DSBs formation, transcription, and chromosome architecture has now been further substantiated by several follow-up studies.

For example, using BLISS, Iannelli et al.[128] intersected DSB maps that were generated in the U2OS-AsiSI system with multi-layered expression profiling (RNA-seq, Bru-seq, CAGE, RNA POLII ChIP-seq), which enabled them to monitor transcriptions in regions where AsiSI cut sites were present. This study revealed that transcription was inhibited around AsiSI-induced DSBs, but not in uncut regions, in an ATM-dependent manner, and that transcriptional repression is progressively weakened by moving away from the vicinity of DSB sites[128]. This transcription repression phenomenon has been attributed to several mechanisms, such as degradation of RNA POL II, condensation of chromatin and recruitment of transcription repressor factor PRC1[129–131].

Recently, Dellino et al.[132] have also applied BLISS to investigate how transcription processes induce DSBs at discrete genomic loci in normal, unperturbed human mammary cells, and address how DSBs are processed and whether they are linked to cancer-associated translocations. They reported that DSBs are accumulated at promoters, 5′ splice sites and active enhancers upon the release of paused RNA POL II. In the absence of canonical DDR, these DSBs are processed by end-joining, and POL II pausing at long genes seemed to be the main predictor and determinant of DSBs.

In another study using BLISS, Gothe et al.[133], studied the interplay between transcriptional activity and the occurrence of *MLL* (Mixed lineage leukemia) chromosome translocations. They reported that TOP2-mediated DSBs were enriched in chromatin loop anchors and associated with a high level of transcription. They observed that transcription is a major contributor of TOP2-associated DSBs, by showing that etoposide-induced DSBs were highly enriched in active genomic regions and POL II- occupied sites. In addition, these DSBs were positively associated with the level of transcription output at promoters and intragenic sites. Furthermore, occurrence of TOP2-induced DSBs substantially decreased in promoters and gene bodies upon inhibition of transcription elongation. They also reported that *MLL* and its fusion partners are highly transcribed, enriched at chromatin loop boundaries, and accumulate TOP2-induced DSBs in a transcription-dependent manner. Thus, the authores proposed that since TOP2-induced DSBs overlap with *MLL* and its fusion partners, they possibly drive the formation of oncogenic *MLL* translocations. Overall, the authors described that interplay between spatial 3D chromosome organization and transcription are major contributors of DSBs at recurrent genomic regions that frequently translocate in cancer.

Recently, in **Paper II**, we also showed that using BLISS-generated data from cell lines, one can construct computational models of DSBs to predict the frequency of expected breakages across the human genome[69]. By using random forest regression models from four DSB datasets, we generated quantitative measures of the relative importance of a variety of epigenetic marks, transcription factor occupancy, gene expression and other features related to DSB susceptibility. Our analysis revealed that the most influential feature in DSB frequency prediction is replication timing across all models. We also demonstrated that comparison of DSB profiles used for the creation of genome fragility models from cell lines with the landscape of rearrangement breakpoints in large tumor cohorts, such as TCGA and ICGC databases, can be utilized to generate a comprehensive list of candidate hot and cold spots fragile loci in different tumor types.

In addition to the studies above, BLISS can also be exploited to unveil new layers of neural regulation and to characterize the molecular basis and functional consequences of the emerging role of so-called activity-induced DSBs[98]. Occurrence of DSBs are particularly deleterious in neurons, since their damage repair capacity is

reduced[134] and as such, accumulation of DSBs might result in cellular senescence or apoptosis of post-mitotic neurons[134], which in turn could account for various neuropathological and age-associated neurodegenerative disorders (for example in Alzheimer's disease mouse model and human brain samples, elevated DNA breaks have been identified)[135]. Interestingly, recent findings demonstrated that neuronal activity specifically induces targeted DSBs involved in transcriptional regulation of a subset of neuronal early response genes (such as *Fos*, *FosB* and *Npas4*), which is mediated by TOP IIb[98]. In addition, it has been shown that the 3D genome is heavily rearranged during stem cells differentiation to favor the transcriptional and epigenetic changes that accompany differentiation[136]. Therefore, using BLISS technology combined with RNA-seq and chromosome conformation capture based methods (e.g., Hi-C) we currently aimed to profile the neural genomic landscape of DSBs and investigate its link to chromatin organization and transcriptional output during neural differentiation, by taking advantage of neuroepithelial-like stem cells derived from induced pluripotent stem cells [137].

In view of the above, the successful integration of data generated from this model system might yield insights into temporal dynamics of DSB formation that affect transcription, chromosome architecture and organization in the nuclear space. In addition, we anticipate that this approach could potentially result in discovery of candidate fragile genes which are suspected to be integral in neurodevelopmental and neurodegenerative disorders. Repeated exposure of these candidate genes to transcription-dependent DSBs as a result of TOP II-activity may result in accumulation of mutations in their regulatory regions, such as promoters. This phenomenon in turn could drive gene expression changes in gene regulatory networks associated with neurological disorders. In order to determine whether fragile promoter regions accumulate mutations, we aim to perform targeted promoter sequencing by making capture probes following our recent iFISH pipeline[138]. We envision that this project could contribute to discover new layers of neural regulation, bringing us closer to understanding the molecular basis of complex neuronal-related diseases.

