CARDIFF
UNIVERSITY

PRIFYSGOL
CAERDYDD

Determining the mechanism of off-target mutagenesis caused by CRISPR-Cas9 genome editing

Felix Michael Dobbs

Division of Cancer and Genetics
School of Medicine
Cardiff University

A Thesis Submitted to Cardiff University
For the Degree of
Doctor of Philosophy

2021

Acknowledgements

First and foremost, I would like to thank my principal supervisor Professor Simon Reed for giving me the opportunity to undertake this PhD, for supporting me wholeheartedly, and for continuously challenging me over the last four years.

I would also like to thank my second supervisor Dr Mick Fellows for providing valuable insight into the workings of industry, and for his hospitality during my visit to AstraZeneca.

Additionally, I would like to thank Dr Patrick van Eijk for our daily energised discussions and for supporting all elements of my PhD work.

Lastly, I would like to acknowledge my partner Katie for listening during endless one-way discussions about the genome and for supporting my ambitious and often risky plans for the future.

Summary

Genome editing using CRISPR-Cas9 holds considerable promise for the generation of cell and gene therapies that can treat human disease. Despite this, the safety of CRISPR-Cas9 has been questioned because it can induce off-target DNA double strand breaks (DSB) in the genome, which often lead to mutagenic outcomes that directly drive oncogenesis. Several next generation sequencing (NGS) approaches have been developed to address this, but are biased by the PCR-based NGS library preparation that they employ. Consequently, these approaches do not provide an accurate measurement of CRISPR-Cas9 off-target activity in the genome. This thesis describes a method, INDUCE-seq, that was developed to solve this problem. In Chapter III, the development, and characteristics of INDUCE-seq are shown. INDUCE-seq combines a PCR-free methodology with a novel use of the Illumina sequencing flow cell for DSB enrichment to eliminate amplification bias. Each INDUCE-seq read is equivalent to a single DSB-end, generating an undistorted, digital readout for genomic breaks in cells. Chapter IV demonstrates the application of INDUCE-seq for measuring induced and endogenous DSBs in live cells. This work reveals the characteristics of AsiSI-induced cleavage and benchmarks INDUCE-seq against alternative methods. Furthermore, distinctive non-random endogenous break distributions are shown for different cell types. In Chapter V, INDUCE-seq is applied to the study of CRISPR off-targets in the genome. INDUCE-seq demonstrates enhanced sensitivity, discovering novel CRISPR off-targets in live cells, in addition to known sites that were previously only detectable using *in vitro* approaches. These off-targets are further analysed in Chapter VI, where it is shown that the sensitivity of DSB detection by INDUCE-seq outperforms amplicon sequencing for measuring the frequency of mutational editing outcomes. In conclusion, this thesis presents the novel method INDUCE-seq for measuring genomic DSBs. The information provided by INDUCE-seq will directly enable the development of safer CRISPR-based cell and gene therapies.

Contents

1 Chapter I - Introduction

1.1 CRISPR-Cas9 genome editing

Ever since the discovery of the DNA double helix, scientists have sought ways to alter the genetic code of genomes. In particular, the ability to make precise, rational edits to an organism's genome offers the ability to probe gene function, introduce or correct mutations, and engineer proteins for a range of applications. Indeed, technologies for making and manipulating DNA have enabled many of the advancements in biology over the last 60 years. Notable examples include solid-phase DNA synthesis, restriction endonucleases and recombinant DNA technology, the polymerase chain reaction (PCR), and next generation sequencing (NGS), all having undoubtedly transformed modern biology. In 2012 it was first reported that a RNA-programmed bacterial nuclease could be designed using the principles of Watson-Crick base pairing to target and cleave DNA (Jinek et al. 2012). This tool, known as clustered regularly interspaced short palindromic repeats (CRISPR) CRISPR associated protein 9 (Cas9), originated from the CRISPR-Cas adaptive bacterial immune system, and would go on to revolutionise and democratise genome editing for the scientific community. CRISPR-Cas9 has since been used to edit the genomes of cultured cells, mammals, and plants, which has accelerated the pace of fundamental research, and enabled novel clinical and agricultural breakthroughs. Nobel laurate Sydney Brenner once remarked "Progress in science depends on new techniques, new discoveries and new ideas, probably in that order." (Robertson 1980). CRISPR-Cas9 is undoubtedly evidence of this and demonstrates unequivocally that explosive scientific progress is made possible by emerging enabling technologies.

1.2 Eukaryotic genome editing before CRISPR

Genome editing is the discipline of converting a specific DNA sequence into a new desired DNA sequence directly within the context of a cell's genome. In doing this, permanent alterations can be conferred to "living" genetic material, which can result in a broad range of phenotypic alterations depending on the context. For the last few decades, the most prominent way to introduce DNA sequence alterations into the genomes of living organisms has been to direct nucleases to a site of interest and introduce a DNA double strand break (DSB) (Urnov 2018). The central principle behind DSB-induced genome editing is as follows: once a DSB is induced at the target site of interest, endogenous DNA repair pathways repair the break to confer genetic changes at the site of the break. In the presence of a donor DNA molecule, DSBs are repaired by homologous recombination (HR, often referred to as homology-directed repair (HDR), within the in the context of genome editing), incorporating the donor sequence at the site of the break. A simpler approach is to induce DSBs in the absence donor DNA, resulting in DSB repair by non-homologous end joining (NHEJ), which can generate variable-length nucleotide insertions or deletions (indels), disrupting the target sequence and resulting gene function (**Figure 1.1**) (Fernandez et al. 2017).

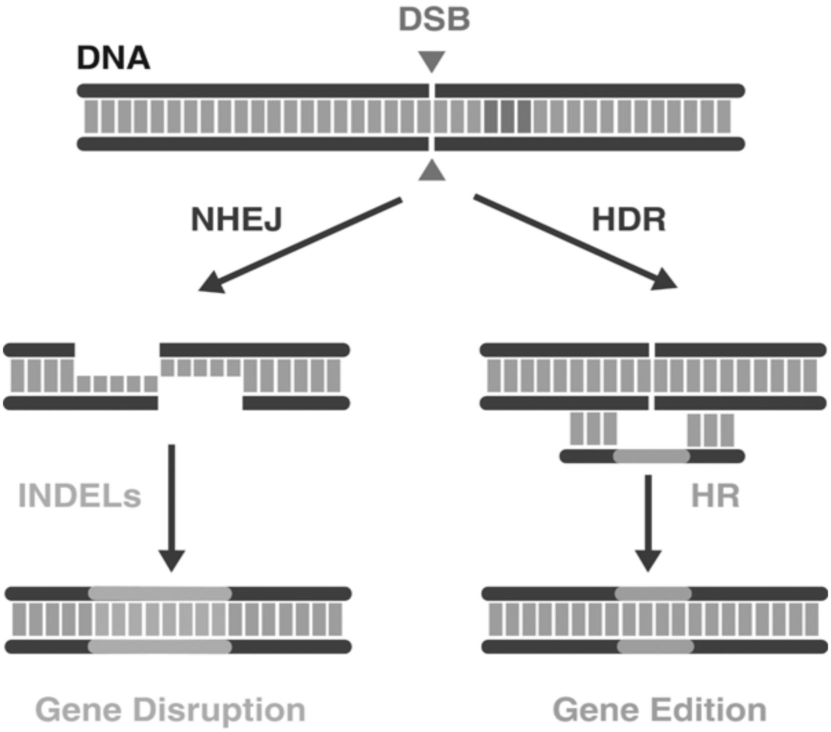


Figure 1.1. The basic paradigm illustrating nuclease-induced genome editing. Following DSB induction at a genomic target of choice, the DSB is repaired either by NHEJ or HR (HDR). NHEJ requires no template for repair and typically results in short insertions and deletions that disrupt gene function. In the presence of a donor sequence, HR can repair in a templated manner to incorporate the desired sequence at the target site. Adapted from: Fernandez et al. 2017.

1.2.1 Meganucleases

Early studies using the I-SceI yeast Meganuclease (**Figure 1.2A**), a rare-cutter endonuclease with a recognition sequence of 18 base pairs in length, demonstrated the promotion of homologous recombination (HR) at a predetermined position in the mouse genome with a frequency two orders of magnitude greater than occurred spontaneously (Rouet et al. 1994; Choulika et al. 1995). This approach has been adapted for gene targeting in mouse embryonic stem (ES) cells, whereby a I-SceI recognition sequence was first inserted at the villin gene locus. Following transient expression of I-SceI, HR was induced at the target locus, resulting in a 100-fold increase in gene targeting efficiency when compared with conventional gene targeted approaches (Cohen-Tannoudji et al. 1998). While the meganuclease system was cumbersome, lengthy, and only functional at pre-defined locations possessing the 18bp recognition sequence, it set the precedent for future genome editing via nuclease cleavage and established the paradigm that DSB induction resulted in efficient editing.

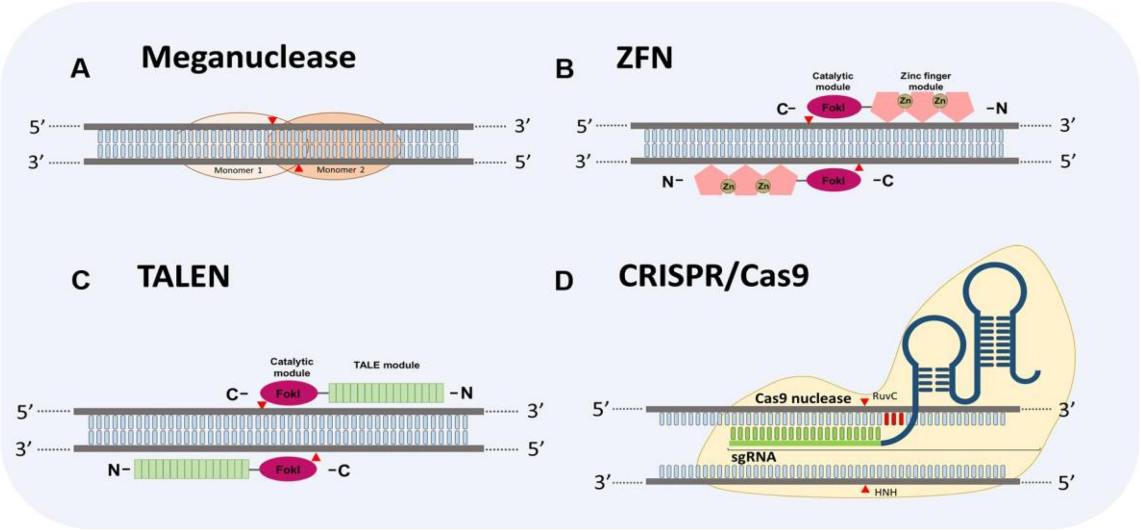


Figure 1.2. The four major classes of nuclease-based genome editing tools. (**A**) Yeast meganuclease comprising a homodimer structure that recognises an 18bp sequence. (**B**) Zinc-finger nucleases (ZFN) targeting DNA in pairs via programmable zinc-finger modules coupled with non-specific FokI domains. (**C**) Transcription factor-like effector nucleases (TALENS) targeting DNA in pairs via programmable transcription activator-like effector (TALE) modules coupled with non-specific FokI domains. (**D**) Clustered regularly interspaced short palindromic repeats CRISPR) CRISPR associated protein 9 (Cas9) (CRISPR-Cas9), targets DNA via a guide RNA component that directs two nonspecific nuclease domains. Taken from: Romay and Bragard 2017.

1.2.2 Zinc-finger nucleases

Several years later, Zinc-finger nucleases (ZFNs) emerged as a novel genome editing tool, addressing the evident limitations of meganucleases in that they can only cleave pre-defined sequences. ZFNs were the first truly targetable genome editing tools and demonstrated that through protein engineering, arbitrary DNA sequences could be cleaved on demand. ZFNs comprise a fusion of a series of zinc-finger based DNA binding domains and the cleavage domain from the bacterial restriction endonuclease *Fok*I. Each ZFNs is capable of targeted cleavage of a single DNA strand, and thus must be designed in pairs in order to cleave two adjacent sites forming a DNA double strand break (**Figure 1.2B**). ZFNs provide an inherent advantage over meganucleases due to the separation of function of the DNA binding component, a series of zinc-binding repeats in it its DNA binding domain, and the DNA cleavage component provided by the *Fok*I cleavage domain. Previous studies had shown that sequences from other transcription factors acted as peptide modules that made contacts with base pair triplets (Pavletich and Pabo 1991). Changing just a few residues in a single zinc finger therefore altered its DNA-recognition specificity, and thus could be simply designed to target a DNA sequence of interest. *Fok*I is a type IIS restriction endonuclease that recognises asymmetric DNA sequences and cleaves outside of their recognition sequence (Romay and Bragard 2017). Unlike the meganuclease system, which possesses a shared DNA binding and cleavage domain that cannot not be easily modified, the combination of a non-specific DNA cleavage domain and an easily configurable DNA binding domain, made ZFNs a neat modular solution for highly targetable genome editing.

Throughout the 2000’s, ZFNs gained traction for genome editing in a range of organisms, culminating in 2009 with a landmark paper reporting the world’s first knockout rats using the system (Geurts et al. 2009). Clinical gene therapy applications were also explored, with several groups reporting selection-free, efficient knockout of mammalian genes in transformed human cells (Maeder et al. 2008; Santiago et al. 2008). Primary human T cells would then follow, demonstrating the potential of nuclease genome editors to edit primary cells for a range of clinical applications (Maier et al. 2013). Despite the relative success of ZFNs, and although certainly an improvement over the previous generation of meganuclease-based editing tools, ZFNs were not easily affordable or accessible to standard molecular biology laboratories, and thus usage for mammalian transgenesis was largely limited to commercial entities. Furthermore, some unwanted consequences of the more tuneable system were observed; sequence specificity of ZFNs was not strict, and similar sequences known as off-targets can be cleaved by the nuclease. The result of this was the unwanted insertion of a single-stranded oligodeoxynucleotide (ssODN) at off-target sites, questioning the safety of such genome editing applications in the clinical setting (Radecke et al. 2010).

1.2.3 Transcription factor-like effector nucleases

Building on the success of ZFNs, transcription factor-like effector nucleases (TALENS) represent a significant improvement over the limitations of the ZFN system. TALENS are also chimeric proteins and comprise the same *Fok*I endonuclease cleavage domain as ZFN, but substitute the zinc-finger binding domain for highly conserved repeats derived from transcription activator-like effectors (TALEs), which are proteins secreted by *Xanthomonas* bacteria to alter transcription of genes in host plant cells (**Figure 1.2C**) (Christian et al. 2010; Li et al. 2011; Joung and Sander 2013). Because of these structural similarities, TALENS also operate in dimers to induce DSBs to adjacent opposing-strand target regions. Where TALEs differ from zinc-finger binding domains is in the simplicity of design; ZFNs must be engineered in a laboratory, each with a defined target based on inferred code that links groups of three amino acids with three nucleotides. TALENS on the other hand, can be simply and rapidly designed by researchers using a protein-DNA code that relates modular DNA-binding TALE repeat domains to individual bases in a target binding-site. These repeats can be assembled in most molecular biology laboratories using plasmid intermediates, standard restriction endonucleases, and golden gate assembly, which reduced the barrier to entry compared with ZFNs (Fernandez et al. 2017). Whilst simple in principle, execution of customised TALENS systems is a time-consuming and cumbersome process, due to the numerous plasmids that are required to assemble each targeting system. If not assembling the system in house, commercial pricing was also a limitation at ~$5000 per target, making it prohibitive for smaller laboratories. As with ZFNs, off-targets are a significant concern with TALENs; a study in which TALENs were used for genome editing in human pluripotent stem cells found low but measurable indels at several of the 19 predicted off-target sites based on sequence (Hockemeyer et al. 2011). Since the first reporting of TALENs for genome editing in 2010, several gene therapy applications have been demonstrated, including the genetic correction of sickle cell disease (SCD) mutations in patient-derived, human induced pluripotent stem cells (hiPSCs), and editing T cells to treat paediatric leukaemia (Ramalingam et al. 2014; Qasim et al. 2017). These applications are few, however, and therefore before TALENs could take hold as a mainstream and viable alternative to ZFNs, a new technology was on the horizon that would radically change genome editing for the foreseeable future.

1.3 The emergence of CRISPR

In the mid 2000's, in a parallel and separate research field to that of genome editing, several microbiology and bioinformatics groups began investigating clustered regularly interspaced short palindromic repeats (CRISPRs) in the genomes of bacteria and archaea (Doudna and Charpentier 2014). These CRISPR loci were first described in 1987 (Ishino et al. 1987) as a series of short direct repeats interspaced with short non-repeating sequences in the genome of *Escherichia coli*. These received little attention until over a decade later. During the human genome project (HGP), the genomes of many other organism were sequenced in line with increasing sequencing capabilities globally. Following *in silico* analysis of bacterial and archaeal genome drafts, CRISPR loci were identified in >40% of bacterial genomes and 90% of archaeal genomes (Mojica et al. 2000). Furthermore, these same elements were almost always adjacent to multiple, well-conserved CRISPR-associated (Cas) genes (Jansen et al. 2002). The most striking observation about these CRISPR loci was that the non-repeating spacer DNA sequences shared near exact homology with sequences derived from mobile genetic elements of viral origin (Bolotin et al. 2005; Mojica et al. 2005; Pourcel et al. 2005). These findings sparked a surge of research activity to study the functional significance of the CRISPR loci. Makarova *et al*. 2006 hypothesised that CRISPR might function as a form of bacterial adaptive immune system based on two observations: 1) these loci are transcribed, and therefore are not so-called 'junk' repetitive elements, potentially serving as a memory of past viral infections, and 2) Cas genes encode proteins with putative nuclease and helicase domains (Haft et al. 2005). The first experimental evidence of CRISPR-Cas mediated adaptive immunity was reported by Barrangou, Horvath, and colleagues in 2007, demonstrating that following infection of *Streptococcus thermophilus* with lytic phages, the host integrates new spacer sequences derived from the viral DNA in to the CRISPR locus of its own genome. Not only this, but these same spacer sequences conferred targeting specificity to the Cas enzymes, which provides resistance to the invading phage DNA (Barrangou et al. 2007). Since this work, many different CRISPR systems have been discovered, broadly separating in to two classes each with multiple sub-types. Class 1 comprise multi-subunit effector nuclease complexes and contain type I and type II CRISPR systems that are mostly identified in Archaea. Class 2 comprise single-protein effector modules and contain type II, IV, V and VI CRISPR systems (Koonin et al. 2017).

1.3.1 The mechanism of CRISPR-Cas9 mediated adaptive immunity

Further investigations to elucidate the mechanism of CRISPR-based adaptive immunity revealed that it occurred across three stages: Adaptation, Expression, and Interference (**Figure 1.3. Outline of the CRISPR bacterial adaptive immune system**) (Knott and Doudna 2018). Although broadly similar, each CRISPR system uses distinct molecular mechanisms with varying complexity to facilitate this process. As it is the most widely recognised and well-characterised CRISPR system to date, this thesis will focus on the biology and application of CRISPR-Cas9, the type II CRISPR-Cas system from *Streptococcus pyogenes* (SpyCas9). The first stage of CRISPR-Cas9 mediated bacterial adaptive immunity is the adaptation stage and occurs following cellular entry of an invading mobile genetic element (**Figure 1.3. Outline of the CRISPR bacterial adaptive immune systemA**). Acquisition of new spacer elements in to the CRISPR locus is mediated by the Cas1-Cas2 complex, which captures and inserts ~30bp of foreign DNA in a polarised manner at the borders of repetitive sequences (**Figure 1.3. Outline of the CRISPR bacterial adaptive immune systemA, diamond sections of CRISPR array**) via a nucleophilic reaction similar to that of transposases (Nunez et al. 2015). Interestingly, this process preferentially inserts newly acquired sequences at the first repeat of the CRISPR locus, following an AT-rich sequence known as the leader sequence. Positioning newly acquired sequences at the start of the CRISPR locus serves as a chronological record of previously encountered foreign nucleic acids, and results in their immediate expression during the second stage of CRISPR-mediated adaptive immunity (van der Oost et al. 2014). During the expression stage, both the CRISPR array and Cas genes are expressed to produce a long precursor transcript known as the pre-crRNA (**Figure 1.3. Outline of the CRISPR bacterial adaptive immune systemB**). Specific to type II systems, in addition to the Cas operon and CRISPR array, a trans-activating crRNA (tracrRNA) is also essential for crRNA maturation (**Figure 1.4A**) (Deltcheva et al. 2011). The tracrRNA comprises a 25 nucleotide sequence that is complementary to the repeat crRNA sequences of the pre-crRNA transcript, which facilitates periodic complementary base pairing of tracrRNA:crRNA adjacent to non-repetitive spacer elements (**Figure 1.4B**). These short sections of dsRNA are recognised and bound by Cas9 (formerly known as Csn1), a large multifunctional protein with two putative nuclease domains, HNH and RuvC-like. This binding event primarily serves as a positional marker for site-specific cleavage by RNase III, which releases the individual repeat-spacer-repeat units from the pre-crRNA. Following processing, the crRNA:tracrRNA duplex remains tightly bound to Cas9 and the 5’ end of the crRNA is trimmed from 24-27 nucleotides, resulting in a spacer which is ~20 nucleotides in length (**Figure 1.4C**). Finally, the Cas9 ribonucleoprotein (RNP) complex facilitates interference of invading DNA via a targeted cleavage mechanism (**Figure 1.3. Outline of the CRISPR bacterial adaptive immune systemC** and **Figure 1.4C**). The non-repetitive spacer element of the crRNA directs the Cas9 nuclease to the complementary protospacer region of the invading DNA. DsDNA cleavage is then initiated approximately three nucleotides upstream of an NGG consensus sequence known as the protospacer adjacent motif (PAM) (Jinek et al. 2012). Requirement of a PAM sequence adjacent to the targeted protospacer sequence protects the DNA encoding the crRNA, preventing self-targeting and cleavage at the CRISPR locus, which bares the same target sequences but lacks the presence of PAM sequences.

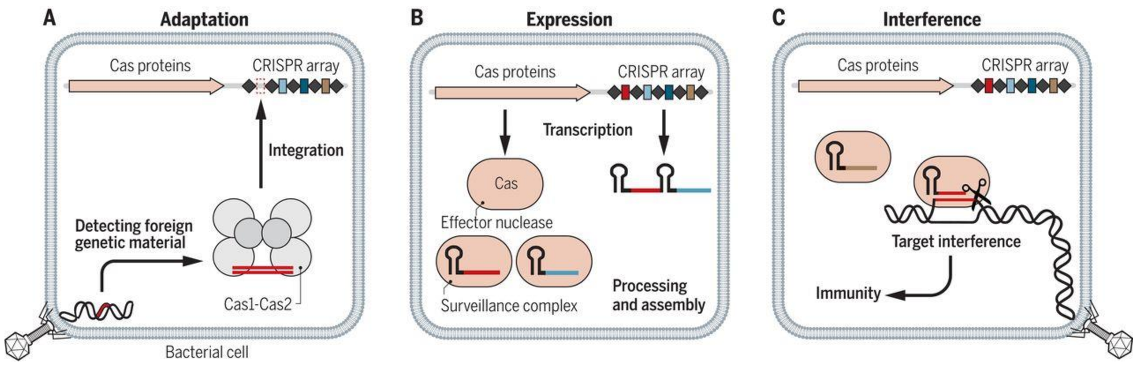


Figure 1.3. Outline of the CRISPR bacterial adaptive immune system. A) Adaptation. Mobile genetic elements are acquired in to the CRISPR array via Cas1-Cas2. **B) Expression.** The CRISPR array and CRISPR-associated proteins are expressed. Resulting RNA is processed to form short crRNA which bind to Cas proteins to form a surveillance complex**. C) Interference.** The Cas effector nuclease, now bound with a guide RNA, targets, and cleaves foreign genetic elements with a complementary sequence to the crRNA. Taken from: Knott and Doudna 2018.

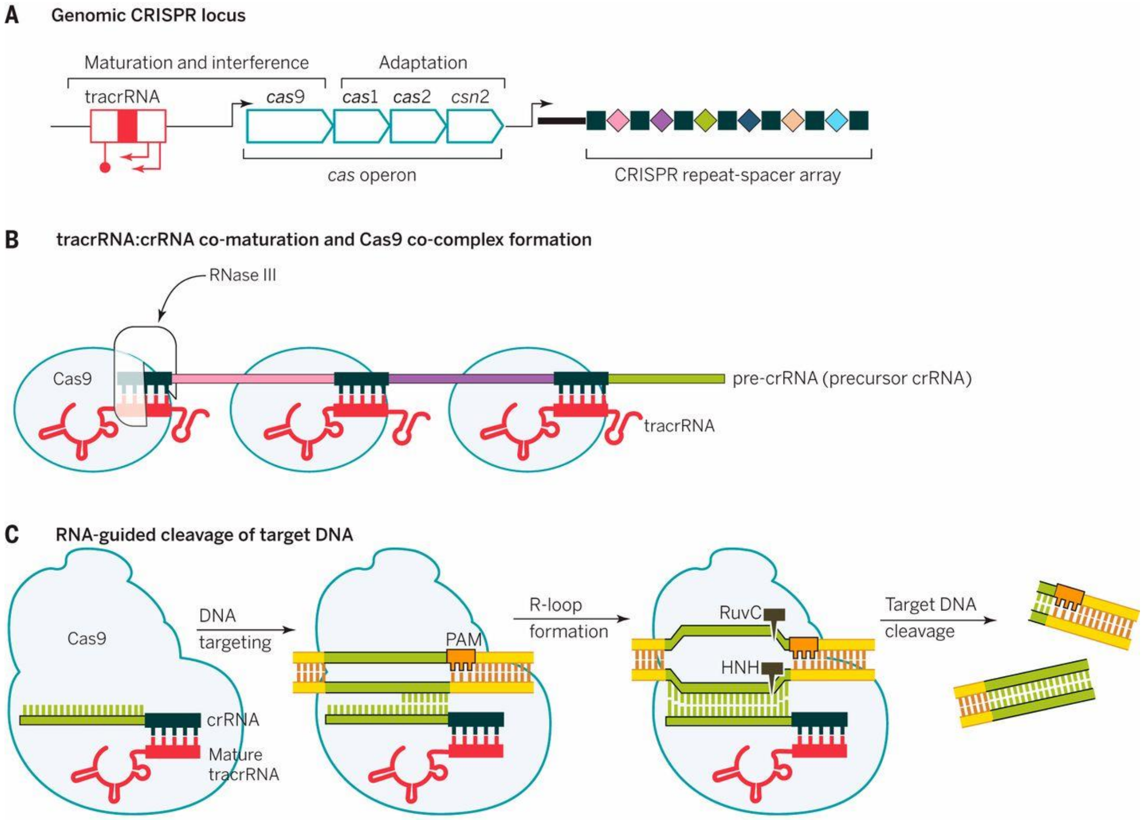

A
Genomic CRISPR locus
Maturation and interference
Adaptation
tracrRNA
cas9
cas1
cas2
csn2
cas operon
CRISPR repeat-spacer array
B
tracrRNA:crRNA co-maturation and Cas9 co-complex formation
RNase III
Cas9
pre-crRNA (precursor crRNA)
tracrRNA
C
RNA-guided cleavage of target DNA
Cas9
DNA targeting
crRNA
Mature tracrRNA
PAM
R-loop formation
RuvC
HNH
Target DNA cleavage

Figure 1.4. Detailed overview of the mechanism of adaptive immunity for the type-II CRISPR-Cas9 system from *S. pyogenes*. (**A**) Structure of the Cas operon with tracrRNA and the CRISPR array. (**B**) Cas9 associates with the antirepeat-repeat structure of the crRNA:tracrRNA duplex and is processed by RNase III to form individual Cas surveillance complexes. (**C**). Targeting is mediated by the ~20nt spacer element of the crRNA, which binds Cas9 to complementary DNA sequences. Cleavage is initiated by RuvC and HNH nuclease domains if a 5’-NGG-3’ protospacer adjacent motif (PAM) is present in the target DNA. Taken from: Doudna and Charpentier 2014.

1.3.2 The CRISPR genome editing revolution

This landmark study, led by Jennifer Doudna and Emmanuel Charpentier, set in motion a genome editing revolution, and would go on to win Doudna and Charpentier the Nobel Prize in Chemistry in 2020. In addition to demonstrating the mechanism of RNA-guided DNA cleavage by Cas9, the same study showed how the system was simply programmable using a ~100 nucleotide single guide RNA for sequence specific cleavage that could be repurposed for genome editing. Within six months, a series of publications followed that demonstrated the system working in a range of eukaryotic models, including mouse cells, transformed human cells, and zebrafish embryos (Cho et al. 2013; Cong et al. 2013; Hwang et al. 2013; Jinek et al. 2013; Mali et al. 2013). Where CRISPR-Cas9 stands apart from the previous generations of genome editing systems is in its simplicity. The ease of design and production of the guide RNA component, which facilitates retargeting of the non-specific Cas9 nuclease, greatly exceeds that of comparable ZFNs and TALENS while offering similar or greater editing efficiencies (Mali et al. 2013). When compared to the technically challenging task of engineering proteins for direct protein-DNA recognition, which is required for ZFNs and TALENS systems, CRISPR-Cas9 offers an elegantly simple solution based on Watson-Crick base pairing that is accessible to any molecular biology laboratory. Furthermore, unlike ZFNs and TALENS, which function as adjacent pairs to generate a DSB (**Figure 1.2B and 1.2C**), CRISPR-Cas9 requires a single targeting and nuclease element to cleave both DNA strands, dramatically simplifying the design process. Early reports demonstrating the use of CRISPR-Cas9 for editing eukaryotic genomes showed that single or multiple targets in cell lines could be simply edited by transfecting cells with plasmids co-expressing codon-optimised versions of SpyCas9, with custom designed guide RNAs (Cong et al. 2013; Mali et al. 2013). Furthermore, based on the requirement of an NGG PAM sequence, Mali et al. 2013 calculated that there are ~190,000 unique guide RNA-targetable sequences in the human exome, comprising 40% of the genes in the human genome. Over the last decade a range of therapeutic targets for CRISPR-Cas9 editing have been identified, and several CRISPR-based cell and gene therapies have since entered early-stage clinical trials. These include *ex vivo* cell therapy editing strategies which generate novel CAR-T cancer immunotherapies, HIV-resistant T cells, and treatments for β-thalassemia and sickle cell disease (Ernst et al. 2020). Furthermore, CRISPR-Cas9 gene therapies that direct edited patient cells *in vivo* have also shown great promise. However, functional delivery of CRISPR-Cas9 *in vivo* remains challenging, and clinical trials have so far been limited to delivery in the eye to correct erroneous gene splicing that treats the inherited disorder Leber's congenital amaurosis (Ernst et al. 2020). Given the vast targetability of CRISPR-Cas9 in the genome, it is evident that these clinical studies are just scratching the surface of the therapeutic possibilities that could be enabled by CRISPR-Cas9 genome editing. Whether this potential is realised, however, will depend on the ability to edit the genome safely. This issue is discussed later in this chapter.