Another area that a modified-version of BLISS can be applied is for direct detection and visualization of DSBs in single-cell. Currently, a common way for quantitative measurement of DNA damage level and DSBs in single cell is by immunofluorescence against activated DNA repair proteins (e.g., γH2AX and 53BP1), which are detectable as nuclear foci. However, this approach is an indirect measurement of DSB, and also repair proteins may be detected even in the absence of actual DNA damage[139,140]. Alternatively, TUNEL (terminal deoxynucleotidyl transferase dUTP nick end labeling)[141] and COMET[142] assays have been used for DNA damage detection. TUNEL enables exposed DNA ends labeling with the addition of deoxynucleotides (either directly with a fluorescent label or with a chemical label) to the 3'-hydroxyl terminus of DNA breaks, using the enzyme terminal deoxynucleotide transferase[141]. In COMET assay, a suspension of single cells is embedded in an

agarose gel and spread onto a microscope slide. Cells are lysed and under alkaline/neutral conditions the DNA unwinding and electrophoresis result in the migration of broken DNA fragments away from the nucleus, appearing like a 'comet'. The extent of DNA damage can be determined based on the size, shape and distribution of DNA within the comet[142]. However, both methods are prone to artifact and have low-sensitivity, and as such are usually used for detecting massive DNA damages, such as apoptosis. Instead, we propose to design a Y shape-like BLISS adapter (termed Y-BLISS), through which, after *in situ* ligation to DSB ends, the flaps of adapter can be used to hybridize several fluorescently labeled detection oligos – similar to single molecule RNA FISH hybridization strategy. Alternatively, we can use our recent FISH based method (RollFISH)[143] strategy, which exploits the padlock probe on the flab region of the adapter, followed by rolling circle amplification to strengthen the signals, if needed. We foresee that if Y-BLISS performs well, for the first time we can directly and precisely visualize and quantify DSBs at a single cell level, which can be combined with genome-wide single-cell BLISS (once optimized) and immunofluorescence to better understand the cell-to-cell variability in the fragility landscape and DNA repair outcomes.

2.4.2 CUTseq and its application for CNAs profiling

CNAs associated with genomic instability are considered to play critical roles in driving cancer initiation, evolution, drug resistance, and a source of cell-to-cell genetic heterogeneity[115]. Thus, the ability to accurately detect and evaluate CNAs is critical in order to identify their origins, their role in cancer pathogenesis, understanding their implications for patient prognosis, and developing novel therapeutics. In **Paper III**, we presented CUTseq, as a technology for CNA profiling in cell lines and FFPE samples.

We used CUTseq as a reduced genome representation method, which employs frequent cutter restriction enzyme gDNA digestion combined with next-generation sequencing to generate sequence data adjacent to the restriction cut sites. In similarity to other restriction-enzyme based reduced-representation sequencing methods, CUTseq targets a small fraction (1%–5%) of the genome, thus providing advantages over whole-genome sequencing, such as reduced sequencing cost, greater depth of coverage per locus, and through multiplexing large numbers of samples the cost is even further reduced.

Despite the fact that CUTseq allows for accurate CNAs calling at high resolution, it is limited to detect SNVs at any position along the genome. However, we demonstrated that CUTseq can reproducibly detect a considerable fraction of high-confidence SNVs that were also detected by a standard exome capture method. Comparing CUTseq to another similar reduced representation genome sequencing method, RAD-seq[144], which is used in population genetics and ecology, CUTseq

needs only one ligation event to barcode gDNA, while in RAD-seq two ligation events are required. The one-step ligation procedure likely makes the probability of proper ligation for a given gDNA fragment higher than with RAD-seq, although we did not perform a side-by-side comparison of these two technologies. In relation to this, CUTseq is advantageous in the case of working with low-input material and fragmented DNA that are extracted from FFPE tissue sections, whereas RAD-seq typically starts with relatively high genomic DNA input before digestion[144]. Furthermore, as in CUTseq and its high-throughput format that we described in the paper, a single library can be prepared from hundreds of samples pooled together, it offers a streamlined and cost-effective way of analysis of many specimens in population genomics and ecology application as well.

One application for reduced representation exome sequencing is in multi-region tumor sequencing, through which one can compare profiles of CNAs and SNVs from multiple regions in the same tumor, in order to improve the phylogenetic reconstruction of tumor evolution. We demonstrated that CUTseq can be used to assess CNA profiles of multiple-small regions of a single FFPE tissue sections of primary and metastatic breast cancer lesions. and also, the extent of genetic intratumor heterogeneity can be revealed with this approach, while most likely CNA analysis of gDNA extracted from larger tissue regions might otherwise go undetected. Importantly, due to the lower cost of multiplexed CUTseq libraries generated from multiple small regions of FFPE tissue sections, it can be adopted as a routine diagnostics tool for assessing CNAs and genetics intratumor heterogeneity in tissue sections that have been used for pathological tests.

Another application of high-throughput CUTseq outside of tumor sequencing can be CNAs profiling for cell line authentication and genetic screens. For instance, as CRISPR systems can cause unwanted mutations such as large deletions and complex rearrangements[127,145], CUTseq can be used to detect these nuclease-induced CNAs. In addition, high-throughput CUTseq can be greatly beneficial for authentication and monitoring genomic stability of cell lines in public repositories. Although, in this study we only used single-end sequencing and short reads, using a frequent cutter restriction enzyme, or a cocktail of different enzymes combined with paired-end sequencing and longer reads would most likely enable higher exome coverage.

An important aspect worth considering is that many traditional methods, such as comparative genomic hybridization (CGH) and single-nucleotide polymorphism (SNP) arrays use population-based averaging for SNV/CNA identification within pooled samples. But this approach is unable to accurately assess and unveil the cell-to-cell genetic heterogeneity contained within the pooled samples[146]. Although, using CUTseq for multi-regional sequencing we strived to retrieve as small as possible tissue sub-regions, but we remain unable to provide single-cell resolution. In relation to this, single-cell whole-genome sequencing (scWGS) from individual

cells has become the default choice to survey the landscape of SNVs/CNAs for a deeper understanding of the genetic diversity and tumor heterogeneity of each individual cells[147–149]. However, the high cost of sequencing is still a major application drawback of scWGS, albeit with the decreasing cost of sequencing this may not be a problem in the future. Therefore, an important alternative would be to implement reduced-representation technologies, such as CUTseq as a cost-efficient choice in single-cells for CNAs profiling. In addition, as spatial context is typically lost in single-cell sequencing approaches, CUTseq combined with laser-capture microdissection (LCM) can potentially be a powerful tool to identify CNAs in specific histological tumor subtypes.

In summary, this thesis highlighted the development and applications of novel genome-wide technologies to profile the landscapes of DSBs and CNAs. We envision that future technology development with integrative approaches – for example combination of DSB detection assays (e.g., BLISS) with, for instance, methods to assess genetic alterations (e.g., CUTseq) and concurrent local key epigenetic features or 3D genomic neighborhood and RNA expression – ideally simultaneously in the same sample and preserving it for downstream microscopy analysis (although seems challenging and very difficult to implement) can broaden our understanding of genomic instability and how their landscapes are shaped and or shape the underlying transcriptome, repair choice, epigenome, and 3D genome architecture.

3 ACKNOWLEDGEMENTS

This is indeed a very important part of my thesis and firstly I would like to thank all of you who have read, corrected and proofread this thesis.

Many people contributed to this work and have been on my side during the past few years in all aspects of my life, from lab and beyond and I would like to take this opportunity to express my sincere gratitude to everyone who have supported me. Without you this thesis would not have been possible.

I would like to thank my main supervisor, **Nicola**, for giving me an opportunity to work in your group and guiding me over the last several years. I have learnt a lot in your lab and grew both personally and professionally. I admire your method-development mindset and wish you all the best.

My co-supervisors, **Theo** and **Pan**, thanks for your time and advises along this journey.

Magda, I worked very close with you (there was no alternative :)), thanks for all the advises you provided ("there is room for improvement", saying it every group meeting, really!!! :)), and also giving me the opportunity to interact with you and your very nice lab members to gain more knowledge on microscopy and beyond my PhD project.

I would like to thank my entire dissertation committee members, **Prof. Ulf Landegren**, **Prof, Mattias Mannervik** and **Prof. Richard Rosenquist Brandell**, and my opponent **Dr. Vicent Pelechano** for finding time to read my thesis and being here to discuss with me today.

I also would like to thank my half-time committee members, **Prof. Thomas Helleday**, **Dr. Anita Göndör** and **Dr. Jordi Carreras Puigvert** for the time they put to read my review and the fruitful discussion and feedbacks on my seminar.

A big thanks to **Prof. Adnan Achour** for being my mentor during these years and whose advice and support has been instrumental for me.

I owe a particular thanks to **Prof. Mats Nilsson**, and his team for introducing me to the world of method development field, which was my inspiration to continue along this path. Your knowledge along the constant smile and calmness is so motivational and wonderful.

To all my **co-authors** and **collaborators**, especially to: **Prof. Feng Zhang** and **Dr. Winston Yan** at Broad Institute for sharing their CRISPR expertise, and working hard together to publish the BLISS story. Thanks for the opportunity of visiting

your amazing lab and all the advises. **Dr. Ivan Dellino** and **Rossana Piccioniat** at IEO, Italy, for inviting me to your lab. Prof. Mats Nilsson's group**, Chenglin** and **Marco**, thanks for all your works and helps for our RollFISH paper. **Prof. Vassilis Gorgoulis**, **Dr. Athanassios Kotsinasat**, and **Christos Zampetidis** at University of Athens, thanks for your hard work, frequent skypes and curiosities on our oncogene-induced DSB story. I hope we can publish this work soon. Thanks to **Prof. Colin Semple** and **Dr. Tracy Ballinger** for amazing work on our genome biology paper. Thanks to all the other collaborators and visiting researchers with whom I directly or indirectly learned and shared many things.

I have had the privilege to have amazing friends and colleagues at SciLifeLab and I would like to thank all of them, and I hope that I do not forgot anyone, but if I do, please forgive me.

Especial thanks go to my friends in **BiCro**, **Simon**, **Oscar**, **Jiri** labs:

Quimitto, (que tal cabronetaa = you know that its definition for me is, I like you :)). Thanks a lot for all the memories and laughter inside and outside of the lab (e.g., skiing, climbing, etc. etc. too much activities out of my KI boarder that I couldn't keep up with you :)). Thank you for never picking up the phone whenever I call you:). Thanks for all the Swedish Fikas (I mean butter and bread :)), and Christmas dinner with **Anna**, last year. Thanks for helping me with correction of my thesis, R coding, brainstorming and science discussion. Thank you for all the emotional supports. There are many more things that I have a hard time to remember now! **Tomechko**, we both were the first lab members and you didn't like me initially, until you fell in love :). Your non-stop smile (unless experiments fail :)), contagious laughter, science ideas, business mindset, and luxury style is amazing. Thank you for being Tomechko, for all the memories and laughter inside and outside the lab (e.g, Barry's :)). Thank you for all your helps in the lab during these years. Thanks for introducing us to fashion world (especially, Louie, you know what I'm talking about, Louis Vuitton), golf, diamond, art, a nice Polish word starting with K and in overall how to have a different quality of life :). Sorry that we bothered you to take care of Mahi, for sure he remembers you forever. Be prepared for baby-sitting Elsa :) and thank you for all the emotional supports. Good luck with your PhD and everything else. **Michi joon**, Allora, thanks for always being around and kind for helping me in and outside of the lab. Thank you for being a real checker ("let me check" :)), commenting on the thesis and scanning it for errors. It was really nice working with you for the RollFISH and CUTseq papers. Thanks for all the memories and laughter, and organizing our KI football team (I was not lucky to get the cup with you during these years, though second place is not bad :)). There are many more things that I have a hard time to remember now! Good luck with your PhD and everything else. **Madonna**, thanks for non-stop laughter that drives some people crazy :), singing Disney songs and lullaby in the lab.