1.3.3 The CRISPR toolbox

In just the last eight years CRISPR-Cas9 has developed from a simple targeted DSB system in to a broad and diverse toolbox for genome modulation. Outside of genome editing, these applications include transcriptional regulation, epigenetic modification, RNA editing, and nucleic acid detection, all of which utilise an ever-expanding repertoire of modified and/or novel CRISPR enzymes (Anzalone et al. 2020). Within the genome editing space, a variety of modifications have been made to improve the original SpyCas9 system, such as high fidelity (HiFi) Cas9 variants with enhanced specificity, inactivation of a single nuclease domain resulting in a targetable nicking enzyme, and Cas9 mutants with less stringent PAM sequence requirements that enable interrogation of a different portion of the genome. Furthermore, alternative Cas proteins with different sequence requirements and biochemical properties have been discovered, offering an exciting opportunity to introduce significantly different types of genome edits hitherto unachievable with SpyCas9. Prominent examples of this are the Cas12 family of type V CRISPR nucleases, including Cas12a (formerly Cpf1), and more recently Cas12e (formerly CasX) (**Figure 1.5**) (Knott and Doudna 2018; Liu et al. 2019a). Cas12 nucleases are guided by a single crRNA, possess a single RuvC-like nuclease domain, recognise a T-rich PAM sequence, and cleave target DNA in a staggered pattern distal to the PAM position (Zetsche et al. 2015). These factors result in a system that is orthogonal to SpyCas9; different regions of the genome are targeted by Cas12 owing to the different PAM sequence requirements. In addition, the staggered DSB results in edits that are more prone to large and more complex insertions and deletions (Sansbury et al. 2019).

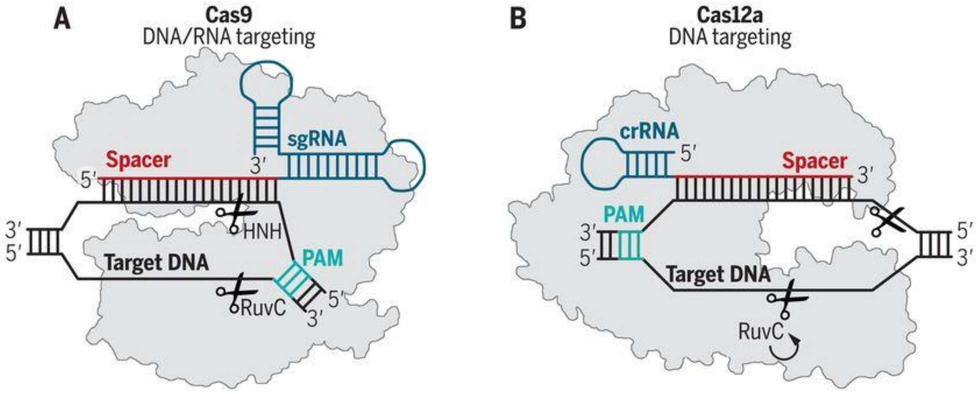


Figure 1.5. **Cas9 and Cas12a target binding and cleavage**. (**A**) Class 2 type II CRISPR-Cas9 system. Cas9 binds to single guide RNA (blue), which comprises the repeat-antirepeat structure from the fused crRNA:tracRNA and a targeting spacer element (red). Cleavage is initiated ~3bp upstream of the PAM by HNH and RuvC nuclease domains generating a blunt DSB. (**B**) Class 2 type V CRISPR-Cas12a system. A single crRNA (blue) binds to Cas12a and confers targeting via a spacer element (red). A RuvC nuclease cleaves both strands of the target DNA distal to the PAM, generating a staggered DSB with overhanging sequences. Adapted from: Knott and Doudna 2018.

Towards the goal of more precise and controllable genome editing, several other classes of genome editors have been adapted from the original SpyCas9 system. The largest and most diverse of these tools are base editing systems, which have been developed to enable precise single-nucleotide alterations (**Figure 1.6B**). Most base editing systems comprise a partially, catalytically inactivated Cas9 nickase, and a single-strand DNA deaminase enzyme which typically catalyses base conversions within a 4-6bp window adjacent to the PAM (Anzalone et al. 2020). Following deamination of a target nucleotide on the target strand, nicking of the opposite non-edited strand causes cellular DNA repair pathways to repair the broken strand via templated repair of the deaminated bases, thus incorporating the alteration into both strands (Komor et al. 2016). To date, this approach has been successfully applied to two classes of base editors: cytosine base editors (CBEs) and adenine base editors (ABEs). CBEs use cytidine deaminases to convert cytosines to uracils, which are read by polymerases as thymines, and catalyses the conversion of G•C base pairs to T•A base pairs. Alternatively, ABEs perform similarly using a deoxyadenosine deaminase to convert adenosines to inosines, which are read by polymerases as guanines, and catalyses the conversion of A•T to G•C pairs. CBE and ABE systems together provide a means to mediate all possible transition mutations (C→T, A→G, T→C, G→A), and have been applied broadly in a variety of cell types and animal models of genetic disease to revert pathogenic single-nucleotide variants (SNV).

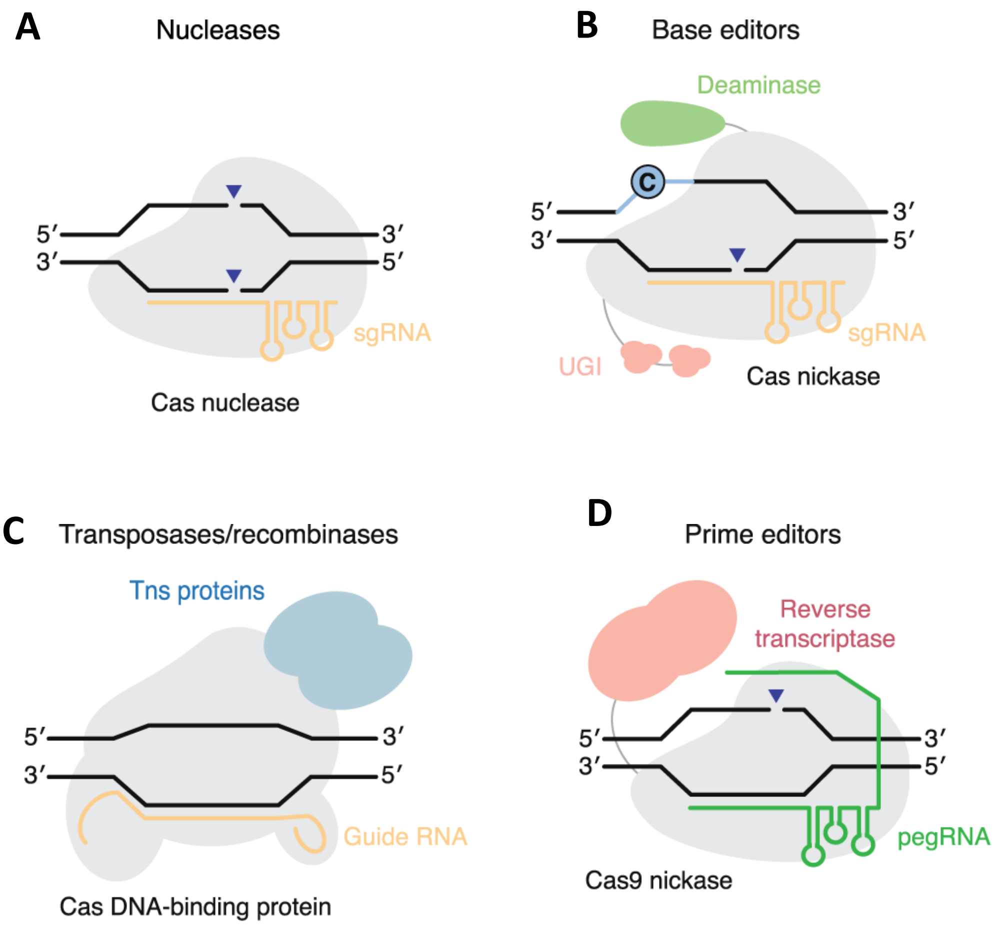


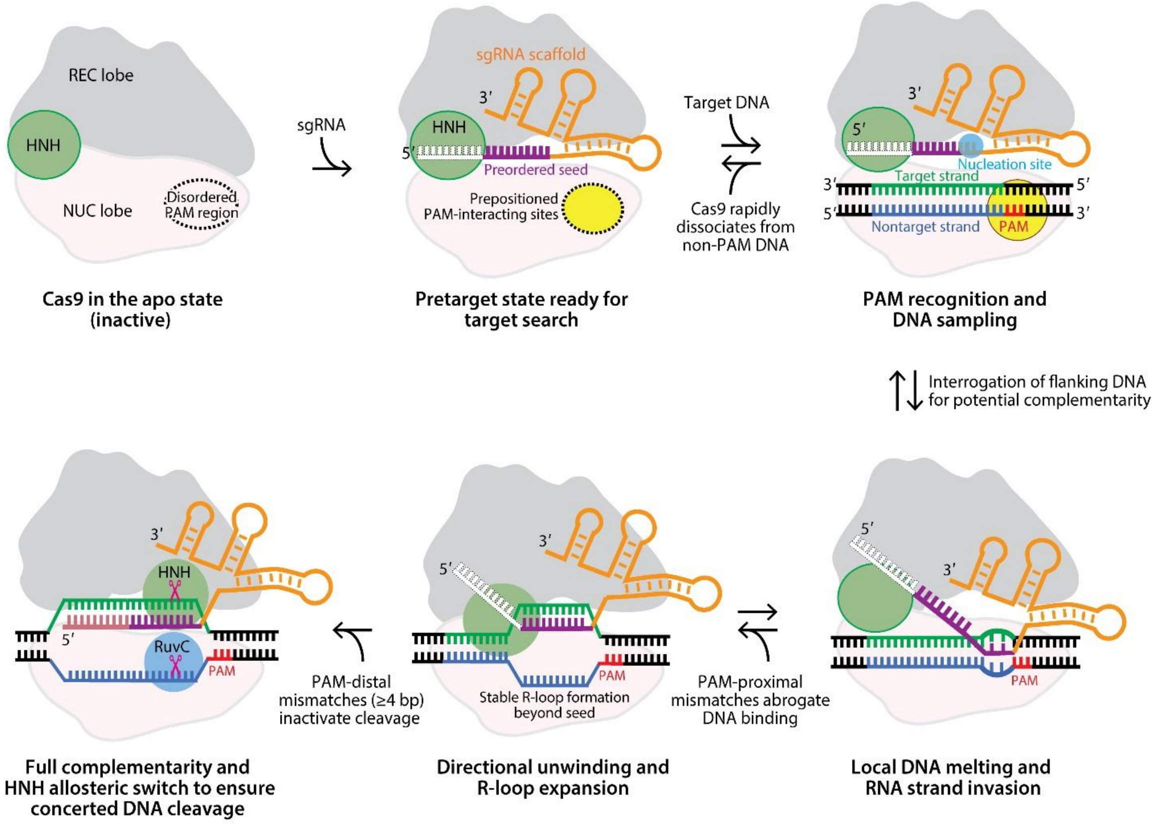
Figure 1.6. The four main classes of CRISPR genome editing tools. (**A**) Cas nucleases cleave target DNA to form a DSB, which is repaired by NHEJ or HR. (**B**) Base editors catalyse single base conversions via a deaminase fusion and strand nicking. (**C**) Transposases insert DNA at desired regions. (**D**) Prime editors use templated DNA synthesis and strand nicking to confer a range of editing outcomes. Adapted from: Anzalone et al. 2020.

More recently, Cas transposases and prime editors have emerged as novel approaches to further expand the genome editing toolbox (**Figure 1.6C**). Cas transposases offer the potential to improve the efficiency and specificity of targeted DNA insertion over traditional nuclease-induced DSB formation and HDR-based approaches. While the latter strategy has been relatively successful in several contexts, key limitations persist such as low efficiency in most therapeutically relevant cell types and producing high frequencies of unwanted indels at the target site (Komor et al. 2016). Briefly, Cas transposases comprise a fusion of an inactivated Cas protein, which is targeted via a guide RNA to a corresponding genomic region with a PAM sequence, and a Tn7-like transposase, which catalyses the integration of cargo DNA at the target site. Although Cas transposases potentially offer an attractive alternative to DSB-based insertion of genetic material, the technology is still in its infancy and has so far only been demonstrated *in vitro* and in bacterial cells (Anzalone et al. 2020). Building on the success of base editors, prime editing was developed to expand the repertoire of targeted point mutations and precise short insertions and deletions (**Figure 1.6D**). Prime editing is conceptually similar to base editing tools, and comprises a fusion of a Cas9 nickase domain and an engineered reverse transcriptase domain (Anzalone et al. 2019). Unlike base editing, however, prime editing systems are targeted via a prime editing guide RNA (pegRNA), which incorporates a template sequence used for editing in addition to the usual targeting spacer sequence. Following binding to the target, the Cas9 nickase domain cleaves the target strand and uses the released 3' end to prime reverse transcription using the pegRNA as a template. The resulting 3' DNA flap containing the edited DNA displaces the 5' sequence containing the non-edited DNA and following a ligation event a heteroduplex structure is formed. Incorporation of the final edited DNA occurs via nicking of the non-edited strand by another guide RNA, promoting DNA repair to replace the non-edited strand with sequence complementary to the edited strand (Anzalone et al. 2019). Prime editing has the potential to be more versatile than previous generations of base editors and traditional nuclease-based genome editors, and edits including insertions, deletions and all 12 types of point mutations have been demonstrated in several cell types. As with Cas transposases, prime editing is in its infancy, and has several limitations such as complex pegRNA design and poor understanding of the DNA repair outcomes that must be resolved prior to widespread adoption of the technology.

1.4 The mechanism of CRISPR-Cas9 mediated DSB formation

Despite the many derivative CRISPR genome editing technologies available, this thesis focuses on the mechanisms associated with the CRISPR-Cas9 system from

Streptococcus pyogenes. SpyCas9 is a large (1,368 amino acid, 158.4 kDa) multidomain endonuclease that is comprised of two lobes: the alpha-helical recognition (REC) lobe, and the nuclease (NUC) lobe containing the well-conserved RuvC-like and HNH nuclease domains, in addition to a variable C-terminal domain (CTD) (Jinek et al. 2014; Jiang and Doudna 2017). While the enzyme is unbound by a guide RNA (crRNA:tracrRNA or a single synthetic guide RNA (sgRNA)) it occupies a catalytically inactive confirmation in the apo state. The CTD, which contains PAM-interacting sites required for PAM recognition, occupies a disordered state in the apo-Cas9 structure that prevents DNA-binding in the absence of a guide RNA (**Figure 1.7, top left**) (Jiang et al. 2015). Following guide RNA binding to the REC lobe via direct interactions of stem-loop secondary structures of the tracrRNA component, apo-Cas9 undergoes a substantial structural rearrangement to form an active DNA recognition complex (**Figure 1.7, top middle**) (Jiang et al. 2016). Most notably, the REC lobe undergoes a conformational change forming a cavity between the HNH and RuvC domains. Cas9 also makes extensive contacts with the crRNA component of the guide RNA, specifically with the first 10 nucleotides of the 20-nucleotide long spacer sequence, known as the seed region (Nishimasu et al. 2014). Pre-ordering the seed region in this manner is thought to make it thermodynamically favourable for target binding (Kunne et al. 2014). The PAM-interacting sites, which were previously disordered in the apo state, are repositioned to recognise a 5’-NGG-3’ PAM sequence and confer stable DNA binding activity to the complex (Jiang et al. 2015). Although not directly bound to Cas9, the remaining ~10 nucleotides at the distal 5’ end of the spacer sequence are positioned inside the cavity formed between the nuclease domains, protecting it from degradation prior to DNA binding.



REC lobe
HNH
NUC lobe
Disordered PAM region
Cas9 in the apo state (inactive)
sgRNA
sgRNA scaffold
3′
HNH
5′
Preordered seed
Prepositioned PAM-interacting sites
Pretarget state ready for target search
Target DNA
Cas9 rapidly dissociates from non-PAM DNA
3′
5′
Nucleation site
Target strand
3′
5′
5′
3′
Nontarget strand
PAM
PAM recognition and DNA sampling
Interrogation of flanking DNA for potential complementarity
3′
HNH
5′
RuvC
PAM
Full complementarity and HNH allosteric switch to ensure concerted DNA cleavage
PAM-distal mismatches (≥4 bp) inactivate cleavage
5′
3′
PAM
Stable R-loop formation beyond seed
Directional unwinding and R-loop expansion
PAM-proximal mismatches abrogate DNA binding
5′
3′
PAM
Local DNA melting and RNA strand invasion

Figure 1.7. The mechanism of DSB induction by CRISPR-Cas9. Cas9 undergoes a large confirmational rearrangement following binding of guide RNA to form a target recognition-structure. This involves prepositioning of the PAM interacting sites for PAM sampling, and preordereding of the guide RNA. The complex is further restructured following successful binding to a complementary site with a matching PAM sequence, and local DNA melting enables RNA strand invasion, forming a stable R-loop structure. Finally, full complementarity of guide RNA binding including the distal region of the spacer, coupled with allosteric regulation of the RuvC domain by confirmational change of the HNH domain, facilitate the concerted cleavage of both DNA strands 3bp upstream of the PAM. Taken from: Jiang and Doudna 2017.

Once Cas9 and the guide RNA have formed a stable complex, the RNP is capable of highly specific DNA binding. This process requires both complementary base pairing between the 20 nucleotide spacer region and the protospacer in the target DNA and direct protein-DNA binding to a PAM recognition sequence directly adjacent to the protospacer target. The search for target DNA begins with PAM recognition, where the Cas9 RNP probes for adequate PAM sequences prior to initiation of spacer binding. This complex rapidly dissociates from sites with the incorrect sequence (**Figure 1.7, top right**). The immediate DNA sequences are interrogated for potential complementarity following successful PAM binding. This stimulates local DNA melting and RNA strand invasion (**Figure 1.7, bottom right**) (Sternberg et al. 2014). Given that no mismatches are present between the DNA and RNA in the PAM-proximal seed region, the target DNA strand then forms an RNA-DNA duplex, thus displacing the non-target strand. DNA duplex unwinding and the resulting R-loop structure facilitates strand invasion across the full 20 nucleotide spacer sequence (**Figure 1.7, bottom middle**). The resulting 20 nucleotide RNA-DNA heteroduplex maintains a pseudo A-structure and is positioned in the central channel between the REC and NUC lobes. Correct hybridisation is detected by Cas9 in a sequence-independent manner based on heteroduplex geometry, enabling the nuclease domains to cleave target DNA regardless of sequence composition (Anders et al. 2014). Interestingly, a much greater degree of structural variation and distortion is observed at the 5’ distal end of the spacer sequence even when all 20 nucleotides are base-paired *without* mismatches. Mismatches in this region are therefore more tolerated than those present in the seed region, which abolish stable binding.

The final stages of Cas9-mediated DSB formation occur once a stable RNA-DNA heteroduplex has formed and Cas9 is activated for cleavage. To limit spurious cleavage of the host genome during extensive target surveillance, Cas9 exhibits decoupled DNA binding and cleavage events (Sternberg et al. 2015). Once target binding and R-loop formation have been completed, the complex undergoes another conformational change that primes the two nuclease domains for cleavage. First, the HNH domain is repositioned in an active conformation and cleaves the target strand three nucleotides upstream of the PAM sequence. This conformational change of the HNH domain induces structural changes of the looped linker regions between the HNH and RuvC nuclease domains. This in turn positions the centre of the RuvC catalytic domains at the non-target strand for cleavage (**Figure 1.7, bottom left**) (Palermo et al. 2016). Allosteric regulation of the RuvC domain by the HNH domain, which is particularly sensitive to mismatching nucleotides at the PAM-distal portion of the spacer sequence, ensures highly precise cleavage of a dsDNA target (Sternberg et al. 2015). Finally, following the endonucleolytic cleavage event, Cas9 remains bound to the dsDNA at the PAM site until it is displaced (Kunne et al. 2014).

1.5 Genome editing via repair of CRISPR-Cas9 induced DSBs

As described briefly in section **1.2**, the approach of genome editing via targeted DSB induction is not a new one. Given its ease of use and widespread adoption, CRISPR-Cas9 confirmed this process as the preferred mechanism to facilitate permanent editing of the genome to date. Part of what makes this approach viable is the diverse range of repair outcomes that are possible following DSB formation. Several cellular DNA repair pathways exist to resolve DSBs, which can result in substantially different genome edits depending on the context such as cell cycle timing, cell type, chromatin environment, and surrounding genomic sequence. In general, DSB repair mechanisms can be broadly separated into four distinct sub-pathways; classical non-homologous end joining (c-NHEJ), homologous recombination (HR), alternative end-joining (alt-EJ, sometimes termed microhomology-mediated end-joining (MMEJ)), and single strand annealing (SSA) (**Figure 1.8**) (Ceccaldi et al. 2016).

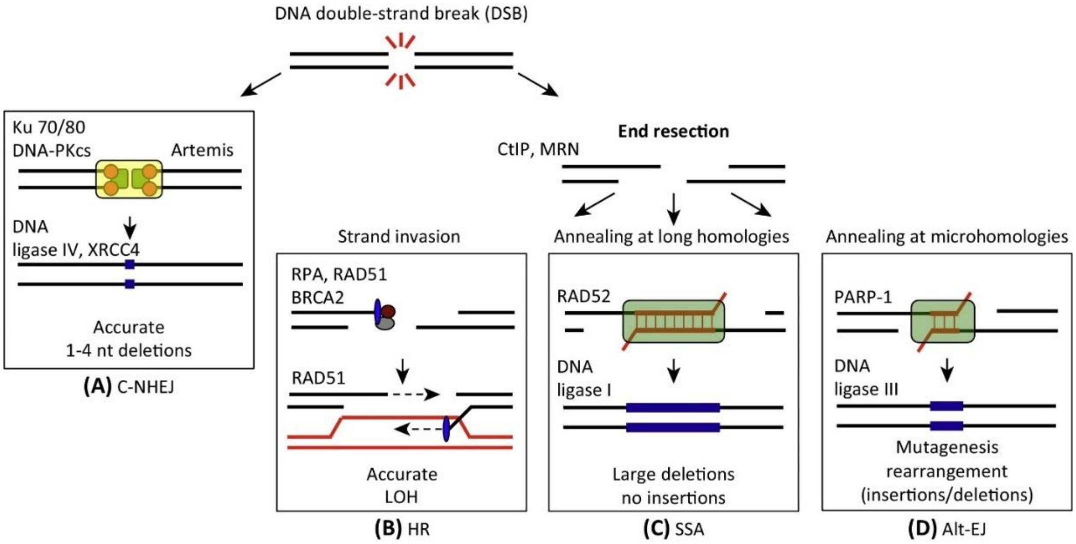


Figure 1.8. **The sub-pathways of DSB repair.** (**A**) When resection is prevented, DSB repair by c-NHEJ is favoured, resulting in short insertions and deletions. When resection occurs, the three alternative pathways compete to repair the DSB. (**B**) HR is the most accurate of the sub-pathways, using a sister chromatid or an exogenous donor sequence for templated error-free repair. (**C**) SSA results in large deletions following the formation of long ssDNA resection products. (**D**) Alt-EJ occurs when shorter resection products are formed, resulting in variable length insertions and deletions. Taken from Ceccaldi et al. 2016.

1.5.1 Classical non-homologous end joining

Following the formation of a DSB, repair pathway choice is influenced by several criteria including the local nuclear environment, epigenetic factors, and the stage of the cell cycle (Densham and Morris 2019). C-NHEJ is the most prolific and rapid of the DSB repair sub-pathways, and repairs around 80% of all DSBs formed (**Figure 1.8A**) (Mao et al. 2008). Distinct from the other three pathways, c-NHEJ does not require end-resection, can occur at any stage of the cell cycle, and involves blunt-end ligation that is independent of sequence homology. This process initiates by first blocking the 5' DNA ends by the Ku70-Ku80 heterodimer (KU). This protein complex ensures that 5' DNA end-resection is prevented and acts as a scaffold, holding the DNA ends in close proximity to recruit the DNA-dependant protein kinase catalytic subunit (DNA-PKcs). DNA-PKcs phosphorylates multiple substrates of the DNA ligase complex such as ligase IV, XRCC4, and XLF, which together facilitate ligation of the broken ends (Krenning et al. 2019). C-NHEJ is error-prone as it lacks a DNA template for repair, often forming short insertions and deletions at the break sites, and occasionally forming translocations to other locations in the genome.

1.5.2 Homologous recombination

If the DNA end at the DSB becomes resected, leaving a 3' single-stranded DNA overhang, the three remaining resection-dependent pathways compete to repair the break. This process is initiated by the MRN (MRE11-RAD50-NBS1) complex, which recruits the C-terminal-binding protein interacting protein (CtIP) to the break site. During the first stage of end-resection the NBS1 subunit generates short single stranded 3' tails (≤20bp in mammals). Further 'extensive resection' commits repair to either the HR or SSA pathways by generating long 3' single stranded DNA (ssDNA) tails. This process is driven by exonucleases and helicases such as exonuclease 1 (EXO1), DNA replication ATP-dependant helicase/ nuclease DNA2 (DNA2), and bloom syndrome protein (BLM). HR predominates in the S and G2 phases of the cell cycle, where replication is highest and sister chromatids are available for error-free templated repair (**Figure 1.8B**) (Jackson and Bartek 2009). Long 3' ssDNA tails are typically highly unstable but are stabilised by rapid binding of replication protein A (RPA). This eliminates DNA secondary structure formation, prevents 3' ssDNA degradation and blocks spontaneous annealing of microhomologous sequences. Following this, RPA is replaced by DNA repair protein RAD51 homologue (RAD51) with the assistance of recombination mediators BRCA1/BRCA2 to form extended nucleoprotein filaments from the ssDNA. The ssDNA-RAD51 nucleofilament then promotes a homology search initiated by strand invasion at the homologous sequence of the sister chromatid (or a template donor strand in the case of genome editing), displacing one strand of DNA to form a D loop. D loop resolution can occur via several mechanisms with different outcomes, but generally cross-over structures (Holliday junctions) are formed and cleaved, resulting in the incorporation of homologous donor DNA at the break site. As a result, HR is considered to be largely error-free, as sequence is replaced by that originating from donor molecule and no genetic material is lost.

1.5.3 Single strand annealing and alternative end joining pathways

In contrast to HR, the two-remaining annealing-dependent pathways, SSA and alt-EJ, are error-prone and introduce insertions and deletions at the break site (Ceccaldi et al. 2016). Similarly to HR, SSA requires the formation of long ssDNA resection products in order to initiate. However, instead of the formation of a nucleofilament with RAD51, RAD52 aligns the RPA-coated ssDNA after end-resection (**Figure 1.8C**). SSA does not require a donor sequence or strand invasion for repair, and instead uses directed annealing of complementary resected sequences typically of around 30bp. Consequently, SSA produces long deletions, potentially up to hundreds of kilobase pairs long. The deletion size is dependent on the distance between the homologous repeats involved (Krenning et al. 2019). Alt-EJ shares features of both NHEJ and HR, and is reliant on the present of microhomologies (2-20nt) within resected 3' ssDNA ends (**Figure 1.8D**). Although alt-EJ has a similar initial resection mechanism to HR and SSA, extensive resection does not occur, and the strands are stabilised by PARP1. Annealing of microhomogolous sequences between the resected strands then follows, which can result in several different repair outcomes, including short insertions or deletions, depending on the repair factors involved in the process.

1.5.4 The influence of repair pathway on editing outcome

In summary, the highly variable repair outcome following DSB formation represents a double-edged sword when performing genome editing using CRISPR-Cas9. Understanding the mechanism of the editing process, in relation to repair pathway activity, is therefore fundamental to editing outcome. Owing to the preference of NHEJ over HR in mammalian cells, CRISPR has been used most widely to generate gene knockouts via the incorporation of indels in exon regions. This can result in a frame-shift and/or the formation of a premature stop codon, thus abolishing gene activity. Error-free, HR-based donor sequence insertion is significantly less efficient than NHEJ (25% vs 75%), but is a much more attractive mechanism for genome editing given the precise control over sequence alteration (Yang et al. 2020). The extent to which each of these four endogenous DSB repair pathways impact CRISPR editing outcome in different cellular contexts, is still not fully understood.