Thanks for constantly being high :), your positive energy really makes the lab too energized that someone should unplug you :). Thanks for all the memories and laughter. Stay fashionable, keep the award-winning Miss. BiCro title and try to finally go to skarpnäck with Tomek :). Good luck with everything. **Massa** joon, Emma, Ann and Meg, Chōshi wa dōdesu ka? My Japanese friend and your lovey family. Thanks for the memories and laughter during your stay in Sweden. You are an amazing person and scientist. Thanks to you and Emma for introducing me to yummy Japanese food. It was so fun to play PS4 with a person coming from PS4 manufacturer :). I'm still waiting for PS5 since ever :). I really miss you Massa and hope to see you soon. Good luck with everything. **Eleni joon**, ti káneis? Thanks for being nice all the time and have constant smile. Thank you for introducing me to TDA protocol and it was great working with you on the iFISH paper. Thanks for all the memories, laughter and supports during these years. Enjoy the parenthood together with Vassilis and beautiful Alice. Looking forward to see you soon. **Britta**, stay goed :), thank you for all your helps in the lab, emotional supports, and commenting and reviewing my thesis. Good luck with the mazing beautiful thing that is happening in your life and I wish you and Lisa all the best. **GG_G** :), thank you very much for all the time being helpful and supportive. Thanks for being patient with me to start learning programing, despite the fact that you had your "don't disturb me" flag up most of the time, but always welcomed me with any questions :). Thanks for all the memories, laughter, Limoncello :), chocolate salami, pizzas and pasta discussions (at the end what pasta was what?! :)). Good luck with your PhD and everything else. **Ana**, my desk and bench neighbor friend, I'm sure you make FRET probes to not be frete :). Thanks for all the small chats about history and culture, and always being kind in the lab and helpful. It was great working with you for iFISH paper. I did my best to keep your bench clean, but sorry if sometimes it wasn't the case :). Good luck with your PhD and everything else. **Xiaolu**, stay Hăo :), thanks for the chats in the lab and discussing the protocols. It was great working with you for the CUTseq paper. Thank you for helping me to get parking permissions :). I wish you all the best with Bingnan and Han. **Xinge**, thanks for being kind and nice all the time, chats in the lab, playing violin in the office :) and your supports. Good luck in your new carrier and everything else with Chen. I would like to thank our computational scientists, **Silvano**, Allora, thanks for always finding times to answer and explain my questions behind the analysis. Thank you for generating 225 BED files for me (together with Fede) :). I was really fed up with your motivational speech (keep up the good work :)). I never felt shy to come to your office and ask about the analysis because of your openness and kindness, thank you. **Fede**, thank you for having your office's door open all the time and answering my questions about the analysis. Thanks for always being kind and patient to explain how I can analysis the data on our neuro-DSB project. **Erik**, thank you for all your helps and guidance for image analysis, and being open to our never-ending questions on analysis. **Marcin**, jak się masz :)?

thanks for helping me for a lot of troubleshooting during your first visit to our lab :). Thank you for all the memories and laughter during the past few years and good luck with everything. **Mr. Roberto**, mannaggia :), thank you for letting me work with you on neuro-DSB project and all the long and tiresome cells preparation. Thanks for commenting on my thesis and your motivations. I wish you all the best on your PhD studies. **Solrun**, thank you for using your nice reply for everything and everyday "it's FINE". Keep up being fine and thanks for always being nice. Good luck with everything. **Ning**, thank you for the discussions on protocol optimization and being nice and helpful in the lab. Good luck with everything. **Carla**, thanks for the small chats and always being nice and smiley. I wish you all the best. **Su**, thank you for taking care of the microscope in the lab and using your own personal scope (camera) to take care of personal photoshoots :). Thanks for always being nice and helpful, and please stop traveling that frequent and posting nice pictures to make us depressed :). **Luuk**, you luuk great :), congrats on starting your PhD studies and thanks for being nice and slimily. I wish you all the best on your journey. Enjoy your lab chore as a TIMER :) **Emma**, thank you very much for helping with my thesis proofreading, organizing our orders and always look for us to ask; where is the packing slip? :). **Merula**, thanks for your everyday smiles and chats about babysitting tips :). Good luck with your master thesis. **Gustaw**, thanks for the small chats and your everyday smiles. Enjoy the new apartment :) and good luck with everything. **Theresa**, you were my first student and thanks for your helps and nice words. Good luck with everything. Also, a big thank you to everyone else in BiCro lab (past and present).

Simon, thanks for always keeping your office's door open, even for me :). Thanks for your kindness and smiles. I will remember all the possible coffee making methods you tried in the kitchen, but I don't know at the end which one you're using now :). Thank you for sharing all the instruments of your lab with us. I could not wish for a better adjacent neighbor lab. I would also like to thank you for accepting to be my defense chairperson. **Birthe**, thanks for small chats here and there.

I was really lucky to have such nice friends in Simon's lab. Thank you all for your helps and supports inside and outside of the lab. **Anna-maria**, ti káneis? We started about the same time, thanks for the memories and laughter, and sometimes complains :). Thank you for the helps and supports during these years. **Jing**, it was great collaborating with you and I hope that it was a good one! Thanks for being nice and helpful all the time, sharing reagents with me and going to MAX to buy burgers while I was stuck with experiments. Good luck with your PhD. **Angelo**, thank you for being so nice and calm all the time, even when you're tirelessly doing experiments. Thanks for the chats inside and outside of the lab, your helps, supports, encouragements ("we can do it" :)) and listening to my words. Good luck with your PhD studies. **Rozina**, thanks for being so nice, kind and always willing

to help out in any aspects from lab and beyond. Thanks for preparing samples for our collaboration back in 2016 when you were in Pan's lab. Thank you for all your supports and advises for both work and life matters. It was great sharing all the baby related stories together :). **Banu**, thanks for being so nice, kind and helpful during these years. Wish you best of luck for your PhD studies. **Rui**, thanks for always being nice and helpful. Thanks for lending your amazing camera to me several times, although I can not take those wonderful pictures that you are able to capture. **Lorenzo**, thanks for being kind and also advising me on baby-related matters. **Philip**, thanks a lot for being open to help out and discuss experiments design and provide some alternative solutions. Thanks for the chats in the lab, especially late evenings and holidays. Thank you for sharing your knowledge in IF with me. **Dörte**, thanks for the small chats around the lab and lunch table.

Big thanks to the whole Thoms, Oscar and Jiri lab members. **Thomas**, I admire your encouraging attitude for letting researches to engage in collaboration, talk and communicate to each other for a better, more efficient and breakthrough science discoveries. Thank you for letting me use your lab's instruments and reagents during the first year of my PhD. **Sabina** and **Flor**, thank you for helping and guiding me for administrative tasks. **Oliver** and **Nina**, thank you for helping me for IF and microscopy questions. **Oscar** and **Jiri**, I was lucky and privileged to have two well-known scientists in the field of DNA damage as our neighbor labs, so I could learn more about the field. Thanks for gathering such nice people in your team as well. **Oscar**, the way you present and communicate science with that sense of humor flavor is amazing and easy to understand. Thanks for always being open to talk and discuss science. **Jordi**, thank you very much for being truly nice, understanding and supportive, especially on my computer crisis. **Ann-Sofie**, thanks a lot for all your helps and supports particularly during my computer crisis. Thanks for always being so nice and kind, and all the chats in the cell lab and during lunch time. **Alba**, your laughter is distantly contagious :) and almost whenever you laugh, I start laughing. Keep up the laughter and spread the happiness. Thanks for adopting me as your honoree lab member :). Good luck with your PhD. **Pele**, thanks a lot for coordinating everything on our floor and all your helps for anything that we had no clue what to do about! **Louise** and **Katie**, thanks for joining our football team and making fun memories. **Valeria, Maria**, and **Dani**, thank you for being nice and chats here and there. **Jaime**, que tal? Plamma o…:) and **Dimitris** (Tikan? Hey yo :)), thanks for being so nice, funny, helpful in the lab and for discussing science all the time. **Asimina**, thank you for small chats in the cell lab. Good luck with your PhD.

To my KTH friends and colleagues on Alfa 4 floor, **Aman**, chetori? :) Thanks for always being smiley, all the laughter and the chats in the kitchen. **Amin**, **Sharath** and **Tharagan**, thanks for your helps and chats in the lab and kitchen.

I also would like to thank the other wonderful people that surrounded me at SciLifeLab, especially those people that helped me with their expertise, instruments and reagents sharing during my PhD studies: **Annelie**, one of the best lab managers that one could wish for. Thank you for always being energetic and smiley. Thanks for letting me and our other lab members use your lab's instruments and also occasional reagents lending :). Thanks for small/long chats inside and outside of the lab. I at least owe you a few bottles of Coca-Cola :). **Maja**, thanks a lot for all the helps with Bioanalyzer, being always nice, smiley and for small chats in the lab. Good luck with you PhD. **Stefania**, thanks for pushing the process of me getting the Covaris access back in 2015. Thank you for small chats and being nice all the time. Good luck with setting up your lab. **Afshin**, thanks for the chats and carrier advises. **Hooman**, thanks for the chats by the Bioanalyzer and good luck with everything. **Mahya** and **Mohammad**, thanks for sharing parenthood experiences. Thanks to single-cell facility members, **Karolina**, **Marcela** and **Simone** for letting me use their instruments.

Thanks to Mats Nilsson's group in Stockholm and Uppsala for sharing their knowledge and being always nice and helpful to me. Thanks to, **Di**, **Chenglin**, **Elin**, **David, Malte, Ivan, Tomasz**, **David**, **Annika**, **Anna** and **Flor**.

Finally, my beloved family; my wonderful wife, **Nana**, thanks for your love and supports during the past several years. Thank you for celebrating me the ups, coaching and encouraging me during the downs, and thanks for your energy, patience, and above all, your love. I love you and this journey was not possible without you. My adorable daughter **Elsa**, thank you for bringing more joy and laughter to our life, daddy loves you forever, no matter what :). **Mom** and **dad**, thanks for everything you have done for me, for always teaching me to believe in myself and supporting me to whatever path that I decided to take. To my twin brothers **Ali**, **Masih**, and my eldest brother **Shahram** for all their supports and love. Love you all.

Wow, the acknowledgment section became quite long! :)

4 REFERENCES

1. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071–1078 (2009).

2. Lindahl, T. & Barnes, D. E. Repair of endogenous DNA damage. *Cold Spring Harb. Symp. Quant. Biol.* **65**, 127–133 (2000).

3. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).

4. Nambiar, M. & Raghavan, S. C. How does DNA break during chromosomal translocations? *Nucleic Acids Res.* **39**, 5813–5825 (2011).

5. Bacolla, A., Tainer, J. A., Vasquez, K. M. & Cooper, D. N. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res.* (2016).

6. Raynaud, F., Mina, M., Tavernari, D. & Ciriello, G. Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. *PLoS Genet.* **14**, 1–18 (2018).

7. Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644–656 (2017).

8. Salehi, F., Behboudi, H., Kavoosi, G. & Ardestani, S. K. Oxidative DNA damage induced by ROS-modulating agents with the ability to target DNA: A comparison of the biological characteristics of citrus pectin and apple pectin. *Sci. Rep.* **8**, 1–16 (2018).

9. Maya-Mendoza, A. *et al.* Myc and Ras oncogenes engage different energy metabolism programs and evoke distinct patterns of oxidative and DNA replication stress. *Mol. Oncol.* **9**, 601–616 (2015).

10. Karanjawala, Z. E., Murphy, N., Hinton, D. R., Hsieh, C. L. & Lieber, M. R. Oxygen metabolism causes chromosome breaks and is associated with the neuronal apoptosis observed in DNA double-strand break repair mutants. *Curr. Biol.* **12**, 397–402 (2002).

11. Shokolenko, I., Venediktova, N., Bochkareva, A., Wilson, G. I. & Alexeyev, M. F. Oxidative stress induces degradation of mitochondrial DNA. *Nucleic Acids Res.* **37**, 2539–2548 (2009).