1.6 CRISPR-Cas9 off-target editing safety concerns

Depending on the context, CRISPR-Cas9 is either a precise pair of molecular scissors, or a dangerous genotoxic agent. For genome editing, a DSB is simply a means to an end to facilitate the desired editing outcome. In the DNA repair and genotoxicology fields, however, DSBs are uniquely characterised by the complete severing of both strands of the DNA duplex and are the most toxic of all DNA lesions. Unlike lesions confined to a single strand that are repaired by excision-based pathways, DSBs do not always possess an undamaged template to facilitate error-free repair (Schipler and Iliakis 2013). Other than directly resulting in cell death, failure to faithfully repair DSBs is an inherently carcinogenic process, resulting in a variety of structural genomic alterations such as deletions, insertions, DNA translocations, and mitotic recombination events in somatic cells (Khanna and Jackson 2001; Vilenchik and Knudson 2003). This phenomenon is most prominently demonstrated by cells which possess a decreased capacity to repair DSBs and can be clearly observed in individuals with inherited DNA damage response (DDR) defects. Examples of this include the loss-of-function mutations in the HR components BRCA1, BRCA2, RAD51C, RAD51D, and ATM, in familial forms of breast, ovarian, and pancreatic cancer (Lord and Ashworth 2012).

From a genotoxicological standpoint, assessing the safety of CRISPR-based therapies presents a unique and complex problem as DSB formation is not an unwanted consequence, but the mechanism of action. For small molecule drugs, direct DNA interactions leading to DSBs are considered unsafe, since just a single mutation at a relevant gene target such as an proto-oncogene or tumour suppressor gene can lead to the development of cancer (Fellows 2016). Further compounding the complexity of the problem, CRISPR-Cas9 can induce numerous ‘off-target’ DSBs throughout the genome in addition to the intended ‘on-target’ site. These sites share spacer and/or PAM sequence similarity, which enables the guide RNA component to erroneously direct the Cas9 nuclease to an off-target site, thus inducing unwanted DSBs at locations other than the target sequence. Not long after CRISPR-Cas9 genome editing was first demonstrated in eukaryotic cells, it was soon discovered that the system could cleave at off-target sites with mismatches in the spacer sequence of up to five nucleotides and even harbour alternative PAM sequences (Fu et al. 2013; Hsu et al. 2013; Pattanayak et al. 2013).

These early attempts at measuring this were based upon biased, targeted methods that identified off-target cleavage using computational prediction of sites with sequence similarity, followed by single-site cleavage assays (Tsai and Joung 2016). As an initial attempt to reveal off-target properties of CRISPR-Cas9, these studies demonstrated significantly high levels of off-target cleavage at many off-target sites, which in some cases were edited more efficiently than the original target site. Furthermore, different sgRNAs showed a range of off-targets from none detected (RNF2 and FANCF) up to 12 (VEGFA site 2), suggesting that certain sgRNAs suffered from more promiscuous cleavage than others (Fu et al. 2013). In addition to short indels, chromosomal translocations that formed between on- and off-target sites were observed following genome editing, highlighting the potential severity of the phenomenon (Cho et al. 2014; Ghezraoui et al. 2014). More recent studies have further demonstrated that severe genotoxic consequences such as large kilobase deletions, and complex rearrangements can arise from on-target cleavage in mitotic cells (Kosicki et al. 2018). For clinical applications, the induction of such transformations at on- or off-target sites at even low frequencies presents a significant problem as *ex vivo* and *in vivo* editing strategies require the modification of great quantities of cells, thus propagating rare, but inevitable off-target editing.

To enable the safe development of therapeutic genome editing, the gene and cell therapy community has established a three-step roadmap to test for CRISPR-associated genotoxic outcome and oncogenesis (Akcakaya et al. 2018; Cheng and Tsai 2018). The first step is the precise and accurate discovery of off-target editing events throughout the genome. The degree of CRISPR off-target activity can differ substantially in number and efficiency between different guide RNAs, is impacted by a variety of factors in a later chapter of this thesis. For these reasons, the full profile of off-target cleavage for any given therapeutic CRISPR-Cas9 guide RNA must be empirically determined prior to undertaking the second step of the roadmap, which involves assessing risk by evaluating the mutational consequences of off-target editing. Functional classification of off-target sites enables the elimination of high-risk guide RNAs with off-targets in protein coding genes, active regulatory regions, cell-type specific topologically associated domains (TADS), or within evolutionarily conserved sequences. If deemed an acceptable level of risk for the benefit, the final step of the roadmap involves *in vivo* validation of editing events and outcomes in animal or human-relevant model systems (Akcakaya et al. 2018; Cheng and Tsai 2018).

1.7 Measuring CRISPR-Cas9 off-targets genome-wide

As it is the first hurdle to overcome in the three-step roadmap for assessing the safety of CRISPR-Cas9 editing, defining the locations and frequency of CRISPR-induced DSBs is critically important in translating the technology into the clinic as a successful gene therapy strategy. In the last decade, there has been a substantial effort within the scientific community to address this problem, culminating in the development of numerous NGS-based methods that measure CRISPR-Cas9 function throughout the genome. These genome-wide approaches can be broadly separated in to two categories: 1. Cell-free approaches that measure Cas nuclease cleavage *in vitro*, and 2. Cell-based approaches that measure DNA DSBs that form in live cells. A summary of all current methods is presented in **Table 1.1**.

Table 1.1. Current NGS-based methods for detecting off-target DSB induction by CRISPR-Cas9.

Category	Method	Description	Reference
In vitro	Digenome-seq	*In vitro* Cas9-digested whole-genome sequencing.	(Kim et al. 2015)
	CIRCLE-seq	Cleavage of circularised DNA by Cas9 followed by ligation capture of exposed DNA ends.	(Tsai et al. 2017)
	DIG-seq	Adapted Digenome-seq methodology using cell-free chromatin DNA.	(Kim and Kim 2018)
	SITE-seq	Cleavage of HWM DNA by Cas9 followed by ligation capture of exposed DNA ends.	(Cameron et al. 2017)
	CHANGE-seq	Adapted CIRCLE-seq methodology that uses tagmentation based library preparation to improve throughput.	(Lazzarotto et al. 2020)
Cell-based Protein	ChIP-seq (γH2AX)	Chromatin immunoprecipitation and sequencing using DDR marker phosphorylated histone variant H2AX.	(Iacovoni et al. 2010)
	DISCOVER-seq	Chromatin immunoprecipitation and sequencing using DSB repair enzyme MRE11. Optimised for tissue samples to detect *in vivo* CRISPR editing.	(Wienert et al. 2019)
Cell-based Indirect repaired	IDLV capture	Off-target detection via the incorporation of lentiviral vectors in DSBs followed by NHEJ.	(Gabriel et al. 2011)
	HTGTS	Translocation capture of a known 'bait' DSB with unknown 'prey' DSBs.	(Chiarle et al. 2011)
	TC-seq	Translocation capture of a known 'bait' DSB with unknown 'prey' DSBs.	(Klein et al. 2011)
	LAM-HTGTS	Translocation capture of a known 'bait' DSB with unknown 'prey' DSBs using LAM-PCR.	(Hu et al. 2016)
	GUIDE-seq	Off-target detection via the incorporation of oligodeoxynucleotides (dsODN) in DSBs followed by NHEJ.	(Tsai et al. 2015)
	PEM-seq	Translocation capture of a known 'bait' DSB with unknown 'prey' DSBs using primer-extension.	(Yin et al. 2019)
	TEG-seq	GUIDE-seq methodology adapted for the Ion Torrent sequencing platform.	(Tang et al. 2018)
	TTISS	Adapted GUIDE-seq methodology that uses tagmentation based library preparation to improve throughput.	(Schmid-Burgk et al. 2020)
Cell-based Direct	dDIP	Damaged DNA immunoprecipitation by 3' end labelling of DSBs with biotinylated dNTPs	(Leduc et al. 2011)
	DSB-seq	Damaged DNA immunoprecipitation by 3' end labelling of DSBs with biotinylated dNTPs	(Baranello et al. 2014)

Unrepaired	BREAK-seq	dDIP labelling performed on embedded chromosomal DNA in agarose beads to reduce artificial DSBs	(Hoffman et al. 2015)
	BLESS	Blunt end ligation capture of breaks *in situ* inside crosslinked and permeabilised cells using biotinylated linker DNA.	(Crosetto et al. 2013)
	END-seq	Adaptation of BLESS. TA ligation capture of breaks *in situ in* cells embedded in agarose plugs using biotinylated T-tailed hairpin sequencing adapter.	(Canela et al. 2016)
	DSBCapture	Adaptation of BLESS. TA ligation capture of breaks in situ inside crosslinked and permeabilised cells using biotinylated full-length T-tailed sequencing adapter.	(Lensing et al. 2016)
	BLISS	Adaptation of BLESS. TA ligation capture of breaks *in situ* inside crosslinked and permeabilised cells bound to a solid surface using T-tailed sequencing adapter.	(Yan et al. 2017)
	i-BLESS	Adaptation of BLESS. TA ligation capture of breaks *in situ in* cells embedded in agarose beads using biotinylated T-tailed hairpin sequencing adapter.	(Biernacka et al. 2018)

1.7.1 *In vitro* methods to detect genome-wide CRISPR-Cas9 off-target cleavage

The first group of methods used to measure CRISPR-Cas9 off-targets are cell-free assays that detect nuclease cleavage *in vitro*. These assays differ from cell-based approaches as they are exclusively designed to measure nuclease-induced cleavage events genome-wide and cannot be used to measure physiological DSB events in cells. The first, and most simple of these approaches is digested genome sequencing (Digenome-seq), which involves whole genome sequencing (WGS) of Cas9-cleaved genomic DNA (Kim et al. 2015). Rather than relying on experimental means to separate and enrich for genomic DNA fragments, arising as a result of Cas9 cleavage, Digenome-seq uses high coverage WGS (~500 million reads per sample) and bioinformatics analysis to define off-target sites based on recurrent read start position. Sequencing reads originating from Cas9-cleaved DNA fragments share the same start and/or end position, and thus are detectable above the 'background' of randomly fragmented DNA during WGS library preparation. The benefits of this approach are its unbiased nature, and relatively simple experimental design, which can be simultaneously used to detect indels if Cas9-edited DNA is used for *in vitro* digestion. As WGS is being performed, many DNA fragments that are unrelated to Cas9 cleavage are also sequenced, resulting in a very high cost of sequencing to identify low frequency editing events. More recently, Digenome-seq was adapted in the form of the method DIG-seq, which follows the same experimental design except that it uses chromatin rather than extracted genomic DNA to represent the nuclear environment more accurately (Kim and Kim 2018).

In an effort to improve the sensitivity and reduce the cost of off-target detection, the other *in vitro* approaches omit WGS in favour of experimental steps to enrich for Cas9 cleaved DNA. The first of these methods is SITE-seq, which involves biochemical cleavage of high molecular weight (HMW) DNA followed by ligation of a biotin-labelled sequencing adapter to the exposed DNA ends (Cameron et al. 2017). The resulting labelled DNA ends are separated via a biotin-streptavidin interaction and prepared for sequencing, resulting in an output corresponding uniquely to the Cas9-cleaved DNA fragments. Using HMW DNA for initial cleavage, reduces the number of mechanically-induced break sites generated during the purification procedure. This significantly reduces the read depth required for sensitive off-target detection. CIRCLE-seq and its more recent derivative CHANGE-seq, circumvent this problem in a similar manner, and generate a sample of fragmented and circularised genomic DNA for Cas9 digestion (Tsai et al. 2017; Lazzarotto et al. 2020). Following cleavage, the circular DNA becomes linearised, which enables library preparation and enrichment of the fragments which were cleaved post-circularisation.

Generally, *in vitro* approaches are more sensitive than cell-based ones as the off-target cleavage by Cas9 is not affected by chromatin structure, subnuclear location or low concentration of enzyme. *In vitro* methods are thought to reveal the ‘superset’ of off-targets for any given guide RNA, and have the potential to identify even the weakest of off-target effects; direct sequencing of sites cleaved *in vitro* using the method CIRCLE-seq revealed >10-fold higher number of off-targets for some sgRNAs than using the cell-based approach GUIDE-seq (Tsai et al. 2017). Despite this enhanced sensitivity, the significance of these off-targets has been questioned, since many may simply represent false positives that would not be cleaved in a cellular environment (Wienert et al. 2019). Furthermore, the lack of a cellular environment may also cause *in vitro* approaches to miss bona fide off-targets that are generated indirectly through Cas9/sgRNA cell treatment, or occur directly as a result of the dynamic chromatin environment. Indeed, factors such as DNA duplex destabilisation through DNA stretching and bubble formation have been shown previously to increase the susceptibility for CRISPR off-target binding, increasing nucleotide mismatch tolerance to as many as 10bp (Newton et al. 2019).

1.7.2 Cell-based methods to detect CRISPR-Cas9 off-target DSBs

The second group of approaches capable of detecting CRISPR-Cas9 off-targets involve measuring DNA DSBs in cells. These methods can be broadly subclassified in to three categories based on the mode of break detection: 1. indirect break mapping using proteins as a proxy, 2. indirect mapping of repaired breaks and 3. direct mapping of unrepaired breaks (**Table 1.1**. **and Figure 1.9**) (Bouwman and Crosetto 2018). The pros and cons of each of these approaches will be discussed in the following sections.

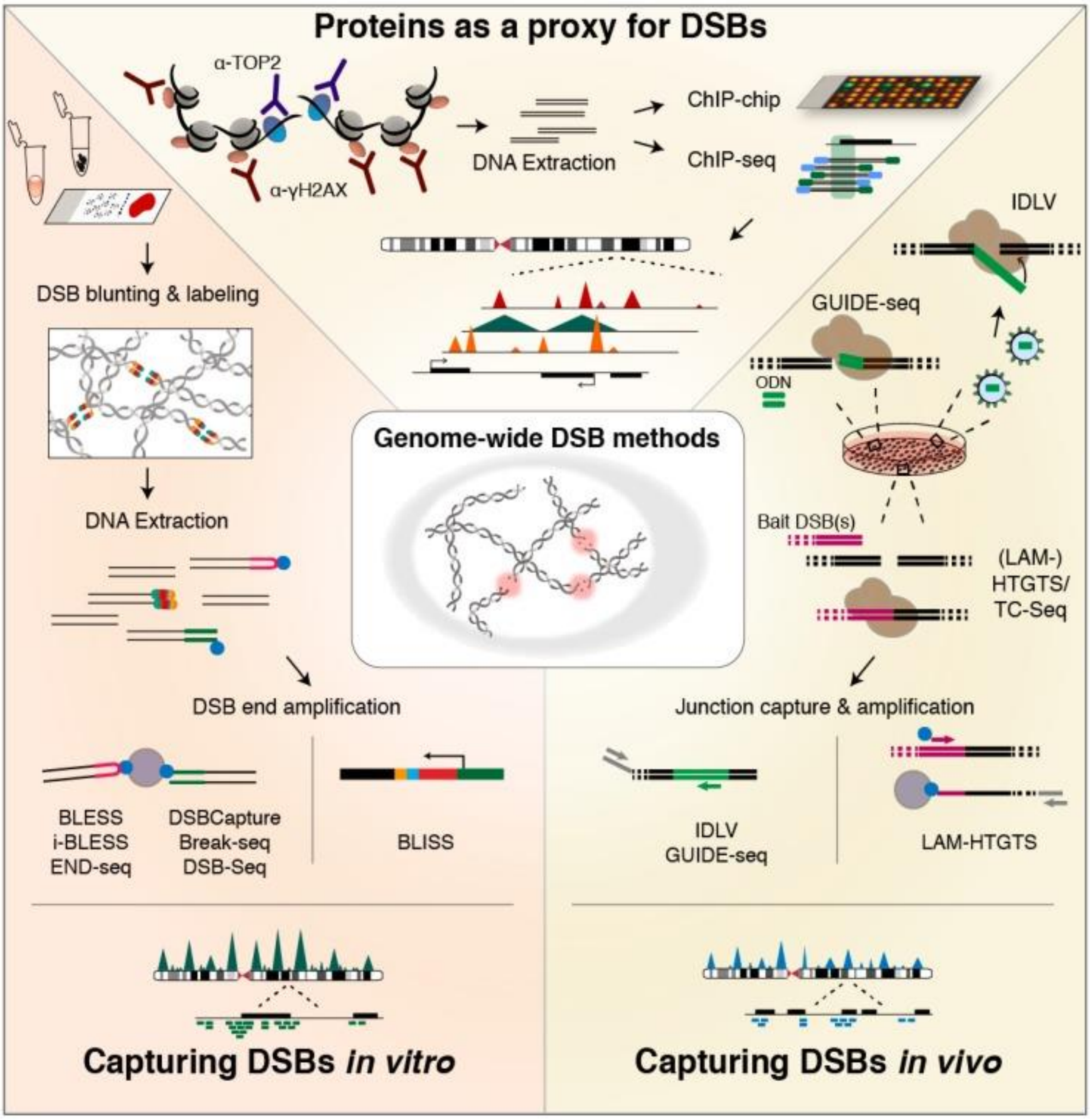


Figure 1.9. Different methodologies to detect genomic DSBs. (**Top**) Proteins used as a proxy for DSBs are captured via immunoprecipitation and DNA that was bound can be analysed using a microarray (ChIP-chip) or NGS (ChIP-seq)**.** (**Right**) Indirect methods to capture repaired DSBs in live cells using NGS. GUIDE-seq and IDLV detect NHEJ mediated integration of short oligonucleotides. LAM-HTGTS and TC-seq assess chromosomal translations using induced bait DSBs at known positions. (**Left**) Direct methods for labelling and enriching break ends in vitro for NGS. Breaks are labelled with biotinylated adapters for enrichment on streptavidin (BLESS, i-BLESS, END-seq, DSBCapture, Break-seq, DSB-seq), or are labelled with an adapter baring the T7 promoter sequence to mediate enrichment by in vitro transcription (BLISS). Taken from: Bouwman and Crosetto 2018*.*

The first of these approaches uses proteins recruited and bound to DSBs to indirectly measure break formation, generating a snapshot of the DSB repair profile (**Figure 1.9, top**). The next approach focuses on using the DSB repair endpoint to measure breaks which accumulate over time in live cells (**Figure 1.9, right**). The last group of methods directly detects unrepaired breaks by immobilising cells/chromatin and captures the exposed ends with a variety of labelling strategies (**Figure 1.9, left**). Unlike *in vitro* methods, cell-based approaches offer the advantage of measuring off-target activities in a directly relevant cell-based system. This enables the discovery of CRISPR-induced off-targets under a variety of experimental conditions. Studies investigating the effects of modified CRISPR proteins and sgRNAs on off-target generation, chromatin environment on break formation and repair, and kinetic analysis using time-course sampling, are all possible using one, or a combination of these approaches. Another advantage of using cell-based methods for detecting DSBs, rather than measuring enzyme cleavage *in vitro,* is that they can simultaneously detect physiological DSBs formed during normal cellular processes such as DNA replication, transcription, chromatin looping, and V(D)J recombination. CRISPR-induced genotoxicity may extend beyond off-targets resulting from guide RNA mismatching, thus defining the natural DSB 'hotspots' and break patterns found in cells could direct more sophisticated target-selection during guide RNA design.

1.7.3 Indirect break mapping using proteins

The first attempts to measure DSBs genome-wide in cells did so using proteins bound to chromatin as a proxy for breaks (**Figure 1.9, top**). These methods utilise chromatin immunoprecipitation (ChIP) of the proteins recruited to breaks during repair, or those that are involved in break formation in order to detect DSB locations. DNA sequences bound to the enriched proteins are then detected by either microarray hybridisation (ChIP-chip) or by high-throughput sequencing (ChIP-seq). Phosphorylated histone variant H2AX (γH2AX), an early response proxy for DNA damage in eukaryotes, has been exploited extensively for this purpose, and was used to generate one of the first genome-wide maps of DNA break induction in mammalian cells (Iacovoni et al. 2010). This marker however presents significant limitations that hinder its use for unbiased and accurate DSB detection. H2AX phosphorylation spreads less efficiently through heterochromatin and is not uniformly distributed throughout the genome. In addition to this, the signal can spread away from lesions up to 2Mb and accumulate at sites of single strand breaks (SSB), further limiting accurate break mapping.

More recently, a novel approach named DISCOVER-seq has also exploited a component of the DDR: MRE11, which promotes 5’ end resection during DSB repair (Wienert et al. 2019). Following CRISPR-Cas9 treatment of cells and mice, ChIP-seq of MRE11 revealed numerous CRISPR-induced breaks at single nucleotide resolution. For protein based DSB detection, this represents a significant improvement over γH2AX-based assays, addressing the lack of precision and biases generated during γH2AX mapping. Although effective for detecting high frequency CRISPR-induced DSBs *in vivo*, DISCOVER-seq was less sensitive than another method GUIDE-seq for detecting DSBs induced in cell culture samples. In contrast to mapping protein binding, GUIDE-seq maps DSBs indirectly using the outcome of repair as the readout. This is discussed in the next section.

1.7.4 Indirect mapping of repaired breaks

To overcome the limitations of protein-associated DSB measurement, several methods have been developed to detect genomic DSBs by utilising DSB repair mechanisms to capture the breaks accumulated in live cells (**Figure 1.9, right**). Generally, these approaches measure either NHEJ-mediated insertion of exogenous DNA, or chromosomal translocations formed following editing.

The integration-defective lentiviral vector (IDLV) capture method was the first NHEJ-based genomic DSBs capture technology, and works via the integration of a lentiviral vector at the DNA break site. Following IDVL insertion, linear amplification-mediated PCR (LAM-PCR) using a primer that exclusively binds to the IDLV sequence enriches for all sites where the vector has inserted. Sequencing of the resulting DNA fragments then reveals the genome-wide landscape of DSBs repaired by NHEJ (Gabriel et al. 2011; Wang et al. 2015). IDLV capture was a promising early attempt to capture breaks generated and repaired in live cells and was applied to the study of genomic breaks induced by CRISPR-Cas9 and TALENS. IDLV capture is estimated to detect breaks with a frequency as low as 1%. Despite this, limitations of the method include: sequence bias, low numbers of informative reads, integration at varying distances from the break site, and high cost (Hu et al. 2016). Improving on IDLV capture, genome-wide unbiased identification of DSBs enabled by sequencing (GUIDE-seq) was developed to address these limitations. Rather than using IDLV dsDNA, DSBs are labelled by GUIDE-seq via NHEJ-mediated integration of short dsDNA oligodeoxynucleotides (dsODN), which are inserted with greater efficiency than IDLVs. Similarly, the dsODN tag serves as a priming site for amplification and sequencing. GUIDE-seq was applied to the study of CRISPR-

Cas9 induced DSBs, revealed hundreds of CRISPR off-targets for a range of guide RNAs throughout the genome that had accumulated over the course of several days (Tsai et al. 2015).

Further development of the GUIDE-seq methodology resulted in novel iterations including: target-enriched GUIDE-seq (TEG-seq) and tagmentation-based tag integration site sequencing (TTISS). These approaches exploit a similar DSB-labelling methodology via oligonucleotide integration, and were developed to enhance assay efficiency. TEG-seq is an adapted version of GUIDE-seq (which was designed using Illumina sequencing platforms) that is compatible with Ion Torrent sequencing, and contains an additional target enrichment step to increase the proportion of sequencing reads originating from the inserted dsODN (Tang et al. 2018). TTISS on the other hand, maintains Illumina sequencing compatibility, but incorporates a tagmentation-based library preparation to vastly increase sample throughput and scalability (Schmid-Burgk et al. 2020). Both of these approaches were used to screen multiple guide RNAs under a variety of experimental conditions. However, they did not address the critical limitations that arise owing to the method of DSB capture via NHEJ integration. Efficient off-target capture via these approaches is directly constrained by the DNA repair proficiency of the cell type used, limiting the study of off-target break induction in some cell types and genetic backgrounds. Furthermore, the requirement of efficient transfection of the dsODN means GUIDE-seq and related assays cannot be applied to *in vivo* and *ex vivo* genome editing, or studies with hard to transfect cell types.

As described previously, another way to identify DSB formation is to measure the structural alterations that arise following DNA repair. Translocation capture sequencing (TC-seq) and high-throughput genome-wide translocation sequencing (HTGTS) were developed for this purpose. Both methods were developed to detect genomic translocation junctions following DSB events using a ‘bait’ DSB induced at an integrated I-SceI meganuclease recognition sequence. The following DSB is prone to forming translocations with genome-wide ‘prey’ DSBs. Bait-prey junctions are then enriched by nested-PCR from the bait sequence and prepared for sequencing (Chiarle et al. 2011; Klein et al. 2011). Several further iterative developments include the LAM-HTGTS technique which substitutes conventional PCR with LAM-PCR, and primer-extension mediated sequencing (PEM-seq), which combines LAM-HTGTS with primer-extension mediated sequencing (Hu et al. 2016; Yin et al. 2019). Collectively, these methods have been applied to identify translocation junctions in B lymphocytes induced by IgH class switching, DSB hotspots in neural stem/progenitor cells, and off-target breaks induced by CRISPR-Cas9 (Hu et al. 2016; Schwer et al. 2016; Yin et al. 2019). Translocation capture, however, is not ideal for CRISPR-Cas9 off-target detection as prior knowledge is required to design primers that hybridise at the ‘bait’ DSB site. Furthermore, capture is prone to a proximity bias, more efficiently capturing DSBs in regions that are topologically associated. Lastly, translocation sequencing methods will only identify DSBs that are prone to forming translocations, underestimating the absolute DSB frequency.

When considering which method to choose for CRISPR-Cas9 off-target detection, the major advantage of determining DSB locations using repair outcomes is the ability to measure DSBs accumulated over time. Compared with indirect ChIP-seq based methods, repair-based break capture offers a more sensitive and direct approach with the potential to measure break points with improved resolution. Despite this, the fundamental limitation of requiring active NHEJ for efficient break capture means they are unsuitable for a broad variety of applications or cell types, detecting only recurrent NHEJ-repaired breaks, and missing breaks repaired via other pathways.

1.7.5 Direct mapping of unrepaired breaks

The final group of DSB detection methods involve direct labelling of exposed DSB ends to capture a snapshot of unrepaired DSBs throughout the genome (**Figure 1.9, left**). Early attempts to do this by damaged DNA immunoprecipitation (dDIP) and DSB-seq, involved 3’ end labelling of DSBs with biotinylated dNTPs in isolated high-molecular-weight DNA, followed by immunoprecipitation, streptavidin enrichment, and sequencing (Leduc et al. 2011; Baranello et al. 2014). The process of extracting HMW DNA however, may introduce high levels of artificial DSBs. BREAK-seq was developed in order to address this limitation, embedding chromosomal DNA in agarose prior to 3‘ end labelling. Despite this, these methods are still limited by 3’ end labelling which is inherently variable at different end structures, and therefore results in disproportionate break detection (Vitelli et al. 2017).

The method called Breaks Labelling, Enrichment on Streptavidin and Sequencing (BLESS) was designed to address the limitation of 3’ end labelling by using ligation to capture DSB ends (Crosetto et al. 2013). The key innovation behind BLESS was to perform all break labelling steps *in situ* inside the nuclear environment, thus eliminating false-positive DSBs arising from DNA extraction. Following formaldehyde crosslinking, cells are permeabilised and DSB ends are blunted and phosphorylated *in situ* prior to ligation with a biotinylated hairpin adapter. DNA is then extracted, and adapter ligated DNA fragments are enriched on streptavidin, ligated with a second sequencing adapter, amplified by PCR, and finally sequenced by NGS. In this manner BLESS has been used to identify the location of fragile sites associated with replication fork stalling, telomere ends, and genome-wide CRISPR-Cas9 induced DSBs (Crosetto et al. 2013; Slaymaker et al. 2016). Despite the advances made by BLESS, several significant drawbacks exist such as: inefficient ligation of break ends, low sequence diversity as a substantial amount of each sequencing read is taken up by the BLESS adapter, high levels of background noise, and a labour-intensive protocol. These limitations have prevented widespread adoption of the method (Lensing et al. 2016).

In recent years BLESS has been used as a benchmark for the development of several other methods based on *in situ* ligation tagging of breaks using adapters. DSBCapture and END-seq, published in the same year, were the first to follow. Both enhanced the efficiency of break capture with the introduction of A-tailing of break ends and the use of a T-tailed adapter. Unlike BLESS, where a proximal adapter is used at the site of the break and ligated with sequencing adapters in a later step, the adapters used to tag breaks in DSBCapture and END-seq contain the first subunit of the Illumina TruSeq adapter. This enables sequencing to initiate at the break site directly rather than at the proximal adapter, thus improving sequence diversity and simplifying the procedure (Lensing et al. 2016). In addition to these changes, END-seq also removes the need for formaldehyde crosslinking by embedding cells in agarose plugs prior to end labelling, supposedly reducing formaldehyde induced breaks and background noise (Canela et al. 2016). As with BLESS, after break tagging using biotinylated adapters, both methods enrich for DSB ends on streptavidin and follow with PCR amplification and sequencing. These improvements implemented by DSBCapture, greatly enhanced the yield of definable DSBs; in a side-by-side experiment, DSBCapture identified 4.5-fold more DSBs than BLESS. Furthermore, in a direct comparison with GUIDE-seq, DSBCapture identified 100-fold more endogenous DSBs in U2OS cells, owing primarily to the differences in sensitivity between direct vs indirect repair-based break capture (Lensing et al. 2016). END-seq is also reported to outperform BLESS. Using AsiSI restriction endonuclease inducible pre-B cells, END-seq found a 319-fold increase in the number of reads at the AsiSI recognition sequence, and a 36-fold increase in the proportion of reads mapped to AsiSI sites (Canela et al. 2016).

Following DSBCapture and END-seq, Breaks Labelling In-Situ and Sequencing (BLISS) was developed to address the original limitations of BLESS and biotin: streptavidin enrichment. BLESS and DSBCapture require >10^6 cells, and END-seq requires >10^7, limiting experimental scalability. BLISS addresses this issue by enriching for break ends with linear amplification rather than streptavidin mediated pull down. DSBs are tagged with an adapter containing the T7 promoter sequence, which allows linear amplification of DSB ends by *in vitro* transcription. Furthermore, the adapter also contains the Illumina

RA5 adapter sequence followed by an 8-12bp unique molecular identifier (UMI) to allow better control for PCR amplification biases and allows the quantification of the absolute number of captured DSB ends at a specific position. The cell number required is further reduced by immobilising cells on to a solid surface, increasing experimental throughput, reducing risk of artificial DSBs through manipulation, and making the process compatible with tissue samples (Yan et al. 2017). BLISS has been applied to the study of induced and endogenously formed DSBs in cell culture and tissue samples, in addition to the identification of CRISPR-Cas9 induced DSBs.