12. Madsen, P. M. *et al.* Mitochondrial DNA double-strand breaks in oligodendrocytes cause demyelination, axonal injury, and CNS inflammation. *J. Neurosci.* **37**, 10185–10199 (2017).

13. Hasham, M. G. *et al.* Activation-Induced Cytidine Deaminase-Initiated Off-Target DNA Breaks Are Detected and Resolved during S Phase. *J. Immunol.* **189**, 2374–2382 (2012).

14. Panchakshari, R. A. *et al.* DNA double-strand break response factors influence end-joining features of IgH class switch and general translocation junctions. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 762–767 (2018).

15. Vuong, B., Nicolas, L., Cols, M., Choi, J. E. & Chaudhuri, J. Generating and repairing genetically programmed DNA breaks during immunoglobulin class switch recombination [version 1; referees: 2 approved]. *F1000Research* **7**, 1–14 (2018).

16. Maas, C. *et al.* The DNA Damage Response Regulates RAG1/2 Expression in Pre − B Cells through ATM-FOXO1 Signaling. (2019). doi:10.4049/jimmunol.1501989

17. Bhattacharyya, T. *et al.* Prdm9 and Meiotic Cohesin Proteins Cooperatively Promote DNA Double-Strand Break Formation in Mammalian Spermatocytes. *Curr. Biol.* **29**, 1002-1018.e7 (2019).

18. Murakami, H. & Keeney, S. Regulating the formation of DNA double-strand breaks in meiosis. 286–292 (2008). doi:10.1101/gad.1642308. Schizosaccharomyces

19. Mazouzi, A., Velimezi, G. & Loizou, J. I. DNA replication stress: Causes, resolution and disease. *Exp. Cell Res.* **329**, 85–93 (2014).

20. Gelot, C., Magdalou, I. & Lopez, B. S. Replication stress in mammalian cells and its consequences for mitosis. Genes 6, 267–298 (2015).

21. Helmrich, A., Ballarino, M. & Tora, L. Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Mol. Cell* **44**, 966–977 (2011).

22. Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat. Cell Biol.* **16**, 2–9 (2014).

23. Aparicio, T., Baer, R. & Gautier, J. DNA double-strand break repair pathway choice and cancer. *DNA Repair (Amst).* **19**, 169–175 (2014).

24. Puc, J., Aggarwal, A. K. & Rosenfeld, M. G. Physiological functions of programmed DNA breaks in signal-induced transcription. *Nat. Rev. Mol. Cell Biol.* **18**, 471–476 (2017).

25. Helmrich, A., Ballarino, M., Nudler, E. & Tora, L. Transcription-replication encounters, consequences and genomic instability. *Nat. Struct. Mol. Biol.* **20**, 412–418 (2013).

26. Sollier, J. & Cimprich, K. A. Breaking bad: R-loops and genome integrity. *Trends Cell Biol.* **25**, 514–522 (2015).

27. Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T. & Cimprich, K. A. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* **170**, 774-786.e19 (2017).

28. De Magis, A. *et al.* DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 816–825 (2019).

29. Pommier, Y., Sun, Y., Huang, S. Y. N. & Nitiss, J. L. Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. *Nat. Rev. Mol. Cell Biol.* **17**, 703–721 (2016).

30. Vitelli, V. *et al.* Recent Advancements in DNA Damage–Transcription Crosstalk and High-Resolution Mapping of DNA Breaks. *Annu. Rev. Genomics Hum. Genet.* **18**, 87–113 (2017).

31. Barlow, J. H. *et al.* Identification of early replicating fragile sites that contribute to genome instability. *Cell* **152**, 620–632 (2013).

32. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182 (2005).

33. Komor, A. C., Badran, A. H. & Liu, D. R. CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell* **168**, 20–36 (2017).

34. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).

35. Jinek, M. *et al.* A Programmable Dual-RNA – Guided. **337**, 816–822 (2012).

36. O'Geen, H., Yu, A. S. & Segal, D. J. How specific is CRISPR/Cas9 really? *Curr. Opin. Chem. Biol.* **29**, 72–78 (2015).

37. Tsai, S. Q. & Joung, J. K. Defining and improving the genome-wide specificities of CRISPR–Cas9 nucleases. *Nat. Rev. Genet.* **17**, 300–312 (2016).

38. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).

39. Daub, H. DNA Damage Response: Multilevel Proteomics Gains Momentum. *Mol. Cell* **46**, 113–114 (2012).

40. Turinetto, V. & Giachino, C. Survey and summary multiple facets of histone variant H2AX: A DNA double-strand-break marker with several biological functions. *Nucleic Acids Res.* **43**, 2489–2498 (2015).

41. Ambrosio, S. *et al.* Cell cycle-dependent resolution of DNA double-strand breaks. *Oncotarget* **7**, 4949–4960 (2016).

42. Maréchal, A. & Zou, L. DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb. Perspect. Biol.* **5**, 1–18 (2013).

43. Lavin, M. F. Ataxia-telangiectasia: from a rare disorder to a paradigm for cell signalling and cancer. *Nat. Rev. Mol. Cell Biol.* **9**, 759–769 (2008).

44. Murga, M. *et al.* A mouse model of ATR-Seckel shows embryonic replicative stress and accelerated aging. *Nat. Genet.* **41**, 891–898 (2009).

45. Huertas, P. DNA resection in eukaryotes: deciding how to fix the break. *Nat. Struct. Mol. Biol.* **17**, 11–16 (2010).

46. Burma, S., Chen, B. P. C. & Chen, D. J. Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. *DNA Repair (Amst).* **5**, 1042–1048 (2006).

47. Malu, S., Malshetty, V., Francis, D. & Cortes, P. Role of non-homologous end joining in V(D)J recombination. *Immunol. Res.* **54**, 233–246 (2012).

48. Downs, J. A. & Jackson, S. P. A means to a DNA end: the many roles of Ku. *Nat. Rev. Mol. Cell Biol.* **5**, 367–378 (2004).

49. Hammel, M. *et al.* Ku and DNA-dependent protein kinase dynamic conformations and assembly regulate DNA binding and the initial non-homologous end joining complex. *J. Biol. Chem.* **285**, 1414–1423 (2010).

50. Gottlieb, T. M. & Jackson, P. The DNA-Dependent Protein Ki for DNA Ends and AssocWbn with Ku Ant. **72**, 131–142 (1993).

51. Yoo, S. & Dynan, W. S. Geometry of a complex formed by double strand break repair proteins at a single DNA end: recruitment of DNA-PKcs induces inward translocation of Ku protein. *Nucleic Acids Res.* **27**, 4679–86 (1999).

52. Paull, T. T. *et al.* A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage. *Curr. Biol.* **10**, 886–895 (2000).

53. Neal, J. A. & Meek, K. Choosing the right path: Does DNA-PK help make the decision? *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **711**, 73–86 (2011).

54. Reynolds, P. *et al.* The dynamics of Ku70/80 and DNA-PKcs at DSBs induced by ionizing radiation is dependent on the complexity of damage. *Nucleic Acids Res.* **40**, 10821–10831 (2012).

55. Gu, J. & Lieber, M. R. Mechanistic flexibility as a conserved theme across 3 billion years of nonhomologous DNA end-joining. *Genes Dev.* **22**, 411–415 (2008).

56. Johnson, R. D. Sister chromatid gene conversion is a prominent double-strand break repair pathway in mammalian cells. *EMBO J.* **19**, 3398–3407 (2000).

57. Gaines, W. A. *et al.* Promotion of presynaptic filament assembly by the ensemble of S. cerevisiae Rad51 paralogues with Rad52. *Nat. Commun.* **6**, 7834 (2015).

58. Wright, W. D., Shah, S. S. & Heyer, W. D. Homologous recombination and the repair of DNA double-strand breaks. *J. Biol. Chem.* **293**, 10524–10535 (2018).

59. Khanna, K. K. & Jackson, S. P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet.* **27**, 247–54 (2001).

60. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 8–11 (2019).

61. Girirajan, S., Campbell, C. D. & Eichler, E. E. Human Copy Number Variation and Complex Genetic Disease. *Annu. Rev. Genet.* **45**, 203–226 (2011).

62. Li, W., Lee, A. & Gregersen, P. K. Copy-number-variation and copy-number-alteration region detection by cumulative plots. *BMC Bioinformatics* **10**, 1–11 (2009).

63. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).

64. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).

65. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4 (2008).

66. Yang, T.-L., Guo, Y., Papasian, C. J. & Deng, H.-W. Genetics of Bone Biology and Skeletal Disease 9. *Genet. Bone Biol. Skelet. Dis.* 123–132 (2013). doi:10.1016/B978-0-12-387829-8.00009-3

67. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).

68. Gamazon, E. R. & Stranger, B. E. The impact of human copy number variation on gene expression. *Brief. Funct. Genomics* **14**, 352–357 (2015).

69. Ballinger, T. J. *et al.* Modeling double strand break susceptibility to interrogate structural variation in cancer. *Genome Biol.* (2019).

70. Alaei-Mahabadi, B., Bhadury, J., Karlsson, J. W., Nilsson, J. A. & Larsson, E. Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 13768–13773 (2016).

71. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

72. Szilard, R. K. *et al.* Systematic identification of fragile sites via genome-wide location analysis of γ-H2AX. *Nat. Struct. Mol. Biol.* **17**, 299–305 (2010).

73. Iacovoni, J. S. *et al.* High-resolution profiling of γh2AX around DNA double strand breaks in the mammalian genome. *EMBO J.* **29**, 1446–1457 (2010).

74. Wienert, B. *et al.* Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science.* **364**, 286–289 (2019).

75. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).

76. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361–365 (2013).

77. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).

78. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: Causes and strategies for minimization. *Clin. Chem.* **61**, 64–71 (2015).

79. Biernacka, A. *et al.* i-BLESS is an ultra-sensitive method for detection of DNA double-strand breaks. *Commun. Biol.* **1**, (2018).

80. Lensing, S. V. *et al.* DSBCapture: In situ capture and sequencing of DNA breaks. *Nat. Methods* **13**, 855–857 (2016).

81. Canela, A. *et al.* DNA Breaks and End Resection Measured Genome-wide by End Sequencing. *Mol. Cell* **63**, 898–911 (2016).

82. Zhu, Y. *et al.* qDSB-Seq is a general method for genome-wide quantification of DNA double-strand breaks using sequencing. *Nat. Commun.* **10**, 2313 (2019).

83. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243 (2015).

84. Kim, D. *et al.* Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**, 863–868 (2016).

85. Tsai, S. Q. *et al.* CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).

86. Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat. Biotechnol.* **29**, 816–823 (2011).

87. Wang, X. *et al.* Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **33**, 175–179 (2015).

88. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–198 (2015).

89. Chiarle, R. *et al.* Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107–119 (2011).

90. Frock, R. L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–188 (2015).

91. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).

92. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220–228 (2010).

93. Berg, S. *et al.* Ilastik: Interactive Machine Learning for (Bio)Image Analysis. *Nat. Methods* (2019). doi:10.1038/s41592-019-0582-9

94. Li, H. & Durbin, R. Making the Leap: Maq to BWA. *Mass Genomics* **25**, 1754–1760 (2009).

95. Yan, W. X. *et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 1–9 (2017).

96. Hoeijmakers, W. A. M., Bártfai, R., Françoijs, K. J. & Stunnenberg, H. G. Linear amplification for deep sequencing. *Nat. Protoc.* **6**, 1026–1036 (2011).

97. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

98. Madabhushi, R. *et al.* Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *Cell* **161**, 1592–1605 (2015).

99. Schwer, B. *et al.* Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proc. Natl. Acad. Sci.* **113**, 2258–2263 (2016).

100. Baranello, L. *et al.* RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription. *Cell* **165**, 357–371 (2016).