Innovating in a different way to BLISS, the method i-BLESS was developed to address several of the limitations of the BLESS and END-seq approaches. Similarly to END-seq, i-BLESS takes an approach that eliminates the need for cell crosslinking prior to *in situ* break labelling by embedding cells in agarose beads (Biernacka et al. 2018). One difficulty associated with the END-seq process is the lengthy experimental procedure caused by the requirement of reagent diffusion into the agarose plugs, which i-BLESS improves on by using agarose beads. Using the enhanced protocol, iBLESS demonstrates improved sensitivity over END-seq, detecting a single DSB in 100,000 cells (compared to one in 10,000 by END-seq). Major limitations of iBLESS, however are that it is optimised for use in yeast and may require further optimisation for mammalian cells. Furthermore, in stark contrast to the innovations made by the BLISS method, i-BLESS requires a greater number of input cells (>10^9), which significantly limits sample throughput.

1.8 Thesis aims

There are now numerous different methods to measure and characterise CRISPR-Cas9 off-target events in the genome, with each assay having its own advantages and caveats depending on the capture and enrichment methodology. None of these methods provide a perfect solution, and no ‘gold-standard’ assay has yet been defined. At present, it is considered that no assay is capable of providing a definitive or fully comprehensive list of off-targets. Furthermore, it has been suggested that the problem necessitates the use of several partially redundant methods to provide an acceptable solution to the accurate identification of CRISPR-Cas9 off-targets (Tsai and Joung 2016). Indeed, combining a repair-based approach such as GUIDE-seq with a direct break labelling approach such as BLISS could enable the thorough characterisation of CRISPR off-target activities genome-wide (Bouwman and Crosetto 2018). Given this, it is generally accepted that none of the currently available methods, when used in isolation, provide a means to accurately quantify off-target genome editing in cells. An ideal assay would provide accurate, unbiased, and sufficiently sensitive DSB detection in a cell-based system while being scalable, simple and cost-effective. For reproducible discovery of off-targets in limited clinical samples, or for assessing many guide RNAs in parallel, scalability becomes particularly important as cell material becomes limiting and running multiple analogous assays is not feasible.

Given its inherently mutagenic mechanism of action, for the translation of *any* CRISPR-based therapeutic modality, the consideration of patient safety must remain at the core of its development. The establishment of a novel method for the accurate, scalable and unbiased detection of DSBs genome-wide, would therefore represent a major advance that would enable the safe and sustainable development of CRISPR-Cas9 based cell and gene therapies. The aim of this thesis is therefore to address the limitations described previously and to develop a novel DSB measurement tool that is capable of quantitatively measuring CRISPR-Cas9 off-target activity throughout the genome. By exploiting such a method, a data-driven approach based on accurate, real-world data can be taken to fully understand the determinants of off-target genome editing. This will ultimately pave the way towards safer cell and gene therapies.

2 Chapter II - Materials and methods

2.1 Wet laboratory methods

2.1.1 Mammalian cell culture and treatments

Human embryonic kidney cells (HEK293) were obtained courtesy of Dr Mick Fellows (AstraZeneca, Cambridge, UK) and HEK293T cells were purchased from ATCC (CRL-3216™). U2OS DIvA cells were obtained courtesy of Professor Jessica Downs (Institute of Cancer Research, London, UK), and originate from the Gaëlle Legube Laboratory (Centre de Biologie Intégrative, Toulouse, France) (Iacovoni et al. 2010). Cells were cultured in DMEM (Life Technologies) supplemented with 10% FBS (Life Technologies) at 37˚C at 5% CO_2. Cells were routinely passaged at ~90% confluence every 3-5 days and reseeded at a density of ~1x10^5/ml. HEK293 cells were nucleofected with 224 pmol CRISPR-Cas9 RNP per 3.5x10^5 cells using a Lonza 4D-Nucleofector X unit with pulse code CM-130. Cells were harvested at 0, 7, 12, 24, and 30 hours post nucleofection for INDUCE-seq processing. To stimulate AsiSI-dependent DSB induction, DIvA cells were treated with 300 nM 4-hydroxytamoxifen (4OHT) (Sigma, H7904) for 4 h.

2.1.2 Cas9 protein and guide RNA

The guide RNA targeting EMX1 (GAGTCCGAGCAGAAGAAGAA) was synthesized as a full-length non-modified single guide RNA oligonucleotide (Synthego). Cas9 protein was produced in-house (AstraZeneca, Gothenburg, Sweden) and contained an N-terminal 6xHN tag.

2.1.3 INDUCE-seq adapters

All modified INDUCE-seq adapter oligonucleotides were purchased from IDT. Single stranded oligonucleotides were annealed at a final concentration of 10 µM in Nuclease-free Duplex Buffer (IDT, 11-01-03-01) by heating to 95˚C for 5 minutes and slowly cooling to 25˚C using a thermocycler. All adapter oligonucleotide sequences are shown in **Appendix A, Table A1**.

2.1.4 Confirmation of adapter ligation activity *in vitro*

Ligations were performed by incubating 0.4µM double stranded P5 and P7 adapter variants with 1,000 units of T4 DNA ligase (NEB, M0202M), 0.8mM ATP (NEB, P0756S), and 50µg Ultrapure BSA (Invitrogen, AM2618), in a final volume of 50 µL 1x T4 DNA ligase buffer. Ligations were incubated overnight at 16˚C and DNA fragment sizes checked using an Agilent 2200 TapeStation using a D1000 ScreenTape (Agilent Technologies Ltd).

2.1.5 INDUCE-seq cell preparation and acquisition

Cell samples were prepared for INDUCE-seq processing by first seeding to 96-well plates pre-coated with Poly-D-lysine (Greiner bio-one, 655940) at a density of 1x10^5/well. Samples were crosslinked in 4% PFA (Pierce, 28908) and stored in 1x PBS at 4˚C. HEK293, HEK293T, and U2OS DIvA cell samples were prepared in the Reed laboratory. Induced pluripotent stem cells, neural progenitor cells, and mature neurons were provided by Dr Ian Tully (Cardiff University, UK) (Tully 2020). RPE1 cells were provided by Professor Jessica Downs (The Institute of Cancer Research, UK). C42 prostate cancer cells were provided by Dr Claire Fletcher (Imperial College London, UK).

2.1.6 *In situ* DSB induction with HindIII

In situ enzymatic induction of DSBs in HEK293T cells was performed using the restriction enzyme HindIII-HF® (NEB, R3104S). This process was the same as described for the full INDUCE-seq method, with the addition of DSB induction prior to end blunting. Following cell permeabilization DSBs were induced using 50U HindIII-HF® in 1x CutSmart® Buffer in a final volume of 50 µL. Digestions were performed at 37˚C for 18 hours.

2.1.7 Proof of concept INDUCE-seq sample processing and sequencing

The INDUCE-seq method was initiated by permeabilising cells. Between incubation steps, cells were washed in 1x PBS at rt. Cells were permeabilised by incubation in Lysis buffer 1 (10 mM Tris-HCL pH 8, 10 mM NaCl, 1 mM EDTA, 0.2% Triton X-100, pH 8 at 4˚C) for one hour at rt, followed by incubation in Lysis buffer 2 (10 mM Tris-HCL, 150 mM NaCl, 1 mM EDTA, 0.3% SDS, pH 8 at 25˚C) for one hour at 37˚C. Permeabilised cells were washed three times in 1x CutSmart® Buffer (NEB, B7204S) and blunt-end repaired using NEB Quick Blunting Kit (E1201L) + 100 µg/mL BSA in a final volume of 50 µL at rt for one hour. Cells were then washed three times in 1x CutSmart® Buffer and A-tailed using NEBNext® dA-Tailing Module (NEB, E6053L) in a final volume of 50 µL at 37˚C for 30 mins. A-tailed cells were washed three times in 1x CutSmart® buffer then incubated in 1x T4 DNA Ligase Buffer (NEB, B0202S) for 5 mins at rt. A-tailed ends were labelled by ligation using T4 DNA ligase (NEB, M0202M) + 0.4 µM Modified P5 adapter in a final volume of 50 µL at 16˚C for 16-20 h. Following ligation, excess P5 adapter was removed by washing cells 10 times in wash buffer at rt (10 mM Tris-HCL, 2 M NaCl, 2 mM EDTA, 0.5% Triton X-100, pH 8 at 25˚C), incubating for 2 mins each wash step. Cells were washed once in PBS and then once in nuclease free H_2O (IDT, 11-05-01-04). Genomic DNA was extracted by incubating cells in DNA extraction buffer (10mM Tris-HCL, 100mM

NaCl, 50mM EDTA, 1.0% SDS, pH 8 at 25˚C) + 1mg/ml Proteinase K (Invitrogen, AM2584) in a final volume of 100µl for 5 mins at rt. The cell lysates were transferred to 1.5ml Eppendorf RNA/DNA LoBind tubes (Fisher Scientific, 13-698-792) and incubated at 65˚C for 1 hour, shaking at 800rpm. DNA was purified using Genomic DNA Clean & Concentrator™-10 (Zymo Research, D4010). DNA yield was assessed using 1 µL sample and Qubit DNA HS Kit (Invitrogen, Q32854) before proceeding to library preparation. Genomic DNA was fragmented to 300-500bp using a Bioruptor Sonicator, and size selected using SPRI beads (GC Biotech, CNGS-0005) to remove fragments <150bp. Library preparation was performed using the with-bead approach (Fisher et al. 2011), using the 0.4 µM Modified half-functional P7 adapter, the NEB Quick Blunting Kit (E1201L), NEBNext® dA-Tailing Module (NEB, E6053L), and T4 DNA ligase (NEB, M0202M). The ligated sequencing libraries were purified using SPRI beads. Libraries were purified twice more using SPRI beads, and size selected to remove fragments <200bp to remove residual adapter DNA. Final clean libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina® Platforms (Roche, 07960255001). Final library DNA fragment distributions were checked using an Agilent 2200 TapeStation using a D1000 ScreenTape (Agilent Technologies Ltd). Samples were pooled and concentrated to the desired volume for sequencing using a SpeedVac. Sequencing was performed using a MiSeq Reagent Nano Kit v2 (300-cycles) (Illumina Inc. MS-103-1001).

2.1.8 Final INDUCE-seq sample processing and sequencing

The INDUCE-seq method was initiated by permeabilising cells. A schematic of the main steps involved in the INDUCE-seq procedure is shown in **Figure 2.1**. Between incubation steps, cells were washed in 1x PBS at rt. Cells were washed in 1x PBS to remove formaldehyde and stored at 4˚C for up to 30 days. The INDUCE-seq method was initiated by permeabilising cells. Between incubation steps, cells were washed in 1x PBS at rt. Cells were permeabilised by incubation in Lysis buffer 1 (10 mM Tris-HCL pH 8, 10 mM NaCl, 1 mM EDTA, 0.2% Triton X-100, pH 8 at 4˚C) for one hour at rt, followed by incubation in Lysis buffer 2 (10 mM Tris-HCL, 150 mM NaCl, 1 mM EDTA, 0.3% SDS, pH 8 at 25˚C) for one hour at 37˚C. Permeabilised cells were washed three times in 1x CutSmart® Buffer (NEB, B7204S) and blunt-end repaired using NEB Quick Blunting Kit (E1201L) + 100 µg/mL BSA in a final volume of 50 µL at rt for one hour. Cells were then washed three times in 1x CutSmart® Buffer and A-tailed using NEBNext® dA-Tailing Module (NEB, E6053L) in a final volume of 50 µL at 37˚C for 30 mins. A-tailed cells were washed three times in 1x CutSmart® buffer then incubated in 1x T4 DNA Ligase Buffer (NEB, B0202S) for 5 mins at rt. A-tailed ends were labelled by ligation using T4 DNA ligase (NEB, M0202M) + 0.4 µM Modified P5 adapter in a final volume of 50 µL at 16˚C for 16-20 h. Following ligation, excess P5 adapter was removed by washing cells 10 times in wash buffer at rt (10 mM Tris-HCL, 2 M NaCl, 2 mM EDTA, 0.5% Triton X-100, pH 8 at 25˚C), incubating for 2 mins each wash step. Cells were washed once in PBS and then once in nuclease free H_2O (IDT, 11-05-01-04). Genomic DNA was extracted by incubating cells in DNA extraction buffer (10 mM Tris-HCL, 100 mM NaCl, 50 mM EDTA, 1.0% SDS, pH 8 at 25˚C) + 1 mg/mL Proteinase K (Invitrogen, AM2584) in a final volume of 100 µL for 5 mins at rt. The cell lysates were transferred to 1.5 mL Eppendorf RNA/DNA LoBind tubes (Fisher Scientific, 13-698-792) and incubated at 65˚C for 1 hour, shaking at 800rpm. DNA was purified using Genomic DNA Clean & Concentrator™-10 (Zymo Research, D4010), and eluted using 100 µL Elution Buffer. DNA yield was assessed using 1 µL sample and Qubit DNA HS Kit (Invitrogen, Q32854) before proceeding to library preparation. Genomic DNA was fragmented to 300-500bp using a Bioruptor Sonicator, and size selected using SPRI beads (GC Biotech, CNGS-0005) to remove fragments <150bp. Fragmented and size-selected DNA was end-repaired using NEBNext® Ultra™ II DNA Module (NEB, E7546L). Fragmented and end-repaired DNA was added directly to the ligation reaction using NEBNext® Ultra™ II Ligation Module (NEB, E7595L) according to the manufacturer's instructions using 7.5 µM Modified half-functional P7 adapter and omitting USER enzyme addition. The ligated sequencing libraries were purified using SPRI beads. Libraries were purified twice more using SPRI beads, and size selected to remove fragments <200bp to remove residual adapter DNA. Final clean libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina® Platforms (Roche, 07960255001). Final library DNA fragment distributions were checked using an Agilent 2200 TapeStation using a D1000 ScreenTape (Agilent Technologies Ltd). Samples were pooled and concentrated to the desired volume for sequencing using a SpeedVac. Sequencing was performed using a NextSeq 500/550 High Output Kit v2.5 (75 Cycles) (Illumina Inc. 20024906)

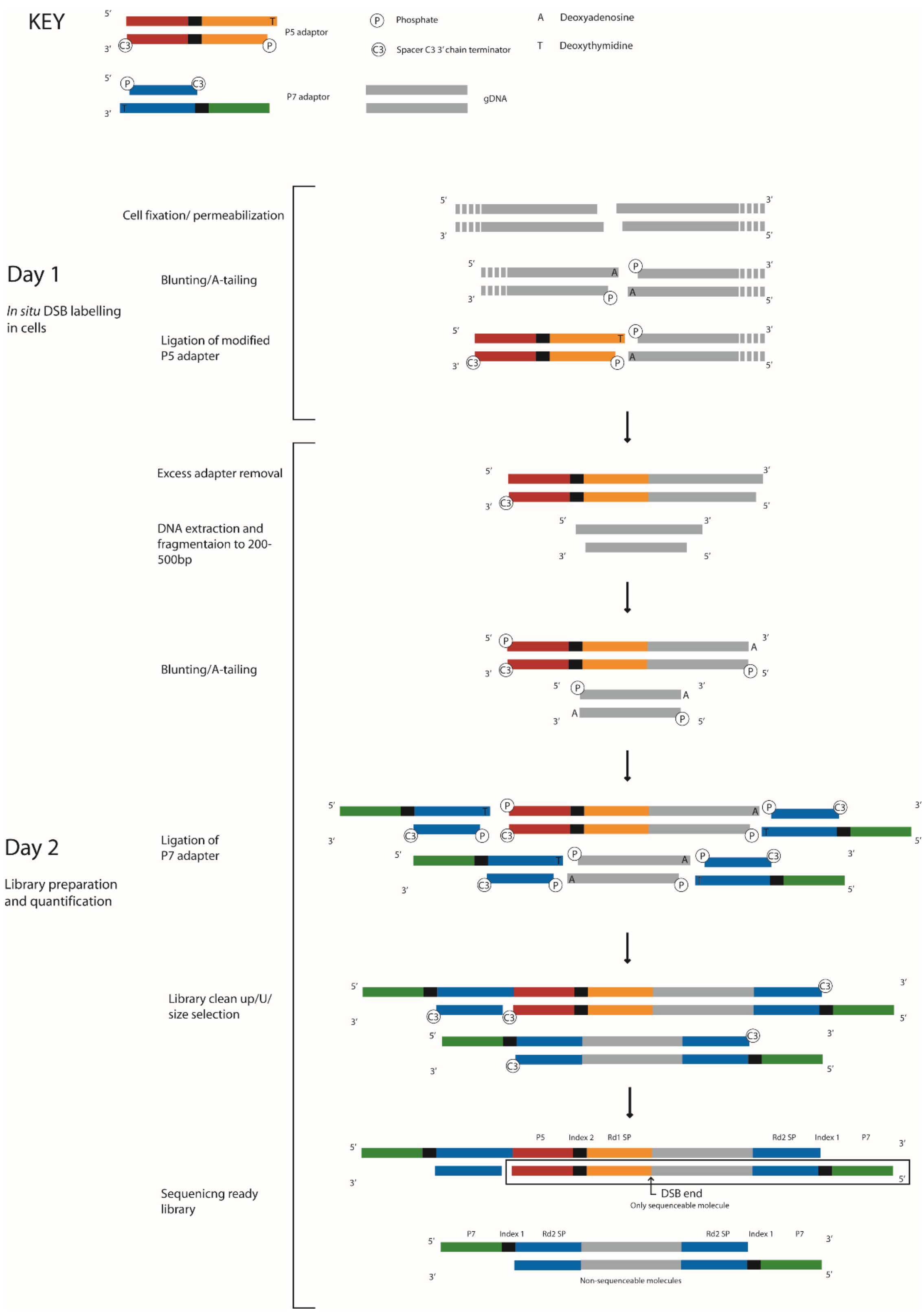

KEY
5′
3′
T
P5 adaptor
P7 adaptor
P Phosphate
C3 Spacer C3 3′ chain terminator
A Deoxyadenosine
T Deoxythymidine
gDNA
Day 1
In situ DSB labelling in cells
Cell fixation/ permeabilization
Blunting/A-tailing
Ligation of modified P5 adapter
Excess adapter removal
DNA extraction and fragmentaion to 200-500bp
Blunting/A-tailing
Day 2
Library preparation and quantification
Ligation of P7 adapter
Library clean up/U/ size selection
Sequenicng ready library
P5
Index 2
Rd1 SP
Rd2 SP
Index 1
P7
DSB end
Only sequenceable molecule
P7
Index 1
Rd2 SP
Rd2 SP
Index 1
P7
Non-sequenceable molecules

Figure 2.1. Schematic of the full INDUCE-seq procedure.

2.1.9 EMX1 on-target editing confirmation by ICE analysis

Confirmation of EMX1 target editing was performed using the Synthego ICE analysis tool. A 639bp target region spanning the EMX1 target site was amplified by PCR using Platinum SuperFi II DNA polymerase (ThermoFisher Scientific) and the primers CCATCCCCTTCTGTGAATGT (Fwd) and GGAGATTGGAGACACGGAGA (Rev). Amplified PCR product was generated for treated and control samples, extracted at 30h, and sequenced by sanger sequencing (Eurofins Genomics). Sanger sequencing.ab1 files were submitted to the Synthego ICE analysis tool (Hsiau et al. 2019) with the EMX1 target sequence (GAGTCCGAGCAGAAGAAGAA) for analysis of the CRISPR target site.

2.1.10 Amplicon sequencing library generation and sequencing

Amplicon sequencing DNA libraries were prepared using a custom panel of rhAmpSeq RNase-H dependent primers (IDT) that flank the INDUCE-seq identified off-targets for EMX1 (**Appendix A, Table A2**). Multiplex PCR was carried out according to manufacturer's instructions using the rhAmpSeq HotStart Master Mix 1, the custom primer mix, and 10 ng of genomic DNA. PCR products were purified using SPRI beads and Illumina sequencing P5 and P7 index sequences were incorporated through a second multiplex PCR using rhAmpSeq HotStart Master Mix 2. Resulting sequencing libraries were pooled and sequenced using a NextSeq 500/550 Mid Output Kit v2.5 (150 Cycles) (Illumina Inc. 20024904).

2.2 Bioinformatics methods

2.2.1 INDUCE-seq pre-processing data analysis pipeline to define break positions

Demultiplexed FASTQ files were obtained and passed through Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove the adapter sequence at the 3' end of reads using the default settings. Following read alignment to the human reference genome (GRCh37/hg19) using BWA-mem (Li and Durbin 2009), alignments mapped with a low alignment score (MAPQ<30) were removed using SAMtools (Li et al. 2009) and soft-clipped reads were filtered using a custom AWK script to ensure accurate DSB assignment. The resulting BAM files were converted into BED files using bedtools bam2bed function (Quinlan and Hall 2010), after which the list of read coordinates were filtered using regions of poor mappability, chromosome ends, and incomplete reference genome contigs, to remove these features from the data. DSB positions were assigned as the first 5' nucleotide upstream of the read relative to strand orientation and were output as a 'breakends' BED file. Care was taken to remove optical duplicates while retaining genuine recurrent DSB events. By maintaining each read ID, flow cell X and Y positional information was used to filter out optical duplicates using a custom AWK script. The final output was a BED file containing a list of quantified single nucleotide break positions. The full INDUCE-seq data processing script is shown in **Appendix B.**

2.2.2 HindIII-induced DSB analysis in HEK293T cells

The positions of HindIII target sites within hg19 were first predicted *in silico* using the tool SeqKit locate (Shen et al. 2016), allowing a max mismatch of 2bp from the HindIII target sequence AAGCTT. The number of breaks overlapping with these predicted sites was calculated using bedtools intersect (Quinlan and Hall 2010). To compare with the DSBCapture EcoRV experiment (Lensing et al. 2016), the same coverage threshold of ≥ 5 breaks per site was used to define each HindIII induced break site. The number of theoretical HindIII mismatch sequence variants was calculated using the following equation (L = String length, M = Number of mismatches):

Number of sequence variants = $(\frac{L!}{M!(L-M)!})3^{M}$

2.2.3 Logo plot generation

Logo plots of the sequence around breaks measured by INDUCE-seq in the HindiII treated and untreated HEK293 cells were generated by first calculating the sequence context +/- 25bp around each break site. Using the resulting FASTA file as an input, the weblogo tool was used to generate the logo plots, using a window size of 50bp and the *H. sapiens* background sequence composition (Crooks et al. 2004). The scripts detailing the steps of sequence analysis are shown in **Appendix B**.

2.2.4 AsiSI-induced DSB detection and analysis in DIvA cells

The positions of AsiSI target sites were calculated in the same way as for HindIII, however with no mismatches allowed and using the sequence GCGATCGC. As DIvA cells are female, sites present on the Y chromosome were removed leaving 1,211 sites for chr1-X. To stringently calculate genuine AsiSI induced breaks, the 8 bp AsiSI site was reduced to 1bp genomic intervals at the predicted break positions. This reduced each 8 bp genomic interval to two 1bp intervals; at position 6 on the plus strand, and position 3 on the minus strand. Direct overlaps were then calculated between 1bp breakend positions and the predicted AsiSI break sites using bedtools intersect (Quinlan and Hall 2010).

Matching strand orientation was required for each overlap to be considered a genuine AsiSI-induced break site.

2.2.5 Defining recurrent breaks and generating randomised datasets

Single-nucleotide endogenous breaks were defined as recurrent break positions using the bedtools merge function (Quinlan and Hall 2010). Distance (-d) values of 1, 2, 5, 10, 20 and 50bp were used to define recurrent breaks across a range of window sizes. **Figure 2.2** demonstrates the merge function with and without the -d argument. Genomic positions with ≥2 breaks were defined as recurrent break sites. Randomised break datasets were generated using the bedtools shuffle function. The list of filtered genomic regions used for defining break positions (**2.2.1**) were excluded from the list of randomised positions using the -excl argument.

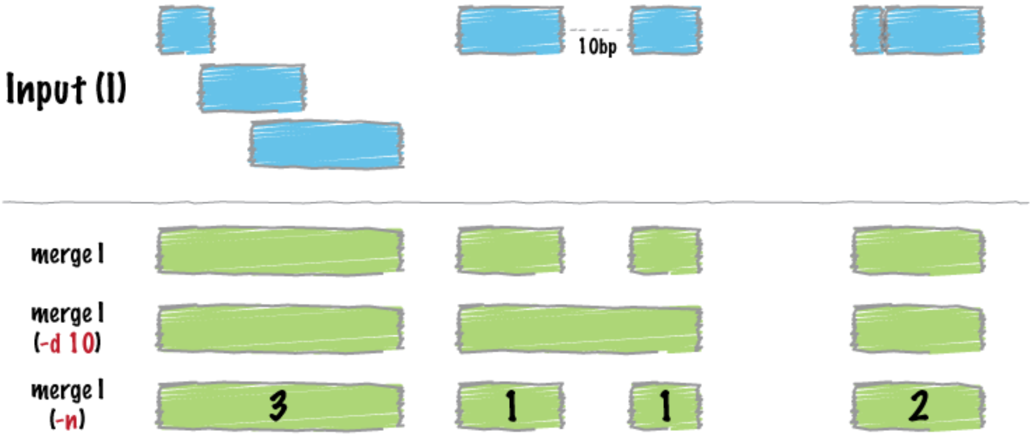


Figure 2.2. Schematic of the bedtools merge function. Input genomic intervals (1bp single break sites) are defined as recurrent break when ≥2 breaks localise to the same position. The distance (-d) argument enables the classification of recurrent break sites when breaks are positioned d base pairs away from each other. The number of breaks per recurrent break site is counted using -n.

2.2.6 CRISPR off-target analysis pipeline

Two sets of potential off-target sites for EMX1 in hg19 were first predicted using the command line version of Cas-OFFinder (Bae et al. 2014), allowing up to 6 mismatches in the spacer and canonical PAM combined for the first set, and up to 7 mismatches for the second. Next, both sets of predicted sequences were filtered based on the mismatch number in the seed region, defined as the 12 nucleotides proximal to the PAM. Each set was filtered for up to 2, 3, 4 and 5 mismatches in the seed, generating a set of 8 files with different mismatch filtering parameters. To define CRISPR-induced DSBs, each 23 bp predicted site was first reduced to a 2bp interval flanking the expected CRISPR break position, 3bp upstream of the PAM. Overlaps were then calculated between these 2 bp expected break regions and the INDUCE-seq 1 bp breakend positions using bedtools intersect (Quinlan and Hall 2010), returning a set of DSBs identified at expected CRISPR break sites. Finally, DSBs overlapping with CRISPR sites were filtered based on the site mismatch number and the number of breaks detected at the site. Sites possessing mismatches >n were required to have more than 1 DSB overlap to be retained as a genuine off-target site. Each set of break overlaps was filtered using a mismatch value of >2, >3, >4 and >5, resulting in a total of 32 filter conditions and off-target datasets for each INDUCE-seq sample.

2.2.7 Calculating overlaps between CRISPR off-target detection methods

EMX1 off-target sites were compared with alternative methods CIRCLE-seq, Digenome-seq, GUIDE-seq, BLISS, and HTGTS (Frock et al. 2015; Kim et al. 2015; Tsai et al. 2015; Tsai et al. 2017; Yan et al. 2017). Genome interval files were generated for each respective off-target detection method. Overlaps of the EMX1 off-targets detected by each method were calculated using bedtools intersect (Quinlan and Hall 2010). Euler diagrams were generated using the R package eulerr.

2.2.8 Amplicon-seq indel analysis

Editing outcomes at the on- and off-targets were determined using CRISPResso software (Pinello et al. 2016) v2.0.32 with the following parameters: CRISPRessoPooled -q30 -ignore_substitutions --max_paired_end_reads_overlap 151. Indel frequencies were compared using CRISPRessoCompare. The full script detailing the used of CRISPResso is shown in **Appendix B**.

3 Chapter III - The development of INDUCE-seq: a novel method to detect genomic DSBs

3.1 Introduction

Although many advancements have been made to date in developing methods to detect CRISPR-Cas9 off-targets throughout the genome, no single approach has been adopted as a suitable ‘gold-standard’. As described in chapter 1, the different approaches have various pros and cons and should be carefully selected depending on the experimental context. For example, *in vitro* approaches offer the highest level of sensitivity at the cost of high levels of false-positive breaks that may not ever exist in cells or *in vivo*. Cell-based assays are usually more technically difficult to perform but can detect putative off-target DSBs in a directly relevant physiological environment. Given that cell-specific factors such as chromatin accessibility, cell cycle timing, transcriptional regulation and DSB repair pathway choice can all alter CRISPR-Cas9 off-target DSB formation and repair outcome, it is now evident that cell-based assays offer the more desirable approach for biologically relevant off-target discovery. As stated in section **1.8**, and given the advantages of a cell-based approach, this study aims to develop a novel cell-based assay for detecting CRISPR-Cas9 off-targets genome-wide that addresses the limitations of existing methods. For the routine profiling of these off-targets, and to better understand the mechanism of off-target induction throughout the genome, the ideal method would involve a simple and scalable procedure that provides a sensitive, unbiased, and quantitative readout.

3.1.1 The limitations of current NGS break detection methods

The key limitation of existing cell-based methods for the detection of DSBs genome-wide is the lack of a quantitative and unbiased measurement of breaks. Most genomic methods to measure breaks are unbiased in one sense, as they do not require prior knowledge of the location of the breaks in order to measure them. Notable exceptions of this are the repair-based methods such as HTGTS, which measure genomic translocations formed from a known and predetermined position in the genome. However, most of the approaches described in section **1.8** do not fall into this category and measure the locations of genomic DSBs in an unbiased manner. Despite this, a different form of bias is inherent in the measurements made by each of these genomic assays and derives from the standard PCR-based library preparation that each of these methods employ for sequencing. Generally speaking, PCR amplification is used during the library preparation procedure to; enrich for molecules with adapters ligated to each end, amplify the amount of library for sequencing, build the full-length adapters required by sequencing, and incorporate different sequence indexes for individual sample identification (Kozarewa et al. 2009). PCR amplification bias is a well-established phenomenon and has been shown to be the principal source of bias in sequencing libraries. The stochastic nature of the PCR is the major force skewing sequence representation after amplification of a pool of unique DNA fragments. This results in; sequence content bias, loss of library molecules, overrepresentation of more abundant sequences, and the inability to accurately quantify the sequencing output (Aird et al. 2011; Jones et al. 2015; Kebschull and Zador 2015). In addition to this, different polymerases can introduce variable biases and error rates depending on the genomic material used (Dabney and Meyer 2012).

Most sequencing applications are not adversely affected by the phenomenon of amplification bias. However, for the quantitative measurement of genomic DSBs, such as CRISPR-Cas9 off-targets, PCR-amplification introduces high levels of bias into a system where the signals (genomic DSBs) are binary and infrequent. Any DSB capture method that utilises a PCR-based library preparation suffers from a degree of PCR bias, therefore distorting the representation of the actual distribution of DSBs present in the genome. This makes the measurement and quantification of DSBs inaccurate and challenging. Further compounding this problem, there is currently no clear consensus definition of what ‘noise’ represents in these assays; some studies report crosslinking as the primary source of background breaks, whereas others report that gentle fixation has little effect on the number of breaks (Canela et al. 2016; Biernacka et al. 2018). The presence of a high background also masks low-frequency DSB events, such as rare CRISPR off-targets or endogenous breaks, thus severely limiting their identification. Furthermore, other DSBs induced randomly by exogenous sources such as ionizing radiation, or ultra-rare off-target DSBs, induced by CRISPR-Cas9, are indistinguishable from noise in these assays, and are therefore not detected. Of the DSB detection methods, only BLISS has utilised an adapter design that incorporates a unique molecular identifier (UMI) in an effort to control for PCR amplification biases (Yan et al. 2017). During the break capture stage of the methodology, each DSB end is labelled with an adapter containing a randomised and unique 8-10bp barcode, enabling bioinformatic filtering of PCR-amplified sequencing reads that share the same barcode. Theoretically, recurrent breaks mapped to the same position in the genome, but that originate from different cells, will have different UMI barcodes making them easily distinguishable from amplified copies. This indeed allows the quantification of break events present in the sequenced sample, but does not fully overcome the biases introduced. This is due to the stochastic nature of PCR amplification, coupled with the repeated cycles of exponential amplification and sample dilution undertaken during sequencing library preparation. This results in both overrepresentation, underrepresentation, and occasionally complete loss of DNA fragments of low abundance and/or with extreme base compositions (**Figure 3.1**) (Hess et al. 2020). UMI correction can only compensate for the overrepresented fragments by filtering of duplicated sequencing reads, thus making it unsuitable for accurate, quantitative, and unbiased DSB measurement throughout the genome.

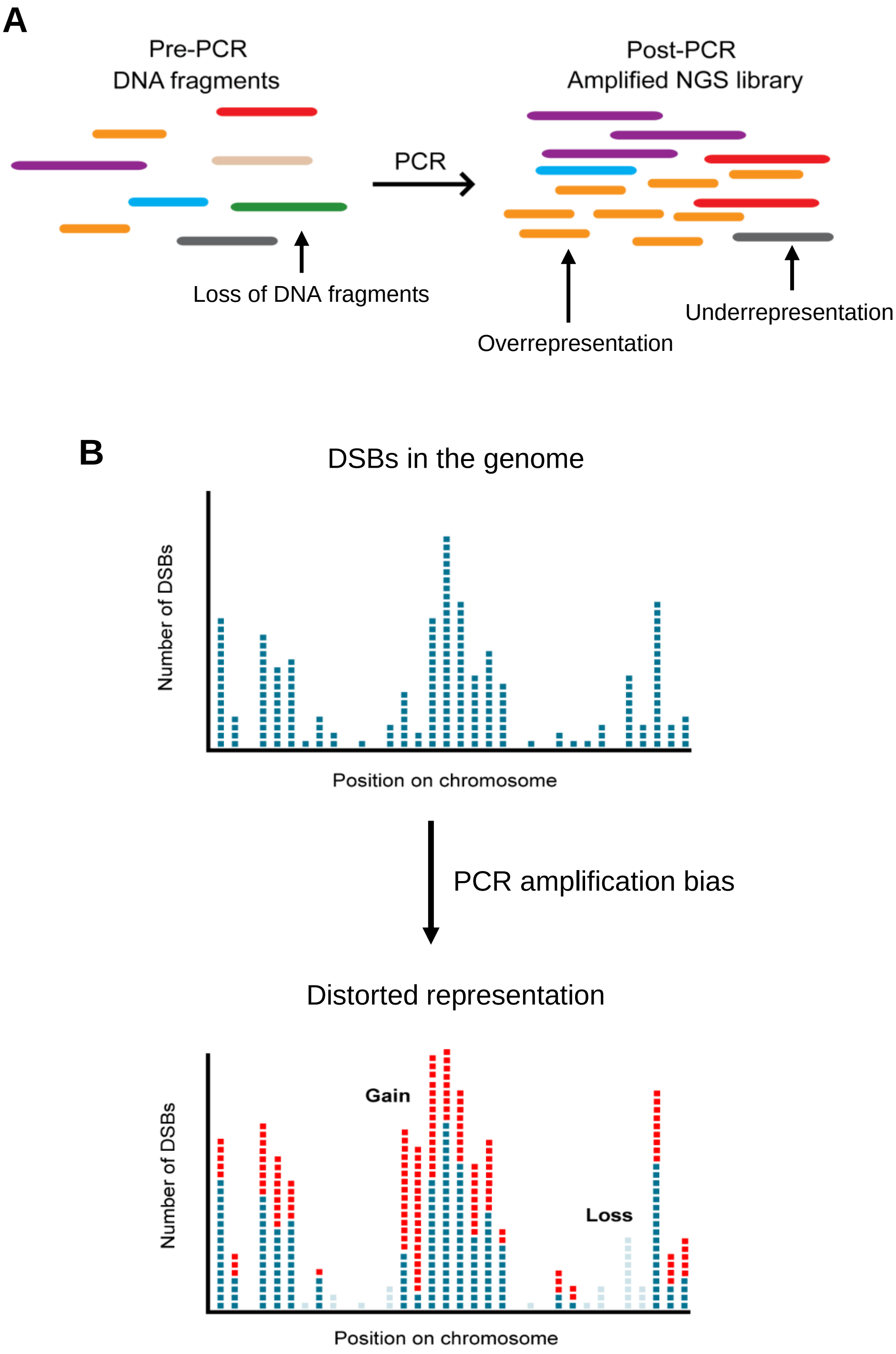


Figure 3.1. Schematic demonstrating PCR amplification bias in NGS libraries. (**A**) PCR amplificant can significantly bias the composition of a heterogenous mixture of DNA fragments. Fragments can become overrepresented, underrepresented, and even complete lost during the amplification process. (**B**) The effect of PCR amplification on genomic DSB measurement. PCR amplification bias results in a distorted representation of breaks in the genome.

In addition to issues with the accuracy of measurement, another key limitation of current genomic methods to detect DSBs is in the scalability and simplicity of the laboratory procedure. Generally, to perform targeted sequencing of any genomic feature of interest, DNA corresponding to that feature must be separated from the bulk of genomic DNA prior to library preparation. In order to selectively capture and sequence DSB sites, the enrichment process is carried out either by enzymatic amplification or by streptavidin-mediated pull-down of break-labelled DNA. PCR or *in vitro* transcription-based amplification protocols allow for the enrichment of DSB sites from a very small number of cells (as low as 10^3), as demonstrated by methods such as GUIDE-seq and BLISS (Tsai et al. 2015; Yan et al. 2017). This aspect makes amplification-based enrichment highly scalable, allowing multiplexing of samples in 96-well format with ease, in addition to making it compatible with many automation platforms. Enrichment of DSBs by PCR amplification, however, suffers from the same caveats associated with PCR-based library preparation, further compounding biases and skewing sequence representation. Furthermore, linear amplification by T7-mediated *in vitro* transcription is prone to premature termination on low-complexity sequences such as homopolymeric poly(A) or poly(T) stretches (Hoeijmakers et al. 2011). The use of streptavidin-mediated pull-down to separate break ends from fragmented genomic DNA offers an unbiased approach for DSB enrichment. This, however, requires at least 10^6 cells per sample and is more labour intensive than amplification-based enrichment. This requirement limits the scalability of such assays, making them unsuitable for the examination of DSBs in materials of limited availability, such as clinical samples or in high-throughput experiments.

3.1.2 Chapter aims

Given these limitations, for accurate, unbiased, and scalable discovery of CRISPR-Cas9 off-targets throughout the genome, there is a need to establish a novel cell-based method to measure genome-wide DSBs which exploits a PCR-free library preparation, and does not rely on amplification or streptavidin-biotin pull-down for break enrichment. The aim of this chapter is therefore to develop a novel PCR-free method to measure genomic DSBs.

3.2 Results

3.2.1 Conceptualisation of the INDUCE-seq methodology

To design an improved break capture method free from the limitations of existing methods, a novel PCR-free methodology was devised called INDUCE-seq (Identification and quantificatioN of DSBs by Unbiased flow Cell Enrichment and sequencing). INDUCE-seq is performed by immobilising cells on to a solid surface, allowing all *in situ* reactions to easily be performed in plate format, and making the process suitable for tissue processing (Yan et al. 2017). For these reasons, INDUCE-seq has the potential to offer the most scalable and least laborious protocol of all available DSB detection methods to date. Furthermore, this break-labelling approach is not limited by the caveats associated with indirect break capture using repair outcome (Yan et al. 2017; Mirzazadeh et al. 2018; Ballinger et al. 2019). To circumvent PCR amplification bias, INDUCE-seq comprises a novel method that allows the enrichment of tagged DSBs without the need for a dedicated enrichment step, or PCR-amplification to generate sequencing libraries. INDUCE-seq uses an enrichment strategy that makes use of elements that are core to the conventional sequencing library preparation, while at the same time eliminating steps that are superfluous. Illumina sequencing is facilitated by the ligation of two modular adapters (P5 and P7) onto the DNA of interest, which enables the capture of adapter ligated DNA on to the sequencing flow cell via hybridisation to an array of P5 and P7 complementary oligonucleotide anchors (**Figure 3.2**) (Bentley et al. 2008). DNA fragments are then 'bridge-amplified' *in situ* on the flow cell to produce monoclonal clusters, each of which will become a sequencing read, initiated from the 3' end of the P5 adapter. For the generation of a paired read, the bridge amplification process is repeated, and sequencing is initiated from the 3' end of the P7 adapter (Bentley et al. 2008).

For the capture of break ends *in situ*, INDUCE-seq uses a full-length double-stranded Illumina P5 adapter to ligation-tag the ends of exposed DSBs, which allows sequencing to initiate directly from the site of the break without any further manipulations such as PCR to build the remaining parts of the adapter for flow cell interaction (Lensing et al. 2016). Following standardised steps for DSB labelling, DNA extraction, and fragmentation, the second stage of the INDUCE-seq library preparation involves the ligation of a half-functional Illumina P7 sequencing adapter. By ligating every DNA fragment with a half-functional P7 adapter, a sequencing library may be generated containing a mixture of functional break-labelled (P5-P7) fragments, and non-functional (P7-P7) sonicated genomic DNA fragments. The use of half-functional P7 adapters has been demonstrated previously for other uses (Zheng et al. 2014). Subsequent loading of the Illumina flow cell enriches for labelled DSB fragments, while all other fragments are discarded during the initial phases of sequencing. INDUCE-seq combines DSB enrichment and sequencing into a single step, and thoroughly simplifies DSB sequencing workflow (**Figure 3.3**). Full details of the INDUCE-seq method can be found in section **2.1.8**.

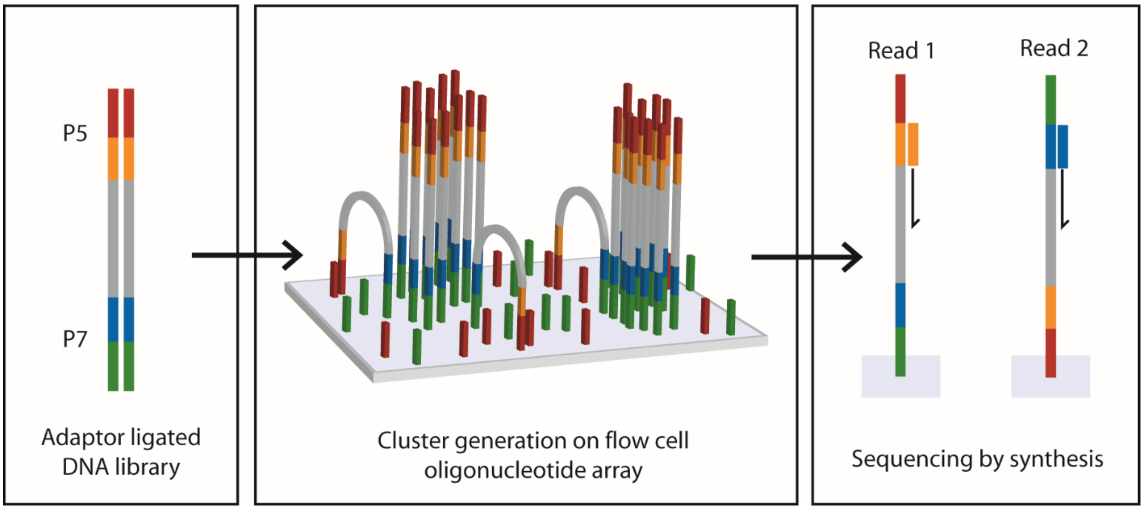

P5
P7
Adaptor ligated DNA library
Cluster generation on flow cell oligonucleotide array
Read 1
Read 2
Sequencing by synthesis

Figure 3.2. Illumina sequencing by synthesis workflow. P5 and P7 adapters are ligated to DNA of interest to generate sequencing libraries. Denatured DNA libraries are loaded on to the Illumina flow cell where they hybridize to an array of P5 and P7 oligonucleotide anchors. Bridge amplification generates monoclonal clusters from each DNA fragment. Sequencing by synthesis is initiated using a sequencing primer that binds to the P5 adapter, generating read 1 from one of the DNA strands, which is sufficient for single end sequencing. For paired end sequencing, bridge amplification is repeated and a sequencing primer that binds to the P7 adapter is used to generate read 2.

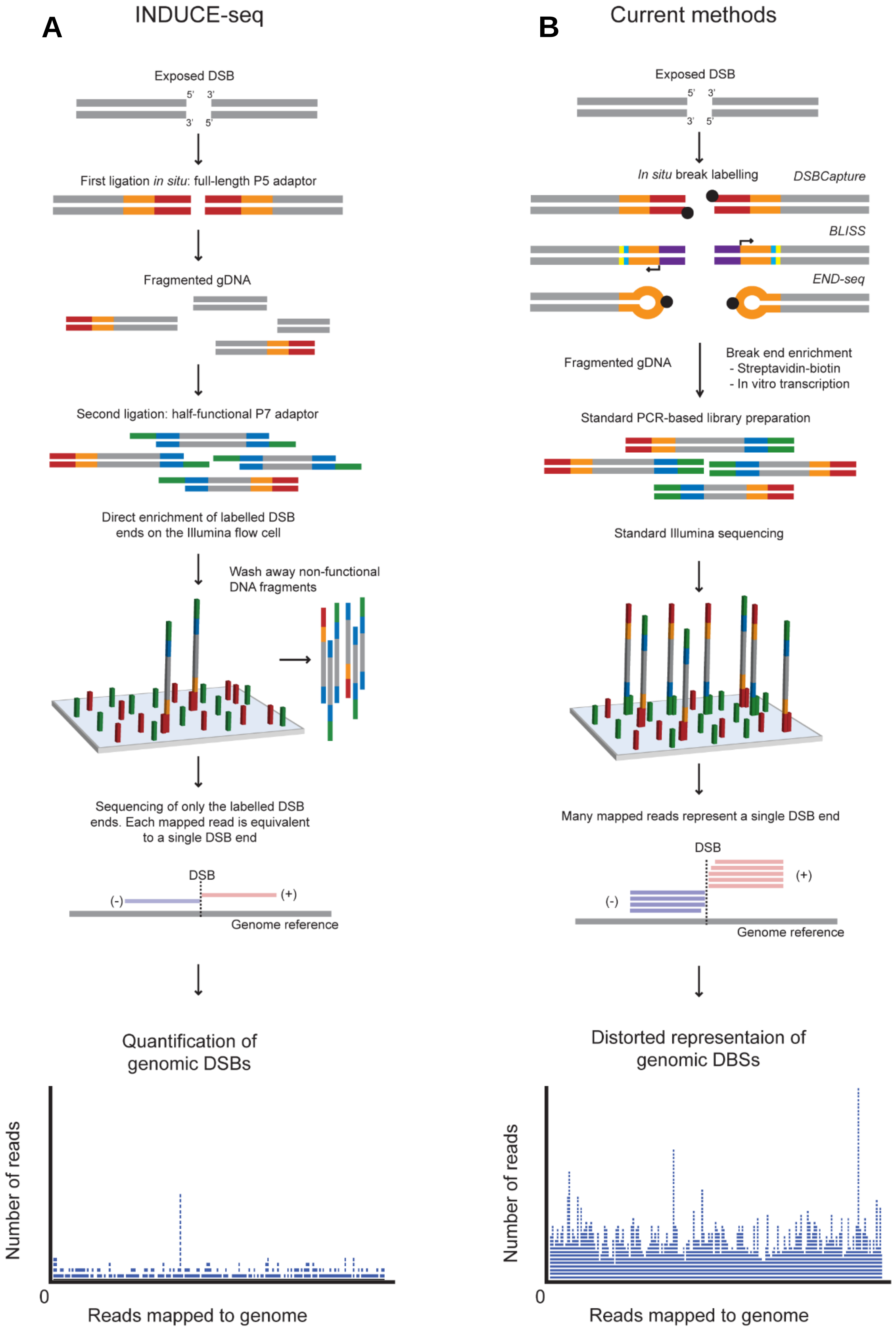

A
INDUCE-seq
Exposed DSB
5'
3'
3'
5'
First ligation *in situ*: full-length P5 adaptor
Fragmented gDNA
Second ligation: half-functional P7 adaptor
Direct enrichment of labelled DSB ends on the Illumina flow cell
Wash away non-functional DNA fragments
Sequencing of only the labelled DSB ends. Each mapped read is equivalent to a single DSB end
DSB
(-)
(+)
Genome reference
Quantification of genomic DSBs
Number of reads
0
Reads mapped to genome
B
Current methods
Exposed DSB
5'
3'
3'
5'
In situ break labelling
DSBCapture
BLISS
END-seq
Fragmented gDNA
Break end enrichment
- Streptavidin-biotin
- In vitro transcription
Standard PCR-based library preparation
Standard Illumina sequencing
Many mapped reads represent a single DSB end
DSB
(+)
(-)
Genome reference
Distorted representaion of genomic DBSs
Number of reads
0
Reads mapped to genome

Figure 3.3. Comparison of INDUCE-seq and current DSB mapping workflows. (**A**) Overview of INDUCE-seq workflow and theoretic sequencing output. Because of the PCR-free design of the protocol, INDUCE-seq should produce quantitative output where one read is equivalent to one break. (**B**) Overview of the DSBCapture, BLISS and END-seq workflows. Sequencing following the standard PCR-based library construction generates a biased output where one read is not equivalent to a single DSB.

3.2.2 INDUCE-seq adapter design

The INDUCE-seq adapter design relies on meeting the requirements for Illumina sequencing on specific DSB labelled DNA fragments while simultaneously leaving all other fragments non-functional. In a standard sequencing library, adapters are comprised of three key components; the P5/P7 flow cell binding sequences, sample indexes of 6-8 variable nucleotides in either the P7 adapter (single index) or both adapters (dual index), and sequencing primer binding sites for read 1 and read 2 (Rd1 SP and Rd2 SP) (**Figure 3.4A**). Following library preparation, all available DNA fragments are prepared for sequencing. Loading of the sequencing flow cell captures properly prepared ssDNA library fragments via the hybridisation of the 3' ends of the P5 and P7 adapters on to complimentary 5' P5 and P7 graft oligonucleotides that are covalently linked to the flow cell (Bentley et al. 2008). As described previously, a full-length adapter containing the P5 binding sequence, index, and Rd2 SP, was utilised for *in situ* DSB labelling to ensure that sequencing could initiate from the DSB site (**Figure 3.4B**). To complete these fragments for sequencing, and without enabling sequencing of all other genomic DNA fragments, the second half-functional P7 adapter was designed specifically to lack the 3' P7 binding sequence required for flow cell hybridisation. For INDUCE-seq sequencing libraries, only the 3' ends on either side of a DSB are correctly labelled for sequencing and contain the 3' P5 binding sequence to hybridise with the Illumina flow cell (**Figure 3.4C**). All other fragments lack the ability to hybridise, and are washed away prior to bridge amplification, leaving only the DSB labelled fragments for sequencing.

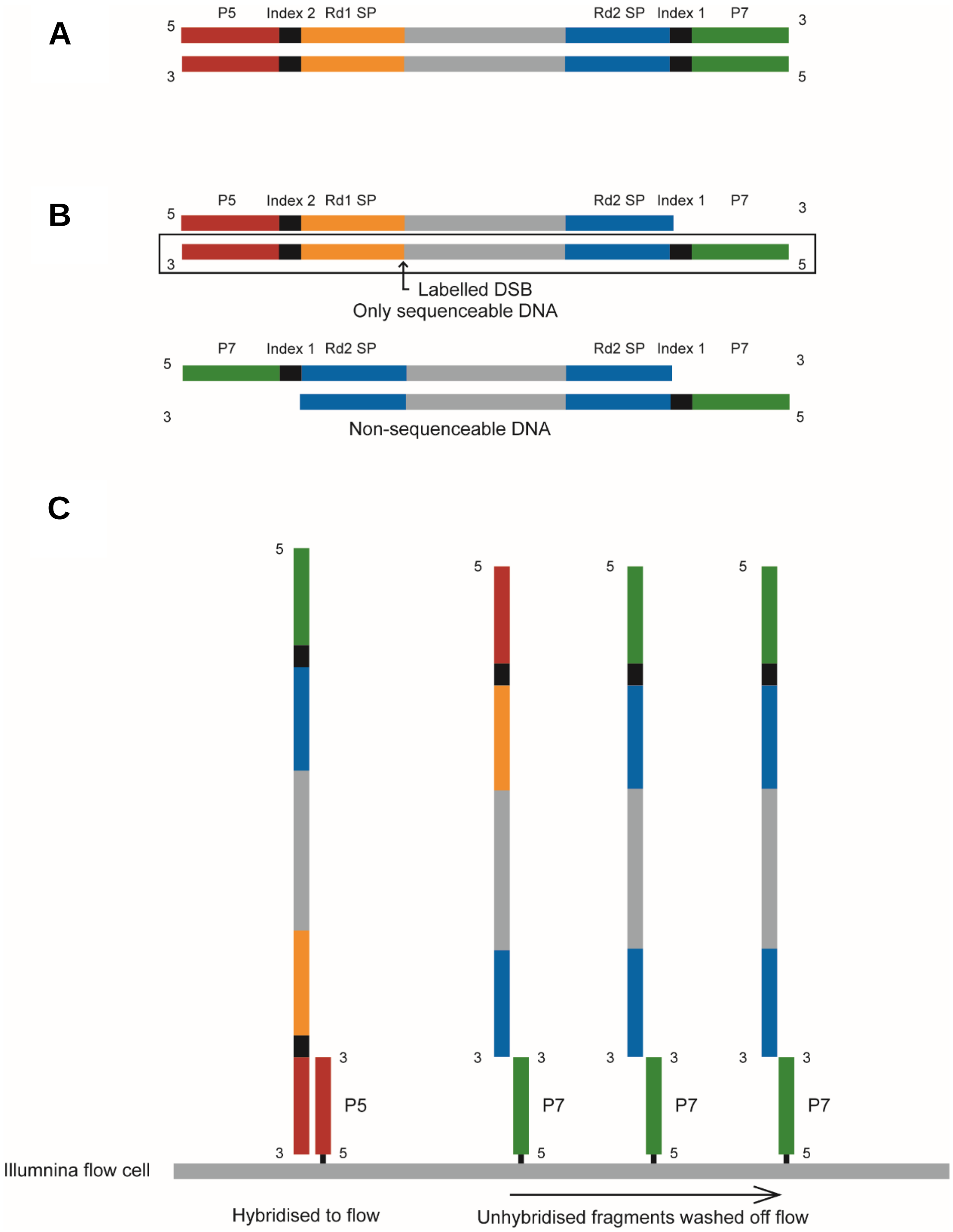


Figure 3.4. Detailed schematic of flow cell enrichment adapter design. (**A**) Structure of a complete adapter ligated dsDNA fragment for sequencing. 3' P5 and P7 adapters hybridize with the flow cell. Sequencing primers bind to the RD1 SP and RD2 SP sequences during the first and second sequencing read. Indexes allow differentiation of different samples. (**B**) Structure of DNA fragments present in an INDUCE-seq library. Only the DSB ligated fragment is comprised of all the adapter components required for sequencing. (**C**) Loading of the INDUCE-seq library on to the sequencing flow cell will enrich for DSB ligated fragments via hybridization of the 3' ends of P5 adapter sequences. No other fragments can interact with flow cell and will be removed.

3.2.3 INDUCE-seq adapter modifications

In addition to sequence modifications described above, several chemical modifications were also made to the adapters to enable the INDUCE-seq process (**Figure 3.5**). Efficient ligation on to break ends and during the library preparation was enabled by the addition of a 5’ phosphate and 3’ deoxythymidine triphosphate ‘T-tail’ to the proximal end of the adapters (**Figure 3.5**). This modification has been demonstrated to increase the ligation efficiency of break capture over blunt-end ligation methodologies, and is used routinely for library preparation (Canela *et al.* 2016; Lensing *et al.* 2016). Phosphorothioate linkages at the T-tails were also implemented to resist exonuclease activity. The incorporation of a C3 Spacer phosphoramidite at the 3’ distal ends of the adapter was adopted in order to maintain the integrity of both the P5 and P7 adapters during the multiple end preparation and ligation steps (**Figure 3.5**). This renders the distal 3’ end of each adapter inert, blocking polymerase extension, exonuclease cleavage, and ligation activity (Dames *et al.* 2007; Lee *et al.* 2011; Wickersheim and Blumenstiel 2013). By blocking the distal 3’ ends of the adapters in this manner, unwanted head to tail adapter:adapter dimers may be prevented during ligations, which can saturate sequencing experiments if not excluded from sequencing libraries. Furthermore, this modification ensures that the 3’ end of the P5 adapter remains unaltered, preventing P7 ligation during the library preparation that would interfere with flow cell hybridisation and sequencing. Taken together, these modifications confer unique properties to the to the INDUCE-seq adapters which enable the efficient capture of break ends, and enrichment and sequencing on the Illumina flow cell.

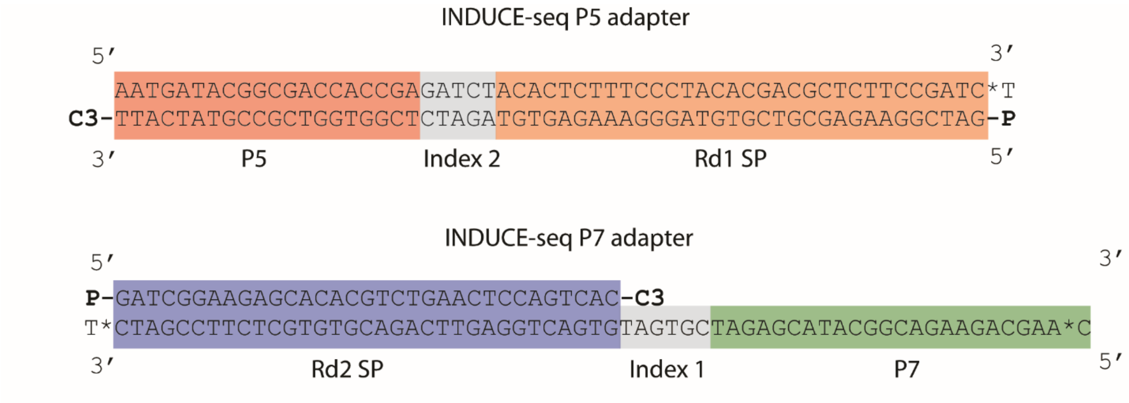

INDUCE-seq P5 adapter
5′
3′
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T
C3-TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAG-P
3′
P5
Index 2
Rd1 SP
5′
INDUCE-seq P7 adapter
5′
3′
P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-C3
T*CTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGTAGTGCTAGAGCATACGGCAGAAGACGAA*C
3′
Rd2 SP
Index 1
P7
5′

Figure 3.5. INDUCE-seq adapter modifications. Proximal ends designed to facilitate ligation are modified with 5’ Phosphate groups (P) and 3’ dTTP overhangs. Phosphorothioate linkages (*) are incorporated at T-tails and the P7 adapter 5’ end to resist exonuclease degradation. 3’ distal ends and are modified with C3 Spacer phosphoramidite (C3) primarily to prevent ligation, but also confer exonuclease resistance and block polymerase extension during quantitative PCR.