101. Baranello, L. *et al.* DNA break mapping reveals topoisomerase II activity genome-wide. *Int. J. Mol. Sci.* **15**, 13111–13122 (2014).

102. Aryal, N. K., Wasylishen, A. R. & Lozano, G. CRISPR/Cas9 can mediate high-efficiency off-target mutations in mice in vivo. *Cell Death Dis.* **9**, 9–11 (2018).

103. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **24**, 927–930 (2018).

104. Currall, B. B., Chiang, C., Talkowski, M. E. & Morton, C. C. Erratum to: Mechanisms for Structural Variation in the Human Genome. *Curr. Genet. Med. Rep.* **1**, 201–201 (2013).

105. Martincorena, I. *et al*. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

106. Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).

107. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

108. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 1–11 (2016).

109. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069–1075 (2008).

110. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7**, 233–245 (2007).

111. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

112. Gerlinger, M. *et al.* Cancer: Evolution Within a Lifetime. *Annu. Rev. Genet.* **48**, 215–236 (2014).

113. Cariati, F. *et al.* Dissecting Intra-Tumor Heterogeneity by the Analysis of Copy Number Variations in Single Cells: The Neuroblastoma Case Study. *Int. J. Mol. Sci.* **20**, (2019).

114. Zhang, X. *et al*. CUTseq is a versatile method for preparing multiplexed DNA sequencing libraries from low-input samples. *Nat. Commun*. **10**, 4732 (2019).

115. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).

116. Arriola, E. *et al.* Genomic analysis of the HER2/TOP2A amplicon in breast cancer and breast cancer cell lines. *Lab. Investig.* **88**, 491–503 (2008).

117. Neve, R. M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515–527 (2006).

118. Yates, L. R. *et al.* Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* **32**, 169-184.e7 (2017).

119. Forbes, S. A. *et al.* COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

120. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

121. Gothe, H. J. *et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Mol. Cell* **75**, 267-283.e12 (2019).

122. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nat. Methods* **10**, 325–327 (2013).

123. Srinivasan, M., Sedmak, D. & Jewell, S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *Am. J. Pathol.* **161**, 1961–1971 (2002).

124. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science (80-.).* **351**, 84–88 (2016).

125. Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

126. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).

127. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, (2018).

128. Iannelli, F. *et al.* A damaged genome's transcriptional landscape through multilayered expression profiling around in situ-mapped DNA double-strand breaks. *Nat. Commun.* **8**, 1–7 (2017).

129. Adam, S. & Polo, S. E. Blurring the line between the DNA damage response and transcription: The importance of chromatin dynamics. *Exp. Cell Res.* **329**, 148–153 (2014).

130. Shanbhag, N. M., Rafalska-Metcalf, I. U., Balane-Bolivar, C., Janicki, S. M. & Greenberg, R. A. ATM-Dependent chromatin changes silence transcription in cis to dna double-strand breaks. *Cell* **141**, 970–981 (2010).

131. Svejstrup, J. Q. The interface between transcription and mechanisms maintaining genome integrity. *Trends Biochem. Sci.* **35**, 333–338 (2010).

132. Dellino, G. I. *et al.* Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations. *Nat. Genet.* **51**, 1011–1023 (2019).

133. Gothe, H. J. *et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Mol. Cell* 1–17 (2019). doi:10.1016/j.molcel.2019.05.015

134. Merlo, D., Mollinari, C., Racaniello, M., Garaci, E. & Cardinale, A. DNA Double Strand Breaks: A Common Theme in Neurodegenerative Diseases. *Curr. Alzheimer Res.* **13**, 1208–1218 (2016).

135. Kanungo, J. DNA-dependent protein kinase and DNA repair: Relevance to Alzheimer's disease. *Alzheimer's Res. Ther.* **5**, (2013).

136. Denholtz, M. & Plath, K. Pluripotency in 3D: Genome organization in pluripotent cells. *Curr. Opin. Cell Biol.* **24**, 793–801 (2012).

137. Falk, A. *et al.* Capture of neuroepithelial-like stem cells from pluripotent stem cells provides a versatile system for in vitro production of human neurons. *PLoS One* **7**, 1–13 (2012).

138. Gelali, E. *et al.* iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture. *Nat. Commun.* **10**, 1–15 (2019).

139. Rybak, P. *et al.* Low level phosphorylation of histone H2AX on serine 139 (γH2AX) is not associated with DNA double-strand breaks. *Oncotarget* **7**, 49574–49587 (2016).

140. Tu, W. Z. *et al.* γh2AX foci formation in the absence of DNA damage: Mitotic H2AX phosphorylation is mediated by the DNA-PKcs/CHK2 pathway. *FEBS Lett.* **587**, 3437–3443 (2013).

141. Gavrieli, Y., Sherman, Y. & Ben-Sasson, S. A. Identification of programmed cell death in situ via specific labeling of nuclear DNA fragmentation. *J. Cell Biol.* **119**, 493–501 (1992).

142. Olive, P. L., Wlodek, D. & Banath, J. P. DNA Double-Strand Breaks Measured in Individual Cells Subjected to Gel Electrophoresis. *Cancer Res.* **51**, 4671–4676 (1991).

143. Wu, C. *et al.* RollFISH achieves robust quantification of single-molecule RNA biomarkers in paraffin-embedded tumor tissue samples. *Commun. Biol.* **1**, 209 (2018).

144. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17**, 81–92 (2016).

145. Shin, H. Y. *et al.* CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome. *Nat. Commun.* **8**, 1–10 (2017).

146. Lepage, C. C., Morden, C. R., Palmer, M. C. L., Nachtigal, M. W. & McManus, K. J. Detecting chromosome instability in cancer: Approaches to resolve cell-to-cell heterogeneity. *Cancers (Basel).* **11**, 1–20 (2019).

147. Van Loo, P. & Voet, T. Single cell analysis of cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 82–91 (2014).

148. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science.* **359**, 555–559 (2018).

149. Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science.* **350**, 94–98 (2015).