3.2.4 Experimental confirmation of adapter function

Before applying INDUCE-seq to the capture of DSBs in cells, initial experiments focused on confirming that the adapters functioned as intended. Because of the two-step ligation strategy employed by INDUCE-seq, the sequencing adapters must resist ligation at their 3’ distal ends via the end blocking C3 spacer modification. Several combinations of adapter variants, with and without the C3 spacer modification, were tested under the same ligation conditions used for break labelling during the INDUCE-seq procedure (**Figure 3.6**). Following incubation with T4 DNA ligase, the addition of the C3 modification to the P5 adapter almost completely inhibited dimer formation when compared to the non-modified version, which generated several linear concatemers between 100 and 300bp in length (**Figure 3.6, lanes 1 & 3**).

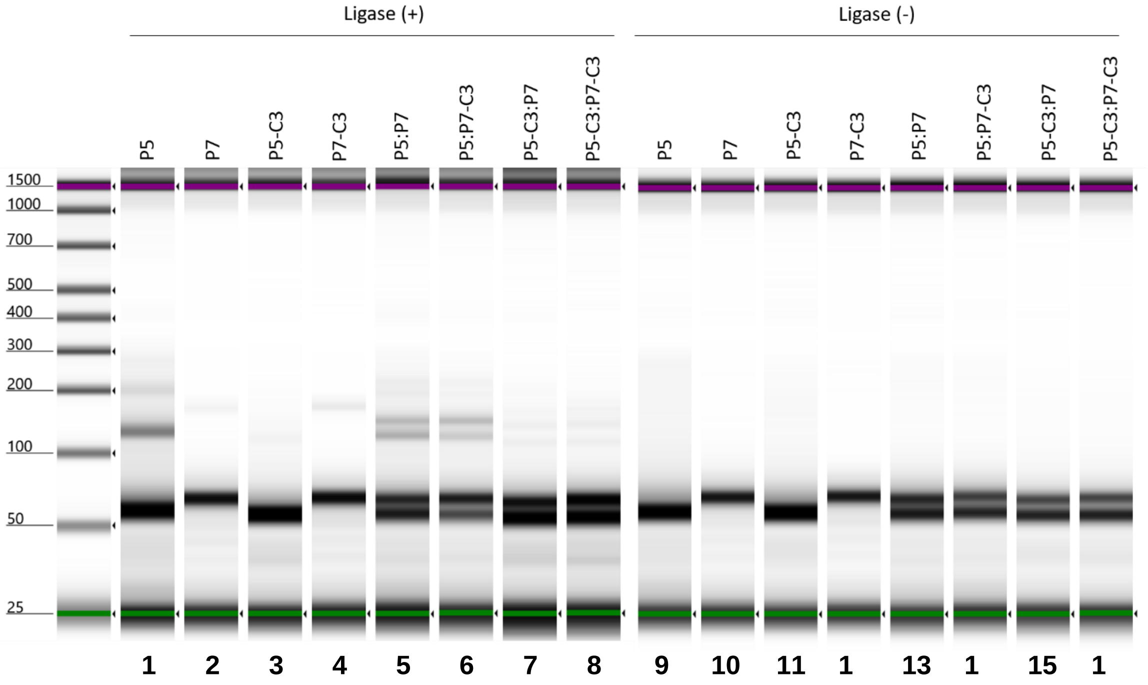


Figure 3.6. ***In vitro*** **P5 and P7 adapter ligations with and without a 3’ C3 spacer modification.** Adapters were ligated individually, and in combinations without additional DNA to allow linear concatemers to form. The lack fragments greater than the length of the adapter monomer indicate that ligation has been prevented by the C3 spacer end modification.

Interestingly, the P7 adapters with and without C3 modification displayed almost identical profiles, with trace amounts of dimer forming for each (**Figure 3.6, lanes 2 & 4**). When P5 and P7 adapters are combined in various combinations, substantial dimerization is only seen in combinations that are comprised of the unmodified P5 adapter, generating proximal:distal dimers of P7:P5 and P5:P5 as seen by the double bands at 140bp and 120bp, respectively (**Figure 3.6, lanes 5 & 6**). These results confirm that incorporating an end blocking modification protects the 3' distal end of the P5 sequencing adapter from further ligations, allowing the ligation of the second sequencing adapter without compromising the integrity of the labelled break ends, ultimately enabling the building of INDUCE-seq libraries.

3.2.5 Demonstrating proof of concept for INDUCE-seq

Having tested the INDUCE-seq adapters *in vitro*, we first validated INDUCE-seq by measuring breaks in untreated (WT) HEK293T cells, and in cells which has been digested *in situ* with the restriction endonuclease HindIII. Controlled enzymatic break induction has been used previously to assess the accuracy and sensitivity of DSB labelling methods (Lensing et al. 2016; Yan et al. 2017). Furthermore, given the non-amplified nature of INDUCE-seq and the relative scarcity of DSB events in cells, it was not clear whether the assay would generate enough material for standard Illumina sequencing. HindIII can target around 830,000 sites in the human genome, and therefore induces several orders of magnitude more DSBs from a population of cells than exist endogenously in a similar untreated population of cells.

Following break labelling, DNA extraction, and fragmentation, WT and HindIII treated samples were assessed for the correct fragment size distribution using capillary electrophoresis before proceeding with library preparation (Figure 3.7, blue and green lines). Correct fragment sizing is critical for library preparation and sequencing; small (<100bp) fragments will be indistinguishable from adapters during the library preparation and large (>1000bp) sequence with a lower efficiency (Kircher et al. 2011). In addition to checking size distribution, this confirmed the complete removal of excess P5 adapter from the sample, which is removed post ligation by washing of the cells prior to DNA extraction. Inadequate removal of excess adapter prior to library preparation would result in P5:P7 adapter dimers, which can bind to the flow cell, and would potentially saturate the sequencing run with non-informative reads. Regardless of treatment, both WT (**Figure 3.7, blue line**) and HindIII treated (**Figure 3.7, green line**) samples were correctly fragmented to around 250bp. Library preparation was carried out using the INDUCE-seq half-functional P7 adapter. DNA ends were prepared for ligation in a similar manner as used for break tagging *in situ*; end blunting, phosphorylation and A-tailing generated suitable ends for ligation of the T-tailed P7 adapter. Size selection was applied to exclude fragments of <150bp to ensure no P7 adapter monomer was left in the sample. This is particularly important as remaining P7 adapter would act as a primer for non-functional fragments during library quantification, resulting in an inaccurate measurement of library concentration. Successful adapter ligation was measured by the shift in library fragments size distributions from ~250bp to ~300bp (**Figure 3.7, orange and red lines**).

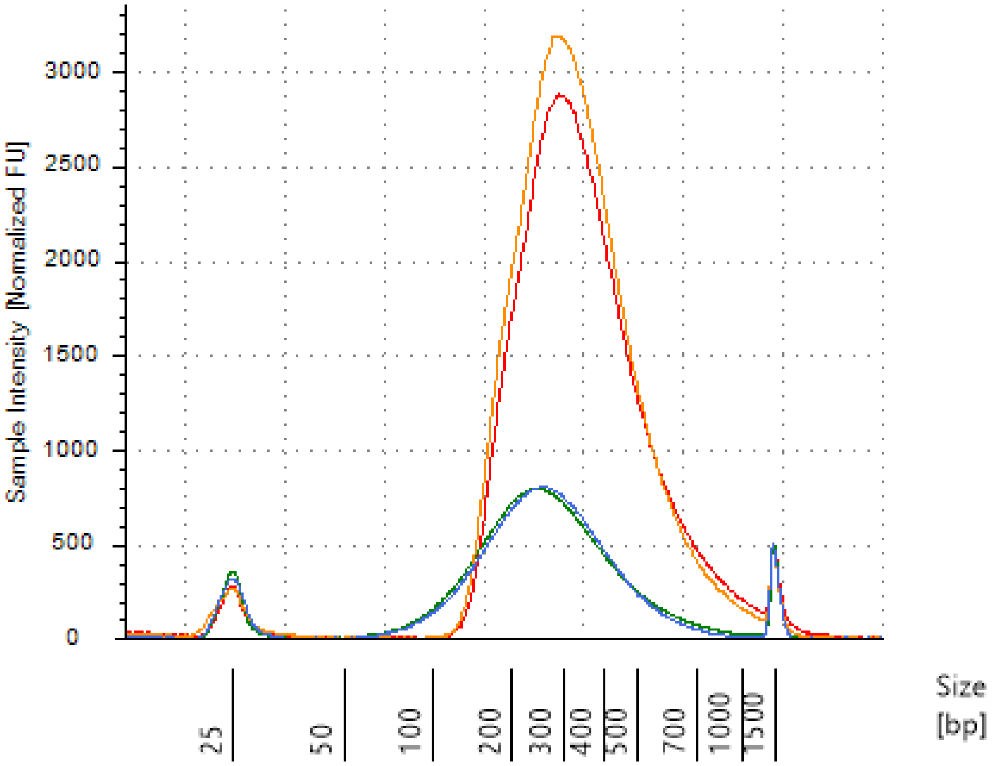


Figure 3.7. INDUCE-seq sample fragment size distributions. Correct fragmentation following sonication of untreated (blue line) and HindIII treated (green line) samples is confirmed by an average size of ~250bp. The fragment size distribution shift (blue to orange) (green to red) confirms successful ligation during the library preparation. Lack of DNA fragments at the 50-100bp mark confirms successful size selection and removal of excess adapter.

Although bulk quantification of DNA fragment sizes provides valuable information about library generation and sizing, it does not inform whether the process has yielded sequenceable material. When using a PCR-based approach that enriches for correctly ligated library DNA fragments this is not an issue and it is assumed that DNA amplified via correct P5 and P7 adapter ligation can be sequenced.

As INDUCE-seq generates complex library mixtures with varying proportions of sequenceable and non-sequenceable fragments, quantification by standard intercalating fluorescent dyes would not be suitable and drastically overrepresent the amount of sequencing library present. Quantification using qPCR enables the accurate measurement of the fraction of sequenceable material in the sample and directly utilises the same design that enables flow cell enrichment. PCR primers used for library quantification anneal to the 3' distal ends of the P5 and P7 adapters, allowing the measurement of only the DSB-labelled DNA fragments containing the P5 primer binding site. Fragments that are non-functional for sequencing and comprise two P7 adapters lack primer binding sites and will not be amplified.

Quantification of the INDUCE-seq libraries revealed a surprisingly low percentage of sequenceable fragments for both the untreated and HindIII treated samples (**Table 3.1**). As anticipated, both libraries were composed of similar quantities of total DNA, 267ng and 216ng, respectively (as measured by bulk DNA quantification). This indicating that the number of DSBs present in each sample does not drastically alter the combined total of functional and non-functional library molecules. The number of sequenceable molecules, however, was substantially lower than the amount of total DNA, even for the HindIII treated sample at approximately 36.8pg, representing merely 0.0171% of the total library DNA. The untreated sample contained even fewer sequenceable fragments, at 0.405pg, representing merely 0.0002% of the library.

These findings confirm that, as determined by qPCR, the novel INDUCE-seq library preparation and adapter design enable the selective enrichment of genomic breaks *in situ*, independently from the total sample DNA amount

.

Table 3.1. Quantification of untreated and HindIII treated INDUCE-seq libraries by qPCR.

	All library fragments	Sequenceable library fragments	
Sample	**DNA amount (pg)**	**DNA amount (pg)**	**% Sequenceable**
Untreated	**267,000**	**0.405**	**0.0002%**
HindIII	**216,000**	**36.835**	**0.0171%**

Due to the vastly higher concentration of sequenceable molecules and the ability to identify known cleavage sites, the HindIII library was selected for sequencing on an Illumina MiSeq Nano flow cell. Consistent with standard Illumina MiSeq loading guidelines (5µl, 4nM library), the HindIII library (5 µl, 0.013nM sequenceable, 75.5nM non-sequenceable) would cause significant over-clustering if all DNA in the sample bound to the flow cell, saturating the sequencer, and leading to failure to define clusters and many unusable reads. If the INDUCE-seq library functioned as designed, only the quantified sequenceable molecules would bind, resulting in underclustering, but still generate a useable sequencing output comprising fewer, and high-quality reads. This outcome means that it is possible to use the Illumina sequencing flow cell to enrich for a very low number of specifically labelled DNA fragments from a complex mixture of genomic DNA fragments.

The resulting sequencing output generated 45,287 sequencing reads, of which 38,264 (84.5%) were successfully mapped to the human genome. This was significantly lower than the upper limit of the MiSeq Nano flow cell (1 million reads) confirming that the so-called non-functional DNA fraction did not cause overclustering. Furthermore, this demonstrated that the break-labelled fraction was exclusively sequenceable. Sequencing quality was not adversely affected by performing flow cell enrichment. At all base positions a median phred score quality of >28 was observed, indicating that the accuracy of base calling was >99.84% (**Figure 3.8A**) (Ewing and Green 1998). Base composition outside of the HindIII sequence, at the first five positions of the read, was not adversely affected by flow cell enrichment, with consistent base content across the remainder of read (i.e. position 6-150) (**Figure 3.8B**).

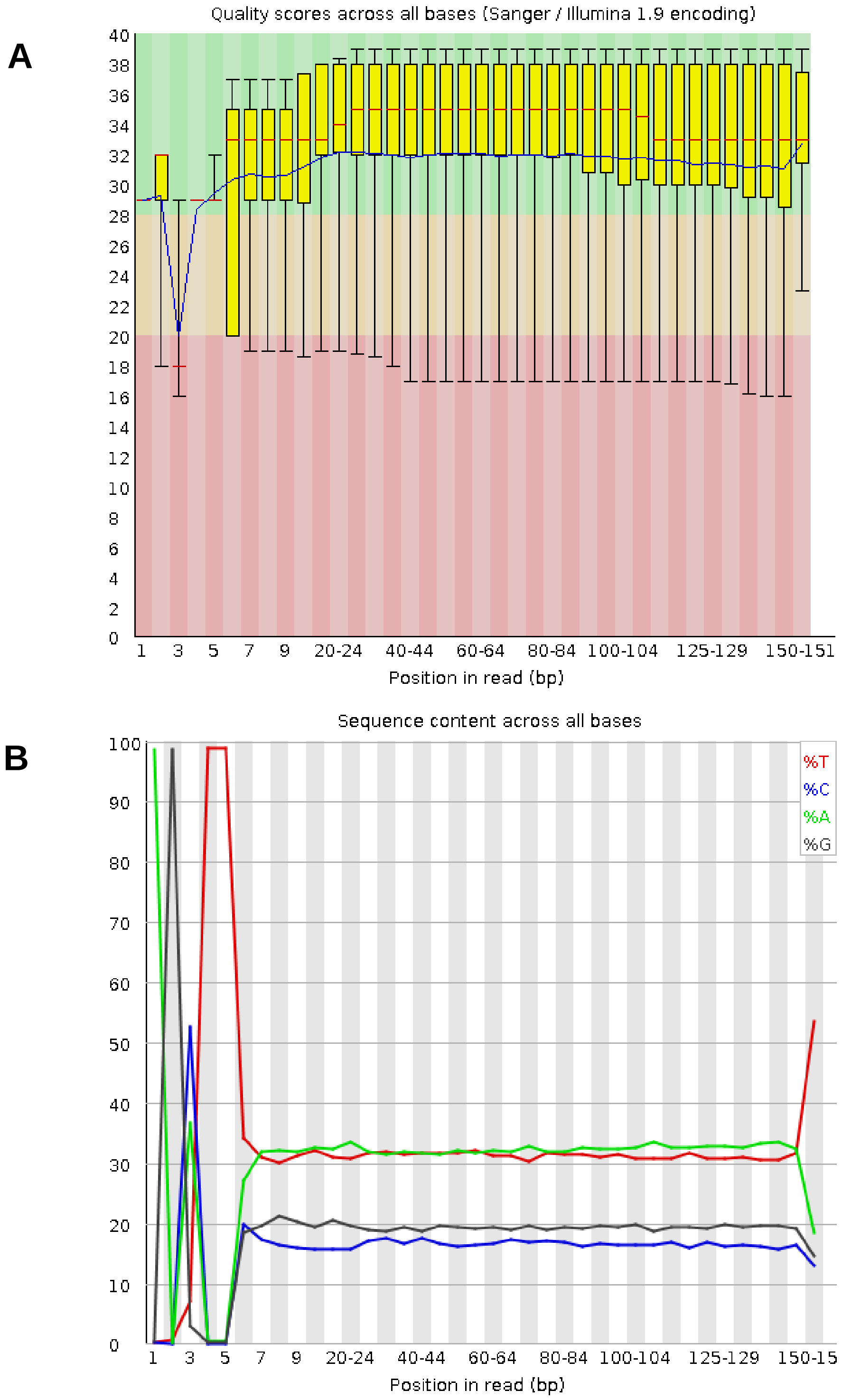


Figure 3.8. Quality checking of INDUCE-seq sequencing output. (**A**) Base quality (phred score) across the whole read is in the acceptable range (median > 28) despite high quantity of unsequenceable material loaded on to flow cell. (**B**) Sequence content across all bases reveals HindIII recognition sequence at start of every read, with uniform base contribution across the rest of the read.

Sequencing reads were mapped to chromosome 1 (**Figure 3.9A**) and reads corresponding to HindIII cleavage sites were visualised using a genome browser (**Figure 3.9B**). HindIII-induced DSBs were easily distinguishable from endogenous breaks by the cleavage motif present at the start of each forward read initiated from the P5 adapter. Following the filtering of poor mappability regions, 37,507 individual break ends were identified. As just 1/3 of the total library was used for sequencing, and samples were generated from ~100,000 cells, this represented an average of 1.14 breaks per cell. Based on the strand that breaks are mapped to, it can be inferred which side of the break was labelled, with breaks labelled on the right, mapping to the plus strand, and breaks labelled on the left, mapping to the minus strand. Both sides of DSBs were labelled in equal proportions; equal numbers of breaks were mapped to the right (18,557) and the left sides of breaks (18,950), demonstrating no strand-bias in break detection. The 37,507 break-ends identified corresponded to 21,569 distinct genomic positions, of which 99% (21,347) were mapped to a site with the HindIII recognition sequence. As INDUCE-seq is amplification-free and quantitative, each read mapped corresponds to a unique break measured in an individual cell from a population. It thus follows that more than one break at a single position can only result from the measurement of a different break at the same position, in independent cells. Multiple breaks were identified at 50.5% of all positions detected, indicating that half of all break sites detected were labelled in more than one cell.

Despite the low number of reads, the identification of multiple HindIII-induced breaks confirmed that INDUCE-seq can be used to detect genomic DSBs via *in situ* tagging, followed by break enrichment directly on the Illumina flow cell. As anticipated, overloading of adapter ligated but non-functional DNA several orders of magnitude over the recommended quantity, did not result in the failure of the sequencing run or poor-quality reads. For the first time, this demonstrates how the Illumina flow cell can be used specifically as a tool for the unbiased enrichment of specific genomic features.

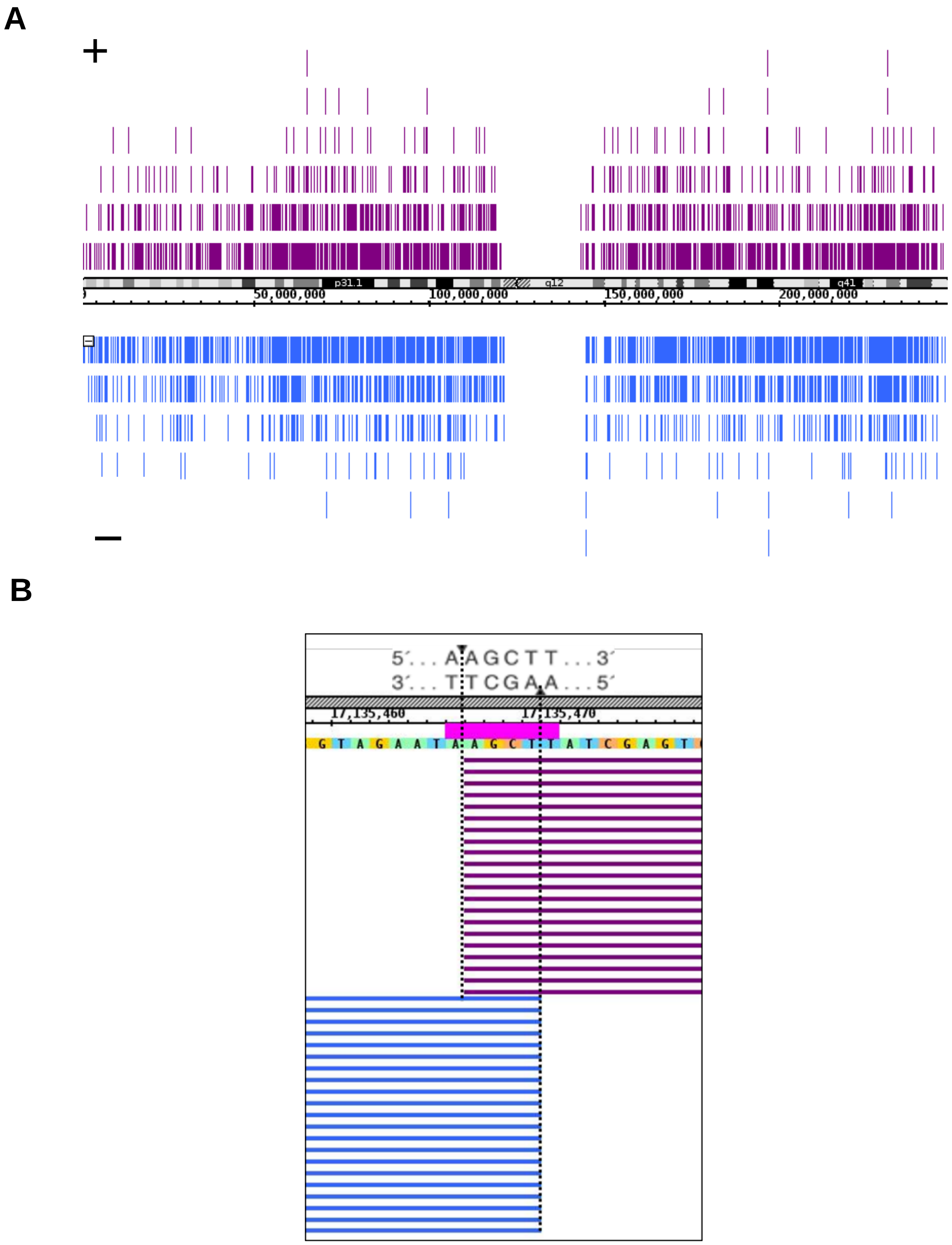

A
+
p31.1
q12
q41
50,000,000
100,000,000
150,000,000
200,000,000
–
B
5´. . . A A G C T T . . . 3´
3´. . . T T C G A A . . . 5´
17,135,460
17,135,470
G T A G A A T A A G C T T A T C G A G T

Figure 3.9. Genome browser (IGB) representation of DSBs induced by HindIII. (**A**) Snapshot of all DSBs identified by INDUCE-seq that map to chromosome 1. Breaks tagged on the right side of the break are mapped to the plus strand (purple) and breaks tagged on the left side of the break are mapped to the minus strand (blue). (**B**) HindIII induced DSB sites are easily distinguishable by their overlapping sequence. 4bp of the 6bp HindIII restriction motif overlaps at the start position of each read, corresponding to each side of the labelled DSB.

3.2.6 INDUCE-seq assay optimisation

Having established that DSBs can be detected by flow cell enrichment, the next step was to improve the workflow in order to increase the yield of DSBs measured from a cell sample. Following ligation of the sequencing adapter to both ends of the DNA fragments, the size distribution is expected to increase by the length of two adapter sequences; the P7 adapter is approximately 60bp in length, so the size distribution should increase by ~120bp. An increase of just 50bp (**Figure 3.7**) suggests that the library preparation protocol used for the proof of concept INDUCE-seq experiment was not efficient. Many fragments were therefore not ligated correctly for sequencing, rendering many labelled DSBs unsequenceable. In order to address this issue, the INDUCE-seq proof of concept library preparation protocol, which used a variety of DNA modifying enzymes, was exchanged for a commercially available library preparation kit (**2.1.8**), and the experiment was repeated. In addition to introducing a new library preparation, multiple samples were processed using different P7 index sequences, which allowed the sequencing of multiple sample repeats simultaneously. Library yield increased significantly using the improved library preparation, as evidenced by the increase in average fragment size from 280bp input sonicated DNA to the 400bp output library (**Figure 3.10**). This shift also had a pronounced effect on the concentration of libraries generated as measured by qPCR (**Figure 3.11**) (**Appendix C, Table A3**). Untreated WT control libraries increased in concentration by an average of 10-fold, and HindIII treated samples increased by an average of 68-fold, showing the effect of the improvement in library yield on the number of sequenceable break ends measured.

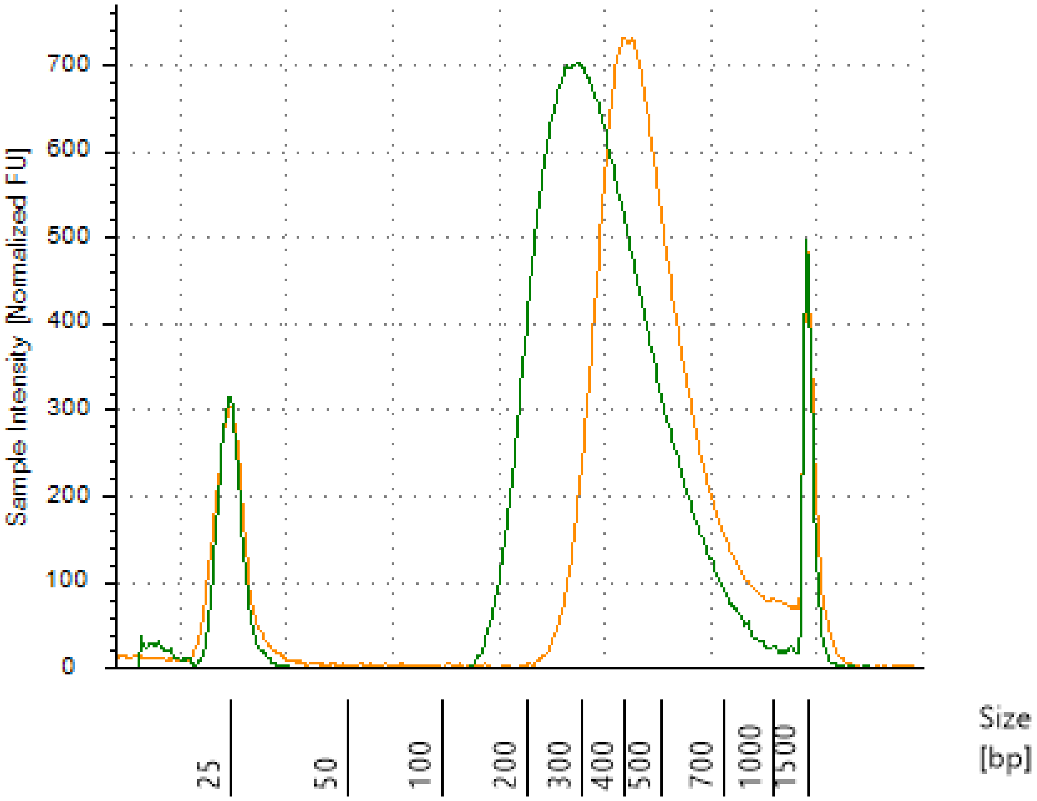


Figure 3.10. Fragment size distribution following modifications to INDUCE-seq library preparation. Enhanced ligation efficiency is shown by shift in fragment size distribution from sonicated DNA (green line, ~300bp average) to the final sequencing library (orange line, ~450bp average).

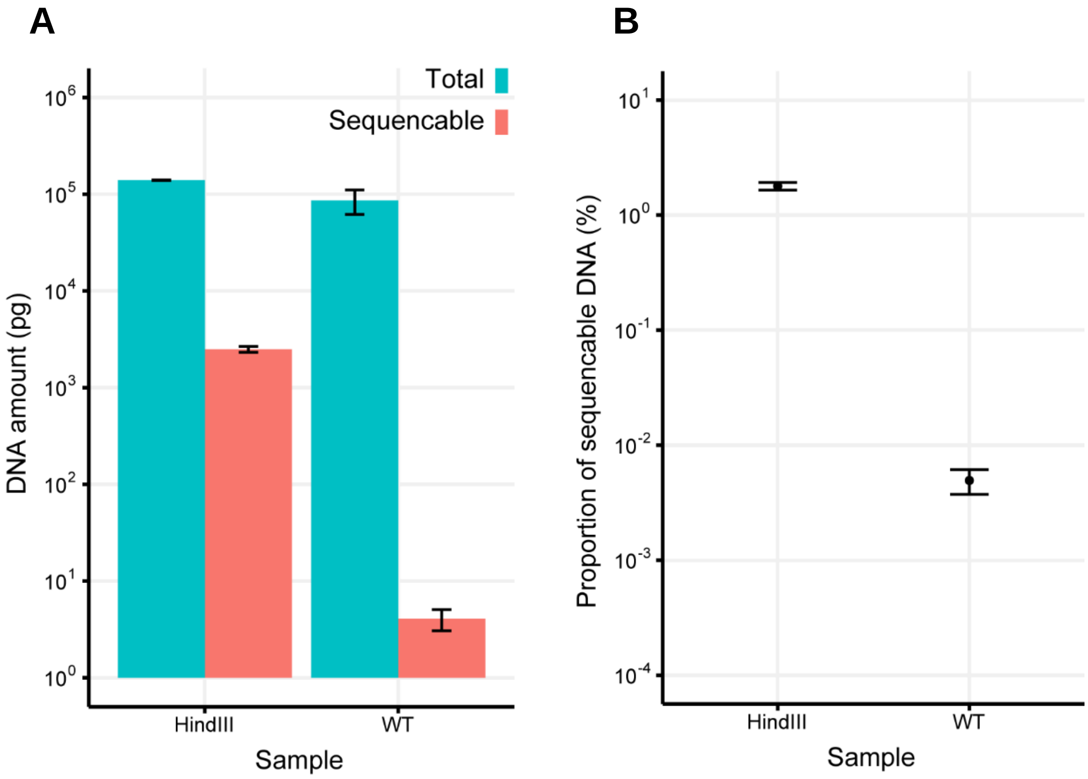


Figure 3.11. Quantification of INDUCE-seq libraries following modified library preparation. (**A**) The total amount of DNA (blue bars) and the sequenceable amount of DNA (pink bars) measured untreated WT INDUCE-seq sequencing libraries (n=5) and HindIII treated INDUCE-seq sequencing libraries (n=2). (**B**) The proportion of sequencable DNA for both library types. Error bars as standard deviation (SD).

In addition to enhancing the efficiency of library preparation, the yield of DSBs measured by INDUCE-seq was further enhanced by altering the choice of Illumina sequencing platform. As the flow cell enrichment component of INDUCE-seq is dependent on the efficiency of the underlying flow cell chemistry to capture sequenceable molecules, the choice of sequencing flow cell has a direct impact on the number of reads that can obtained. Each Illumina flow cell has a different hybridisation efficiency, which can be calculated from the number of recommended input molecules loaded, and the number of reads generated (**Appendix C, Table A4**). Of the available sequencing options, the MiSeq nano, which was used for the proof of concept INDUCE-seq experiment, is the least effective at enriching on the flow cell. Theoretically, using the MiSeq nano flow cell loading protocol, at best only 0.03% of all sequenceable molecules can hybridise to the flow cell and subsequently sequenced. The NextSeq and HiSeq sequencers offer vastly more efficient use of input material, with relative efficiencies of up to 31% and 33%, respectively. This indicates that both of these platforms offer flow cell loading strategies that can hybridise more material, and therefore more effectively facilitate flow cell enrichment. This allows the sequencing of libraries, without prior amplification or high levels of input material.

For these reasons, the NextSeq 550 sequencing instrument was selected, which provides vastly improved sample hybridisation efficiency and yields reads at a scale suitable for a small number of samples. The final stage of optimisation involved altering the denaturation and dilution protocol to load as much of the INDUCE-seq library onto the flow cell as possible. Briefly, sample denaturation and dilution prepares DNA for flow cell hybridisation by melting DNA strands (typically using NaOH), followed by diluting to the correct flow cell loading concentration in a hybridisation buffer. Because standard sequencing libraries are generated via amplification in excess, often in the nM concentration range, the standard protocol for denaturation incorporates a final dilution step where only 12% of the final sample is loaded on to the sequencer. This does not represent a problem for standard PCR-amplified libraries, which contain many duplicates of each DNA fragment. For INDUCE-seq, however, this would result in the loss of almost 90% of the labelled DNA breaks in the library, because all of the starting material is taken through to sequencing. To mitigate against this, the dilution step of the standard protocol was omitted so that INDUCE-seq libraries could be denatured, diluted, and loaded on to the flow cell without sample loss (**2.1.8**).

3.2.7 The detection of *in vitro* induced DSBs using INDUCE-seq

To test the optimised INDUCE-seq library loading and sequencing on the NextSeq 550, six untreated samples, and one sample *in situ* digested using HindIII, were selected for sequencing. Due to the vastly greater library concentration observed for the HindIII treated sample, only 25% (~25,000 cells) was used for sequencing in order to ensure that the concentration of the library pool was not greater than the recommended loading concentration. For the HindIII treated sample, the improved workflow vastly increased the number of breaks detected per cell from 1.14 to 5979, representing a 5,245-fold increase (**Table 3.2**). In untreated cells, INDUCE-seq detected endogenous breaks consistently between samples averaging ~2 breaks per cell. Interestingly, in the untreated samples recurrent break positions within the population made up the minority of breaks detected, with just 2.6% of break positions detected in more than one cell. In contrast, the number of recurrent break positions detected in the HindIII treated sample made up the majority, with 71.6% of all positions detected in multiple cells. Perhaps most striking is the vast difference between the total number of breaks detected from the HindIII treated sample compared to the number in the untreated samples, which is visualised in **Figure 3.12**. INDUCE-seq measured almost 150 million break ends from ~25,000 HindIII treated cells, compared to an average of just 195,000 endogenous breaks from ~100,000 untreated cells. This represents a cell number-adjusted difference of 3,000-fold more breaks measured in the induced sample. By removing PCR and enriching for breaks using the flow cell, INDUCE-seq can simultaneously detect hundreds of millions of highly recurrent HindIII-induced DSBs, in addition to hundreds of thousands of lower-level endogenous DSBs from within the same sample, over a range of three orders of magnitude (**Figure 3.12, green highlight**). The problem of high levels of background breaks that limits alternative PCR-based DSB detection methods is no longer present as a result of this, generating an output with a ‘background’ that is representative and proportional to the actual endogenous break landscape present in the cells at the time of crosslinking.

Table 3.2. The number of DSB identified in untreated and HindIII treated cell samples.

Sample	Break ends	Break positions	Cells	Approximate breaks per cell
WT1	213,725	206,837	100,000	2.1
WT2	253,296	245,233	100,000	2.5
WT3	196,147	189,800	100,000	2.0
WT4	121,437	117,432	100,000	1.2
WT5	190,330	184,594	100,000	1.9
HindIII	149,477,504	2,205,884	25,000	5979.1

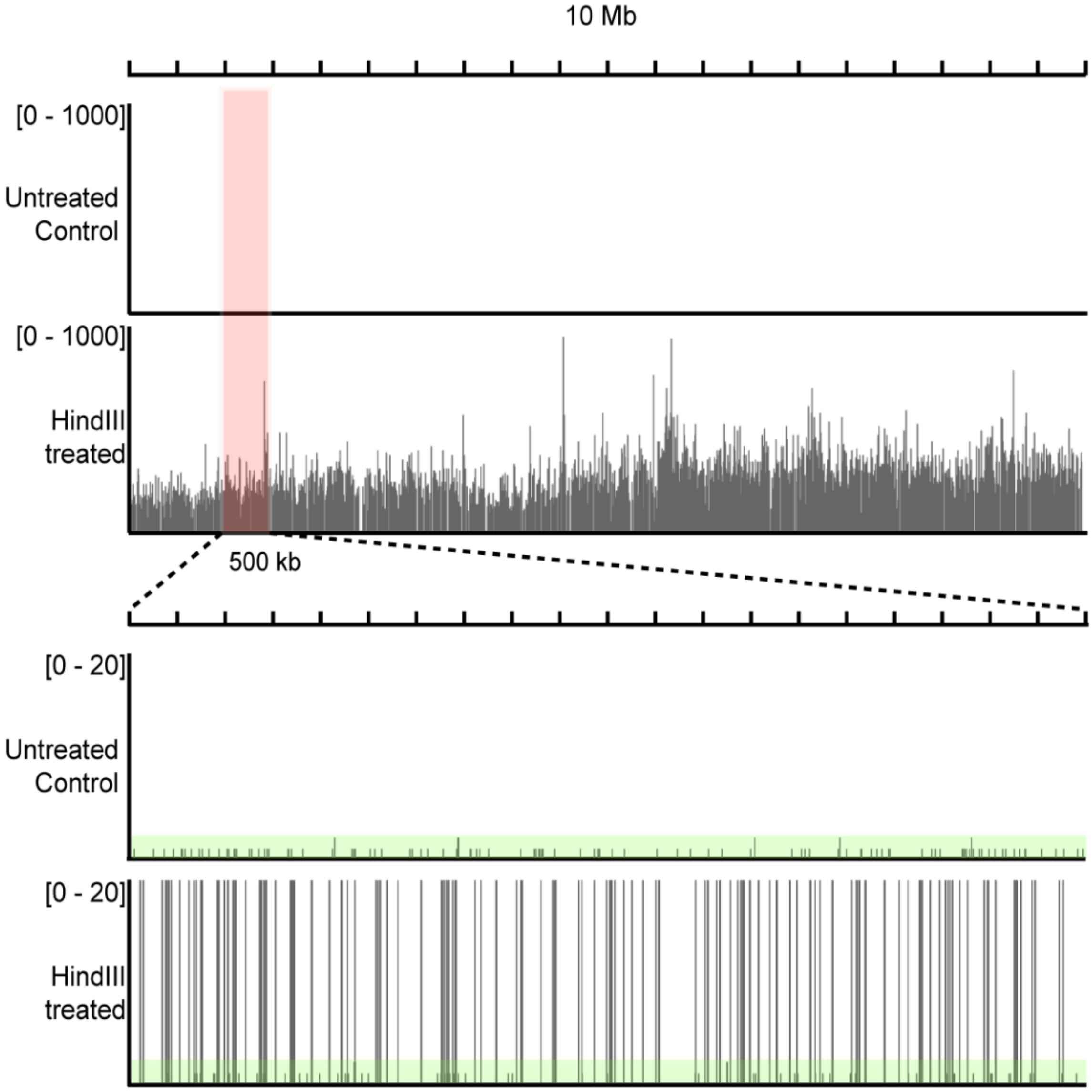


Figure 3.12. INDUCE-seq detects both highly recurrent induced DSBs and low-level endogenous DSBs simultaneously with high resolution. Genome browser view (IGV) of INDUCE-seq reads mapped to a 10Mb section of the genome from HEK293T cells following *in situ* cleavage with the restriction endonuclease HindIII. (**Top panel**) Highly recurrent enzyme-induced breaks represent the vast majority of reads when viewed at low resolution (10Mb, 0-1000 reads). (**Bottom panel**) A high resolution view (pink highlight, 500kb, 0-20 reads) reveals low level single endogenous breaks present in both the untreated sample and amongst the recurrent HindIII-induced breaks (Green highlight).

To better understand the sensitivity of INDUCE-seq, the readout generated from HindIII treated cells was compared with an equivalent experiment where the DSB labelling methodology DSBCapture was used to measure EcoRV induced breaks (Lensing et al. 2016). To achieve a representative comparison with DSBCapture, HindIII-induced breaks measured by INDUCE-seq were determined based on the same parameters used in the DSBCapture study. Direct overlapping of ≥5 reads, with a start position mapped precisely to the expected cleavage site was used to filter low confidence / 'background' reads, resulting in a set of high-confidence restriction endonuclease-induced break sites. Of the total raw reads generated from sequencing, a greater proportion of INDUCE-seq reads were mapped to HindIII restriction sites than DSBCapture reads mapping to EcoRV sites (**Figure 3.13A**). A similar pattern was observed when comparing reads that were aligned to the genome. In this case, 96.7% of aligned INDUCE-seq reads were mapped to restriction sites, representing a 25% improvement in fidelity of break detection over DSBCapture. Significantly, INDUCE-seq uses 800-fold fewer cells than does DSBCapture (25,000 vs 20M, respectively), whilst identifying a similar proportion of HindIII restriction sites (92.7%, 778,563 / 838,710) to the EcoRV sites identified by DSBCapture (93.7%, 403,750 / 430,897) (**Figure 3.13B**). This result demonstrates that by removal of the amplification and enrichment steps that introduce background noise, INDUCE-seq is able to achieve similar levels of induced-break detection sensitivity using a sample derived from 800-fold fewer cells.

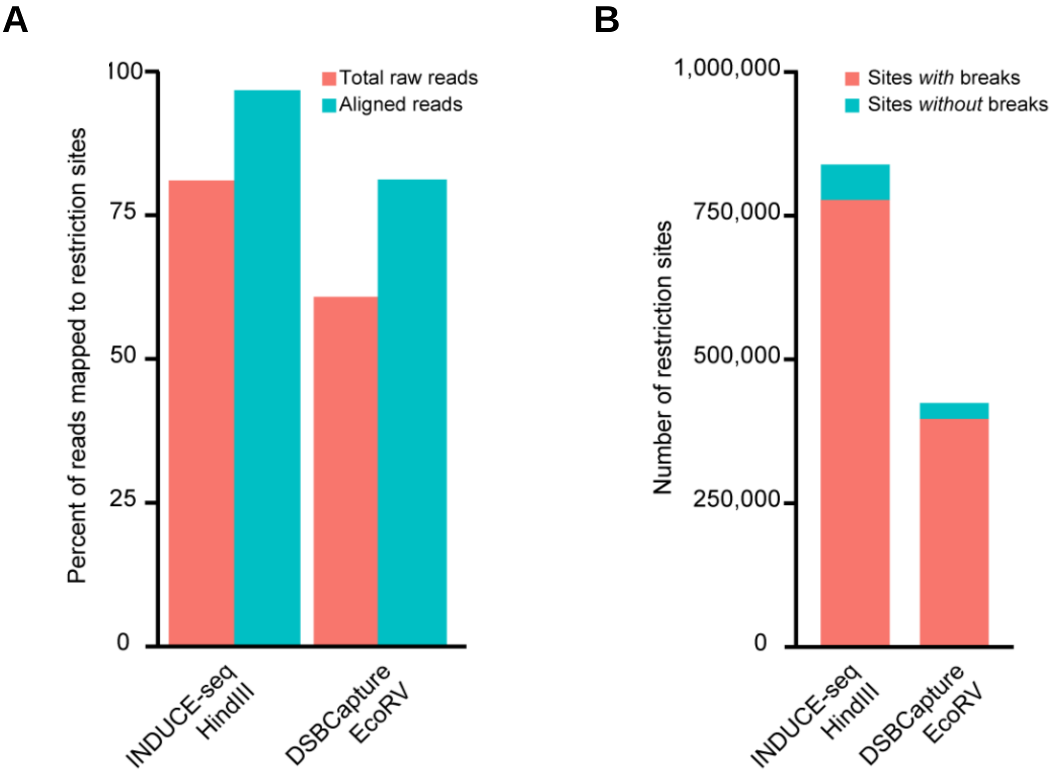


Figure 3.13. Comparison between INDUCE-seq and DSBCapture in detecting *in vitro* cleaved restriction sites by the enzymes HindIII and EcoRV. (**A**) A greater proportion of reads sequenced and aligned to the genome were mapped to restriction sites using INDUCE-seq. (**B)** Using 800-fold fewer cells, INDUCE-seq identifies a similar proportion of HindIII restriction sites (92.7%) to that identified by DSBCapture for EcoRV (93.7%).

To further characterise the sequence context around breaks detected by INDUCE-seq, the frequencies of each nucleotide 25 bp either side of breaks were calculated and represented as logo plots (**Figure 3.14**). Consistent with previous findings, plotting breaks in this manner using the HindIII treated sample revealed a perfect HindIII recognition sequence, consistent with the fact that 96.7% of the breaks in sample were mapped to on-target HindIII break sites (**Figure 3.14A**). Interestingly, a weak asymmetric sequence motif can also be seen around endogenous break sites in the untreated sample (**Figure 3.14B**). This is most prominent within 5bp up- and downstream of the break and, the 5' > 3' sequence around breaks was extremely consistent between technical repeats, showing the same pattern regardless of the side of the break that was labelled. This observation may indicate a degree of sequence bias around endogenous break sites, potentially revealing a preference for break formation at particular sequences in the genome.

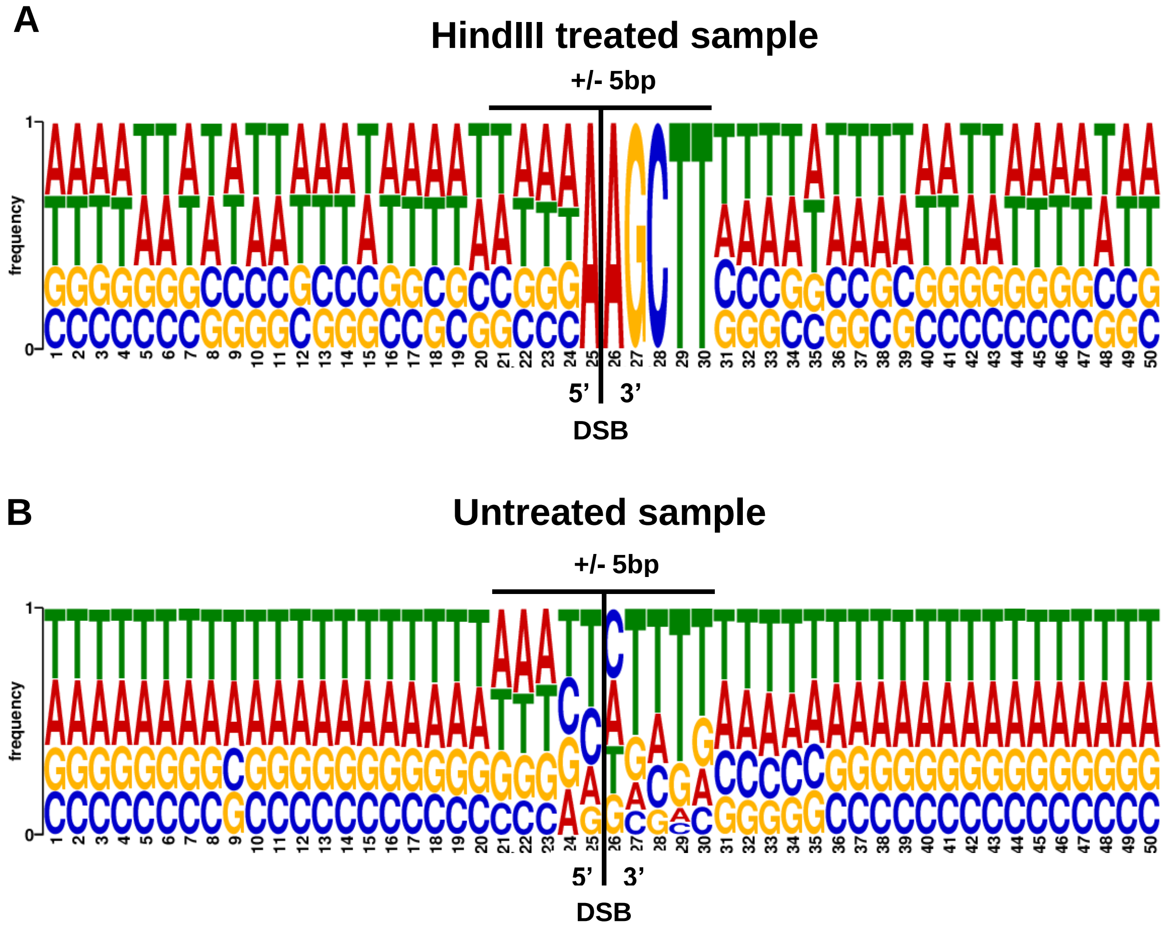


Figure 3.14. Sequence logo at induced and endogenous breaks. The frequency of each nucleotide 25bp upstream and downstream of break positions for the HindIII treated and untreated samples. The DSB position is depicted between base positions 25 and 26, and all sequences are corrected for 5' > 3' orientation based on the strand that the breaks were mapped to. (**A**) HindIII treated sample representing induced and endogenous breaks. Most breaks (96.7%) are mapped to the HindIII restriction sequence AAGCTT. (**B**) Untreated sample representing endogenous breaks. Within 5bp of the break site the biggest differences in base frequency are seen, whereas the extremes (positions 1-20 and 31-50) show a base frequency that averages the sequence composition of the genome, at approximately 40% GC content.

3.2.8 Quantification of off-target break induction by HindIII

Having confirmed the ability of INDUCE-seq to detect on-target break induction by HindIII at the AAGCTT recognition sequence, it is now possible to examine the fidelity of these enzymes by precisely quantifying the off-target sites comprised of redundant recognition sequences. To determine the dynamic range of off-target discovery using INDUCE-seq, cleavage events at 1bp and 2bp mismatching off-target sequences for HindIII were quantified (**Figure 3.15**). HindIII off-targets were identified using the same criteria for the on-target sites (i.e. ≥5 reads mapped to the restriction site). For a restriction endonuclease with a 6bp recognition sequence, there are 18 possible 1bp mismatching sequence variants, and 135 possible 2bp mismatching sequence variants (**2.2.2**). INDUCE-seq detected 496,200 off-target DSBs at sites bearing a single mismatch to the HindIII recognition sequence, which corresponded to all 18 theoretical 1bp mismatching sequence variants and was spread across 4,702 sites in genome (**Figure 3.15, orange bars**). The number of breaks corresponding to off-target sequences with only 1bp mismatch ranged from 8.0x10^4 to 6.5x10^5, the lowest of which represents a ~2,200-fold reduction compared to the number of breaks found at canonical HindIII sequences. In addition to sites with a 1 bp mismatch, 708,138 breaks were identified at 2 bp mismatching sites, which corresponded to 65 of the 135 possible 2 bp sequence variants, and was spread over 7,843 sites in the genome (**Figure 3.15, green bars**). The variance of breaks found for each sequence type was greater with 2 bp than the 1 bp set, ranging from the highest at 3.0x10^6, to the lowest represented by a single site with 5 breaks. This demonstrates that INDUCE-seq is capable of measuring induced breaks with a sensitivity of over six orders of magnitude.

The simultaneous identification of ~150 million breaks at on-target sites, in addition to a variety of off-target sites occurring with as few as 5 reads demonstrates the exquisite sensitivity of INDUCE-seq for detecting both highly recurrent, and rare events simultaneously within the same complex sample. This represents a dynamic range of quantitative break detection of seven orders of magnitude using INDUCE-seq. This is uniquely enabled by the PCR-free nature of the assay, which eliminates the effects of amplification-induced noise, and prevents saturation of the Illumina sequencing flow cell by many amplified copies of highly frequent events. These results demonstrate that INDUCE-seq has the potential to simultaneously measure both rare endogenously formed and highly abundant induced breaks throughout the genome within the same sample. The ability to do this has not before been possible with existing PCR-based approaches.

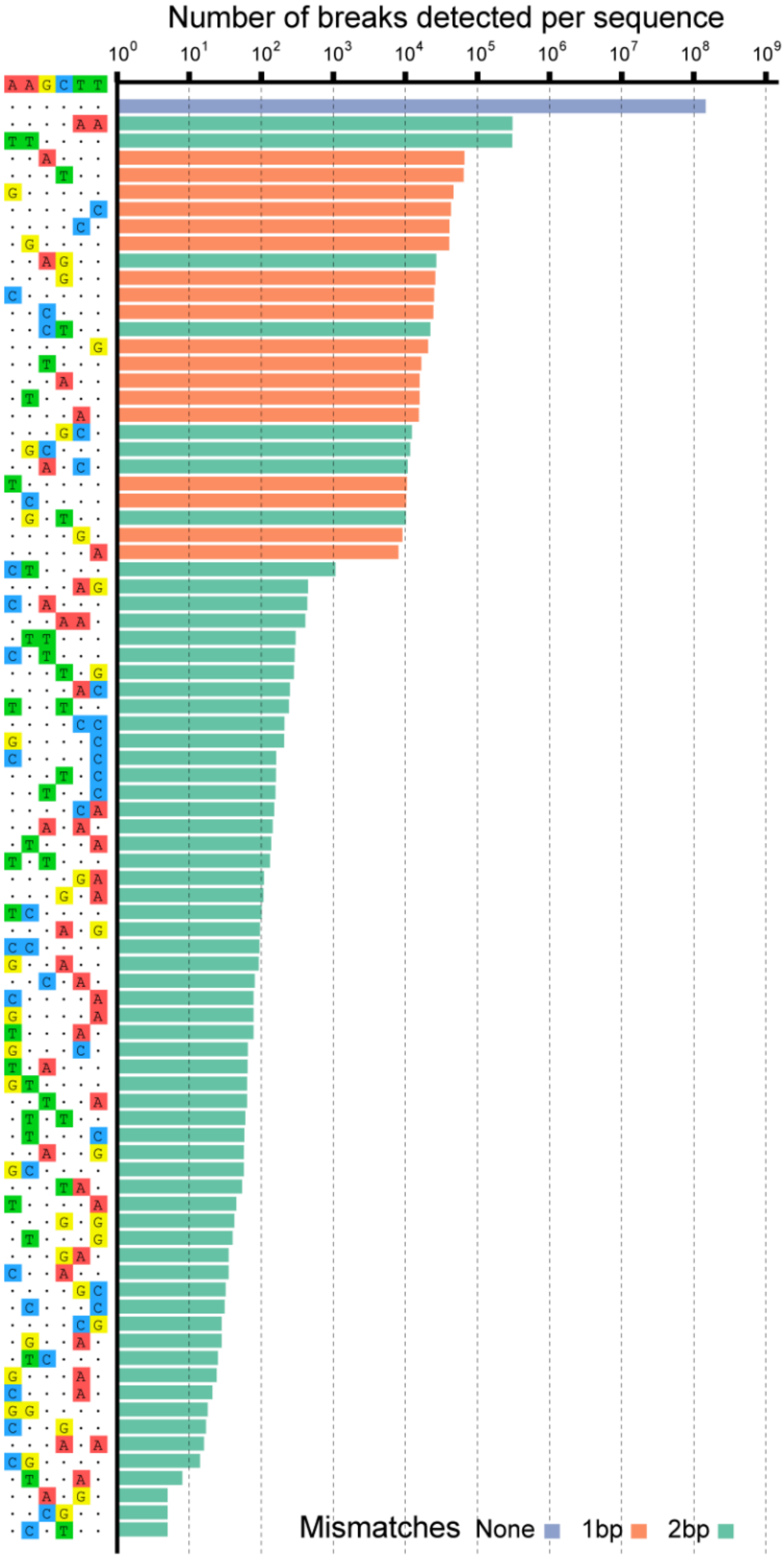

Number of breaks detected per sequence
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8
10^9
A A G C T T
Mismatches
None
1bp
2bp

Figure 3.15 The dynamic range of induced DSB detection using INDUCE-seq. In addition to breaks identified at HindIII on-target sequences (AAGCTT), multiple 1bp and 2bp mismatching off-target sites were also identified. INDUCE-seq measured induced break events spanning 8 orders of magnitude, from ~150 million breaks identified at HindIII on-target sites, to 5 breaks identified at the least frequent off-target.

3.3 Discussion

In this chapter, the development of a novel method, INDUCE-seq, to detect genomic DSBs has been established. INDUCE-seq takes advantage of an unconventional use of the Illumina flow cell as an enrichment tool, which enables the measurement of genomic DSBs without suffering from the same limitations associated with alternative methods that are currently available. The INDUCE-seq adapter design, combined with the spatiotemporal separation of ligation reactions, and subsequent flow cell enrichment, eliminates the need for a dedicated amplification or affinity capture-based enrichment of DSBs up-steam of sequencing. This circumvents the need for PCR-based library amplification to generate sequencing ready libraries. It is noteworthy that INDUCE-seq's novel strategy of library preparation and flow cell enrichment is not limited to the detection of DSBs. With careful adapter design, a range of applications can be conceived of that selectively sequence a variety of specific genomic features. The initial sequencing experiments described here confirm that this novel approach is effective for the sequencing of DSBs, and demonstrate how the process significantly improves the quantitative nature of NGS-based break detection. Contrary to PCR-based methods, which by definition sequence many copies of break-tagged DNA fragments, INDUCE-seq generates a unique digital sequencing output where each read exclusively represents an individual DSB end. In addition to this quantitative enhancement, a lack of prior signal amplification enables the measurement of events across a wider dynamic range than is achievable using PCR-based library preparation methods. Within a single sequencing run, both rare and frequent DSB events can be detected simultaneously without the need for excessive sequencing read depth to capture rare events, or the risk of oversaturation by amplification of high frequency events. Furthermore, as each DNA fragment that was sequenced represents a single labelled DSB event, we bypass the need for unique molecular barcodes to control for PCR duplicates and other amplification artefacts, greatly simplifying the bioinformatics analysis pipeline (Yan et al. 2017).

INDUCE-seq also addresses the caveat of poor scalability that limits alternative approaches. The design of the modified P5 and P7 INDUCE-seq adapters facilitate this - most significantly by allowing DSB enrichment directly on the Illumina flow cell. This eliminates laborious upstream enrichment steps prior to library preparation. Furthermore, as INDUCE-seq is amplification free, each sequencing read that is generated represents a single break end, maximising the cost effectiveness of Illumina sequencing which is commonly cost limiting for NGS experiments. Further improvements to the INDUCE-seq adapter design in the form of end-blocking modifications alleviate the formation of adapter:adapter dimers, which otherwise require specific washing or enzymatic treatment steps to remove. Starting with fixed cells or tissue samples in 96-well format, the process requires just 2-3 days of labour to generate sequencing ready libraries. INDUCE-seq has been optimised for automatable, high-throughput processing, and provides the fastest turnaround time of any cell-based genome-wide DSB detection method available to date.

The establishment of INDUCE-seq demonstrates that flow cell enrichment is undoubtedly a powerful tool for the separation of differentially labelled DNA fragments, which may be expanded for a variety of applications. Innovative modifications to the INDUCE-seq P5 adapter and initial DSB end-processing steps would allow the capture of subsets of DSBs based on their end structures. For example, using an adapter with a 5' overhanging 'sticky end' for the detection of 3' single-stranded ends that are generated as a prerequisite for HR. This same methodology may also be applied to the detection of different types of DNA damage. Many types of DNA lesions affecting single strands, such as UV induced cyclobutane pyrimidine dimers (CPDs) or 8-oxogunaine (8-oxoG) induced by reactive oxygen species (ROS), can be excised enzymatically to produce DNA nicks, which can subsequently be ligation tagged in a similar manner to 3' resected ends (Mao et al. 2016; Wu et al. 2018). Furthermore, single strand breaks (SSBs) or abasic sites would also be susceptible to capture in this manner. In addition to DNA damage detection, target enrichment of specific regions of DNA may be another application of flow cell enrichment. Using the CRISPR-Cas9 system, targeted DSB induction at genomic locations of interest would allow them to be ligation captured for flow cell enriched sequencing. Broad genomic features such as gene coding regions in addition to single positions such as known mutation hotspots could be targeted and sequenced in this manner, dependant on the availability of appropriate CRISPR platforms and targeting design.

To demonstrate the features of the INDUCE-seq methodology, these first experiments examined how INDUCE-seq measures DSBs genome-wide following the induction of defined DSBs in fixed and permeabilised cells using a high-fidelity HindIII restriction endonuclease. Treating cells in this manner was a simple approach to benchmark INDUCE-seq, as it produces very high numbers of breaks at predefined positions in the genome. It was particularly important for proof of concept INDUCE-seq experiments to start with very large numbers of breaks at known positions, as it was not clear how efficient the flow cell enrichment strategy would be. During the proof of concept experiments, HindIII treated samples provided a means to optimise the experimental conditions of INDUCE-seq and enhance the efficiency of break capture. Having optimised the INDUCE-seq protocol, the HindIII treated sample produced a distinctive dataset containing millions of on- and off-target positions that were cleaved throughout the genome. These induced break sites ranged from as few as the 5 break threshold, to some that were represented >1,000 times. These findings show that INDUCE-seq has an unparalleled dynamic range and is capable of the simultaneous measurement of single events such as endogenous breaks and rare off-target induced breaks, as well as millions of high- frequency events from the same sample. This feature of INDUCE-seq cannot be achieved by a PCR-based approach as high-frequency events amplify exponentially and saturate the sample fragment composition, while rare events are prone to being underrepresented. Because of this, the INDUCE-seq readout is proportional to the capacity and efficiency of the Illumina sequencer being used. These findings suggest that the quantitative improvements made by INDUCE-seq should also allow the identification of widespread and random break induction throughout the genome by genotoxic agents such as chemotherapeutic drugs and ionizing radiation, of which a genome-wide, high-resolution measurement does not currently exist.

With the improved characteristics described in this section, INDUCE-seq shows great potential for measuring genome-wide off-target break induction by CRISPR-Cas9 used in genome editing. In particular, the enhanced sensitivity, lack of signal distortion, and wide dynamic range, should enable INDUCE-seq to profile off-target breaks in a proportionate and quantitative manner for the first time. In addition to an accurate measurement of induced breaks, robust CRISPR off-target discovery relies on the characterisation of the endogenous break background for any edited cell. This is particularly important if very low frequency induced breaks are to be identified, which can potentially be represented by an individual INDUCE-seq read. In the next chapter, INDUCE-seq is applied to the measurement of endogenous breaks in a variety of cell types as well as a live-cell restriction endonuclease inducible system to validate its use for measuring DSB induction in live cells.

4 Chapter IV - Using INDUCE-seq for the live cell analysis of endogenous and induced breaks

4.1 Introduction

In the previous chapter, the development of INDUCE-seq was described and demonstrated using HindIII *in situ* digested HEK293T cells. While useful for establishing the method and demonstrating the characteristics of break capture, these experiments do not represent breaks formed at physiologically relevant frequencies, as the cells were crosslinked prior to chromatin digestion. Prior to the application of INDUCE-seq for the measurement of CRISPR-Cas9 off-target DSBs, it was important to benchmark the assay and to validate its use for the capture of enzyme-induced breaks in live cells. This is because benchmarking is not easily done directly using CRISPR off-target DSB induction, as for off-target discovery there are two unknown elements to consider, the location and frequency of off-target events. For experiments using the same guide RNA, CRISPR off-target locations and frequencies can differ greatly between different cell types, cellular delivery format, and time points of sampling (Zhang et al. 2015; Modrzejewski et al. 2020). Furthermore, each off-target discovery assay uses different and bespoke bioinformatics pipelines for off-target discovery, which directly impacts off-target discovery. These factors, in addition to possible differences in the background rate of endogenous break formation in different cell types used for CRISPR experiments make it very difficult to compare assays in this manner.

Given the inability to appropriately benchmark the sensitivity of INDUCE-seq using CRISPR off-targets, studies described in this chapter seek to do this using an easily controllable DSB inducible system known as DSB Inducible via AsiSI (DIvA) cells. The DIvA cell system (transformed U2OS cells) is a commonly used model for the study of DSB biology and repair kinetics and has been used by several of the NGS-based DSB capture approaches to assess assay sensitivity (Lensing et al. 2016; Yan et al. 2017). DIvA cells constitutively express the restriction endonuclease AsiSI, which is a rare 8 bp cutter fused to an oestrogen receptor. Following cellular treatment with 4-hydroxy tamoxifen (4OHT), AsiSI is localised to the nucleus where it cleaves ~100 of the ~1,200 AsiSI restriction sites in the human genome (Iacovoni et al. 2010; Massip et al. 2010; Aymard et al. 2017). A more recent version of the DIvA cell system, known as AID-DIvA, further improves the control of DSB induction by introducing an auxin inducible AsiSI degradation which enables ‘switching off’ of DSB induction (**Figure 4.1**). Although AsiSI has the potential to target ~1,200 sites in the genome, fewer cut sites are observed because the enzyme is CpG methylation sensitive and is potentially blocked by chromatin structure. The benefits of using the DIvA cell system for benchmarking INDUCE-seq are twofold; firstly, comparison between methods can be performed using the same cell line, treated in an identical manner, and secondly the location of AsiSI cleavage sites is easily determined *in silico* simply by scanning the genome for the 8 bp restriction sequence (**2.2.4**).

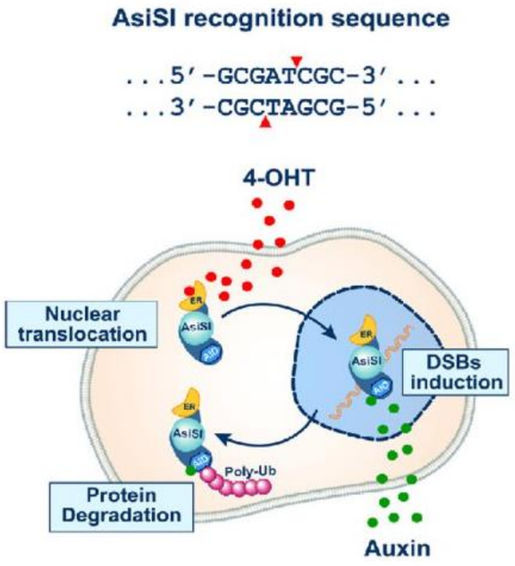


Figure 4.1. Schematic of DSB induction in the AID-DIvA cell system. Following cellular treatment using 4-OHT, the constitutively expressed AsiSI enzyme is translocated to the nuclease where DSB induction occurs at the 8 bp restriction sequence. DSB induction can then be inhibited by the addition of Auxin, which results in AsiSI protein degradation. Adapted from: Mladenova *et al.* 2016.

In addition to benchmarking INDUCE-seq using the DIvA cell system, and given the fact that the background rate of endogenous break formation may influence CRISPR-off-target DSB discovery, the work described here also aims to better characterise the endogenous DSB landscape measured in a variety of cell types. In the previous chapter, a single cell type (HEK293T) was used to establish INDUCE-seq, which demonstrated a strikingly low frequency of endogenous recurrent break sites (<3%). Correspondingly, most breaks identified were individual 'background' break sites, which were seemingly distributed evenly throughout the genome. Because of the limited data available from a single cell type, it is not clear how endogenous 'singleton' breaks occur outside of clearly defined recurrent positions (>1 break at a 1bp genomic locus). This work in this chapter therefore also aims to better define how recurrent breaks are distributed throughout the genome, which in turn will aid the robust discrimination of genuine recurrent break regions from those which arise stochastically.

4.1.1 Chapter aims

The work in the first half of this chapter aims to demonstrate that INDUCE-seq accurately and reproducibly measures enzymatically-induced DSBs in a live cell system. Furthermore, using the DIvA cell system, these experiments will benchmark INDUCE-seq against published DSB measuring approaches. The second half of this chapter aims to characterise the endogenous DSB landscape using a variety of cell types. This work will focus on understanding distribution of recurrent endogenous breaks in the genome.

4.2 Results

4.2.1 The capture of AsiSI induced breaks in DIvA cells by INDUCE-seq

To benchmark INDUCE-seq accurately, comparable experiments to those conducted using DSBCapture and BLISS were designed. DIvA cell culture was performed under identical conditions to achieve this, which included an AsiSI induction period of 4 hours using 4OHT prior to break capture (Lensing et al. 2016; Yan et al. 2017). Following break induction, INDUCE-seq break capture and sequencing was performed using 4 technical repeats of treated cells and 3 repeats on untreated cells, with each sample representing ~100,000 cells. The first difference observed from this experiment with previous INDUCE-seq runs was that the number of breaks measured in both treated and untreated DIvA cells was substantially higher than those detected in HEK293T cells, averaging 2.3 million breaks across the treated samples, and 1.9 million breaks for the untreated samples (**Figure 4.2A and Table 4.1**). This represents a ~10-fold increase in break number in DIvA cells compared to an equivalent number of HEK293T cells in the previous experiment. Importantly, there was no significant difference between the total number of breaks detected (p = 0.22857, Mann Whitney test), when comparing the treated, AsiSI induced, and untreated DIvA cell samples. This demonstrates that AsiSI break induction does not increase overall break induction genome-wide. The number of breaks mapping to AsiSI restriction sites was calculated to determine the frequency of induced breaks relative to endogenous breaks and to compare between the treated technical replicates (**2.2.4**). The number of breaks occurring at AsiSI restriction sites was reasonably consistent between treated sample replicates, ranging from 2,325-3,612. This was considerably higher than the number of breaks detected at AsiSI restriction sites in the uninduced control samples, which ranged from 23-34 (**Figure 4.2B and Table. 4.1**). Furthermore, the number of AsiSI sites identified with breaks was strikingly consistent between the treated replicates, ranging from 225-233 (**Figure 4.2C and Table. 4.1**). Interestingly, in the untreated samples the number of AsiSI sites identified with breaks was very similar to the number of breaks at AsiSI sites, resulting in an average number of breaks per site of just 1.2 across the datasets. This demonstrates the lack of consistent break induction at these positions. Surprisingly, the number of AsiSI induced breaks (n = ~3,000) appeared very low as a proportion of the total number of breaks measured (n = ~2.2M). As a result, on average 10-16 breaks were found per AsiSI site, and just 0.11-0.15% of the total breaks measured were mapped to AsiSI sites (**Figure 4.2D**).

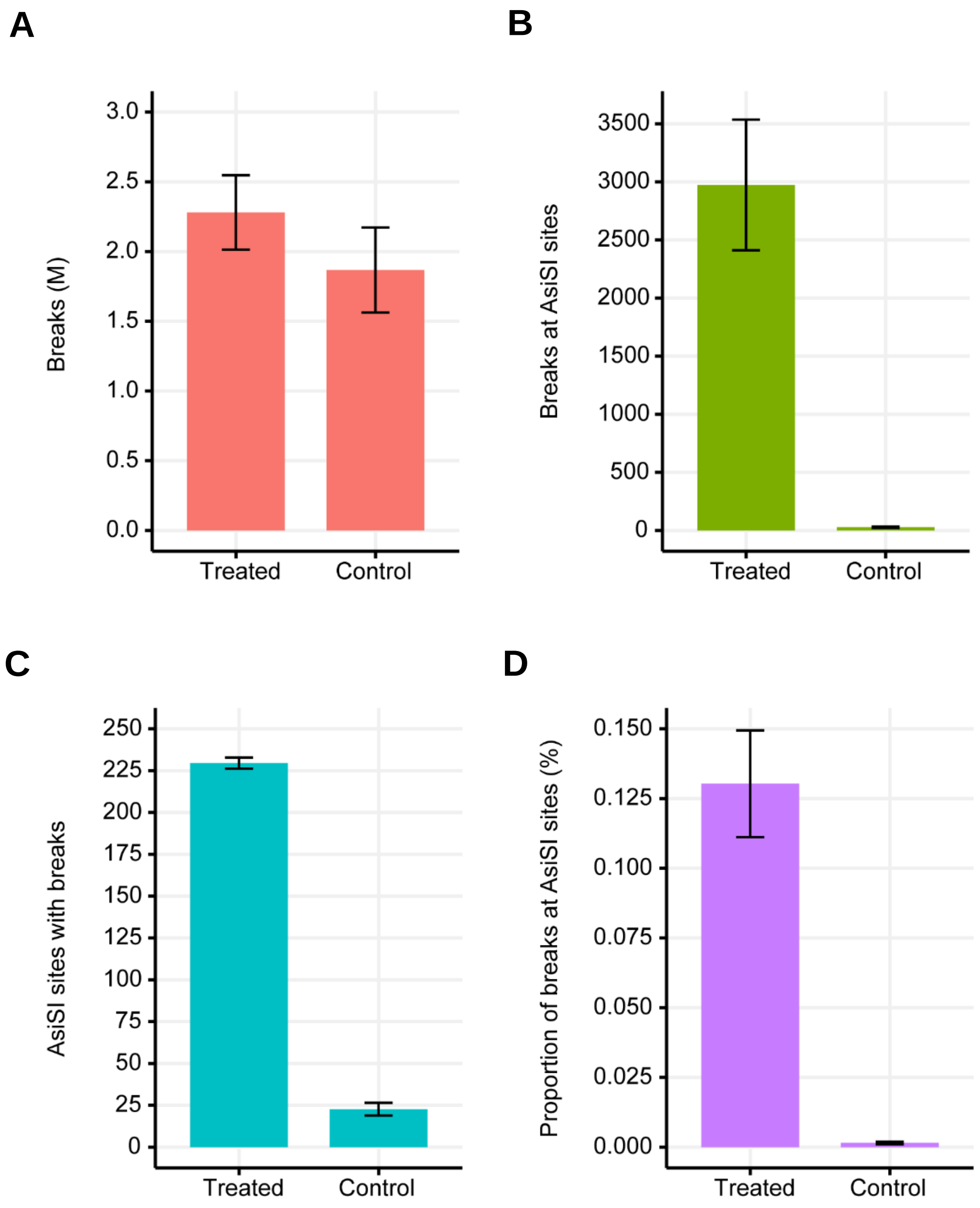


Figure 4.2 The characteristics of INDUCE-seq break capture in treated and untreated control DIvA cells. (**A**) A similar number of total breaks was observed between the treated and control samples, showing no significant difference. (**B**) Substantially more breaks were identified at AsiSI sites for the treated samples compared to the control. (**C**) Similarly, many more AsiSI sites were identified with breaks in the treated samples, which was more consistent than the break number. (**D**) Only around 0.13% of the total breaks were positioned at AsiSI sites for the treated samples, with a negligible proportion positioned at AsiSI sites for the control samples. Treated (n=4) and Control (n=3), error bars as SD.

Table 4.1. The number of DSB at AsiSI sites in treated DIvA cell samples.

Replicate (+/- treatment)	Breaks	Breaks at AsiSI sites	AsiSI sites detected with breaks	Average breaks per AsiSI site	Proportion of breaks
1 (+)	1,989,347	2,732	230	11.9	0.137%
2 (+)	2,606,023	3,222	230	14.0	0.124%
3 (+)	2,367,387	3,612	233	15.5	0.153%
4 (+)	2,157,913	2,325	225	10.3	0.108%
1 (-)	1,720,087	34	27	1.3	0.002%
2 (-)	2,217,780	24	21	1.1	0.001%
3 (-)	1,663,868	23	20	1.2	0.001%

In addition to providing a live cell example of enzymatic break induction, AsiSI induced breaks in DIvA cells also demonstrate the characteristics of break labelling by INDUCE-seq (**Figure 4.3**). The first stage of break labelling includes a blunting step which processes 3' overhangs with an exonuclease and fills 5' overhangs using a polymerase. AsiSI and HindIII therefore produce distinctive patterns of reads mapping at the break site: HindIII cuts with a 4 bp 5' overhang resulting in a 4 bp overlapping region (**Figure 4.3A**), and AsiSI cuts with a 2 bp 3' overhang, resulting in a 2 bp gap between the forward and reverse reads (**Figure 4.3B**). These results show that end structure at enzymatically induced breaks can be easily derived from the positioning of forward and reverse reads, which will enable the characterisation of a variety of CRISPR-induced breaks in the future.

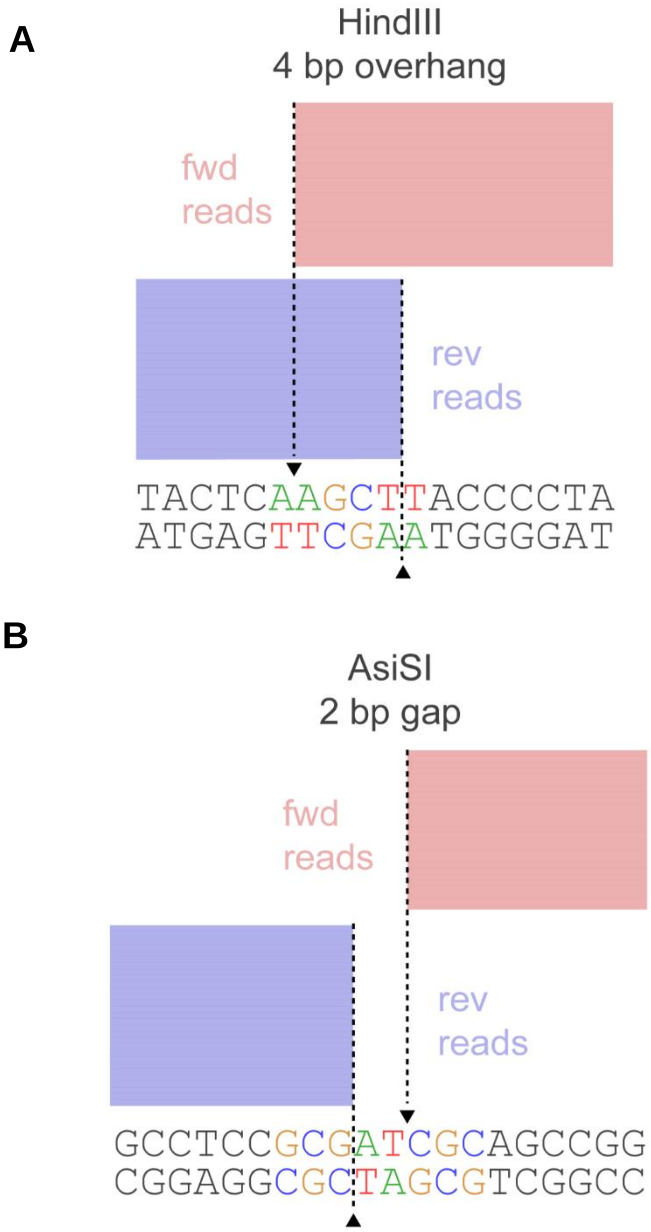


Figure 4.3. The different structures of induced strand breakage as measured by INDUCE-seq. (**A**) HindiII cleaves DNA with a 4 bp 5' overhang, which is consequently represented as a 4bp overlapping region in the INDUCE-seq sequencing reads. (**B**) AsiSI cleaves DNA with a 2bp 3' overhanging region, which is represented as a 2 bp gap between the INDUCE-seq sequencing reads.

4.2.2 Comparing INDUCE-seq DIvA replicate reproducibility

Although the total number of breaks and the number of induced AsiSI breaks and sites was similar between technical replicates, these high-level measurements do not reveal whether the distribution of AsiSI breaks at individual sites was consistent between replicates. Therefore, to further demonstrate the reproducibility of INDUCE-seq between technical repeats, the number of breaks found at each individual AsiSI break site was compared in a pairwise manner between the four replicates (**Figure 4.4A-F**). A strong linear correlation was observed between the number of breaks detected at each AsiSI site for each pair of replicates, with R^2 values > 0.88 for all pairs. The high correlation observed demonstrates that AsiSI break induction is consistently measured at equivalent levels across the same sites between different replicates. Furthermore, the correlation appears independent of total break number and/or the number of breaks found at AsiSI sites. Indeed, the strongest correlation exists between replicates r3 and r4 (R^2=0.91611), which are the two datasets with the most (r3, n = 3,612) breaks and least (r4, n = 2,325) breaks at AsiSI sites, respectively. This is also reflected by the skew of the linear correlation towards r3 on the x-axis shown in **Figure 4.4F**. The consistently high correlation between replicates that is independent of absolute break number (and hence starting cell number), shows that INDUCE-seq reproducibly and proportionally measures induced breaks in live cell populations.

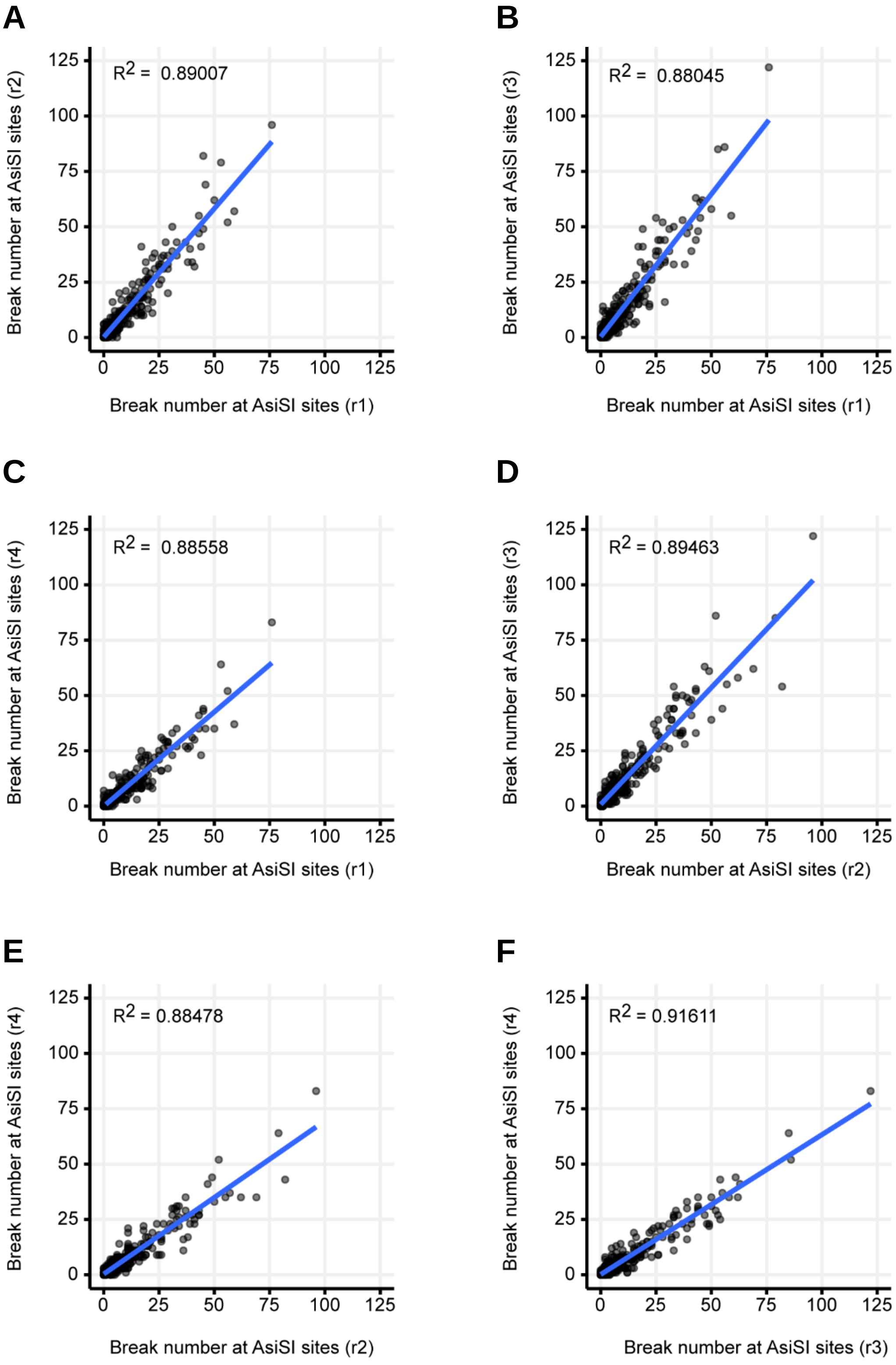


Figure 4.4. Scatterplots comparing the break number found at each AsiSI site. Technical replicates are compared in a pairwise manner: r1-r2 (**A**), r1-r3 (**B**), r1-r4 (**C**), r2-r3 (**D**), r2-r4 (**E**), r3-r4 (**F**), showing a strong correlation between all pairs.

4.2.3 Comparing the INDUCE-seq DIvA cell readout with BLISS and DSBCapture

To further validate the performance of INDUCE-seq to DSBCapture and BLISS methods, the equivalent recently published DIvA cell datasets generated were compared. This was first done by visualising the data using a genome browser depicting the aligned sequencing reads from a single replicate produced by each approach across chromosome 1 (**Figure 4.5**). For reference, a list of AsiSI restriction sites throughout chromosome 1 are displayed in the top panel. When performing a qualitative comparison of the break datasets in this manner it is important to note the differences in starting cell number and number of reads sequenced. INDUCE-seq generated ~2.5M reads from 100,000 cells, DSBCapture generated ~120M reads from 20M cells, while BLISS generated ~70M reads from 30,000 cells. Furthermore, all three methods employ different data analysis pipelines for read-duplicate removal and bias correction. Therefore, this comparison only represents sequencing reads that are aligned using equivalent settings for each method, with little postprocessing.

INDUCE-seq outputs sequencing reads with no background or noise compared to the other two methods. This results from a proportional representation of the low frequency endogenous breaks and recurrent breaks that align with the AsiSI restriction sites in the top panel. The BLISS readout by comparison displays high levels of noise, outputting signal throughout the chromosome at many positions that do not correspond with AsiSI sites. DSBCapture, on the other hand, shows an improved readout compared to BLISS, displaying far fewer highly recurrent breaks. Although reduced in number, some of these peaks do not align with AsiSI sites and the endogenous break background signal is much higher than observed for INDUCE-seq. For reference an AsiSI site detected by all three methods is highlighted (**Figure 4.5, red highlight**).

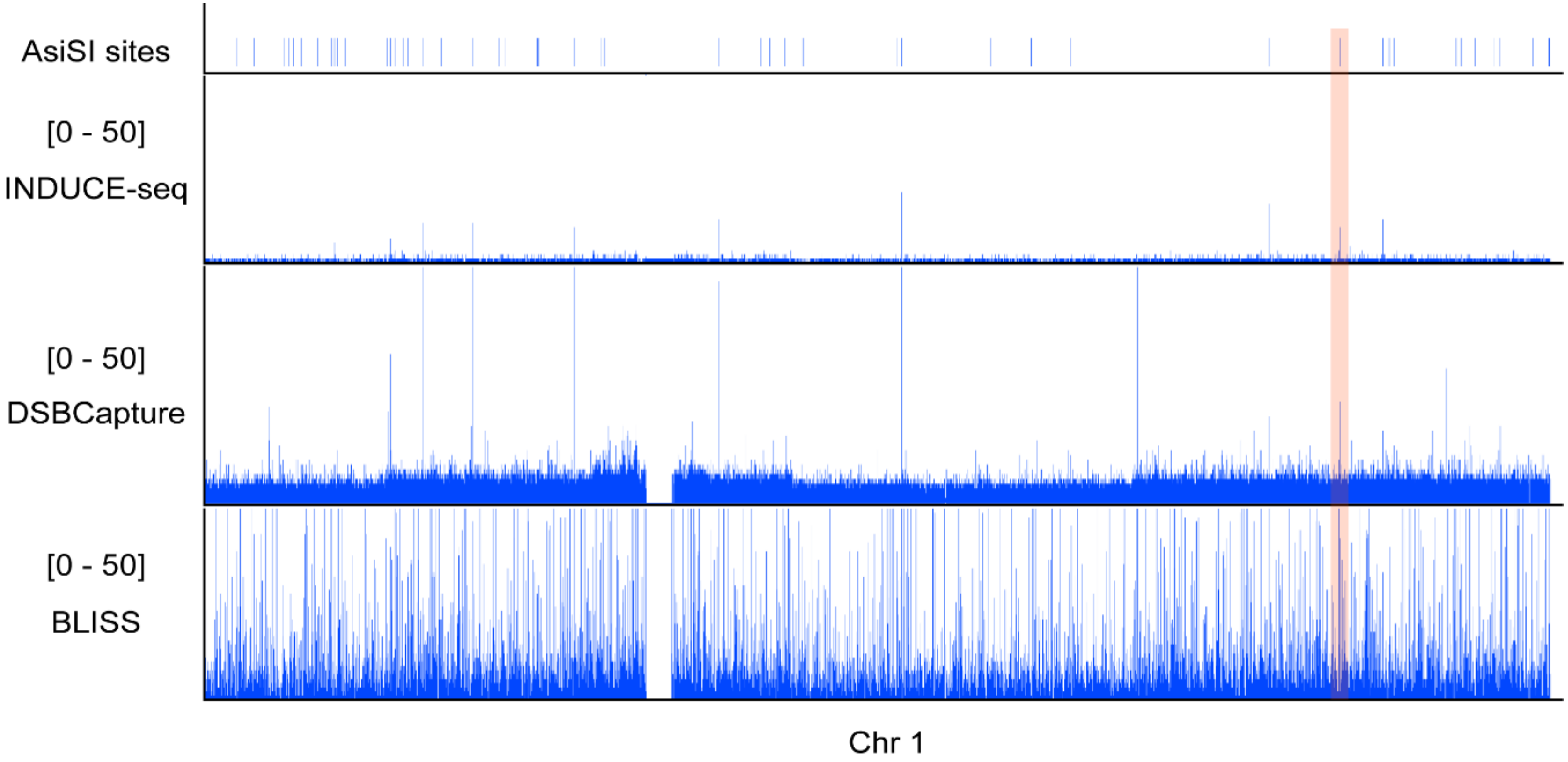


Figure 4.5. Genome browser representation of DIvA cell DSB readout by INDUCE-seq, DSBCapture and BLISS. The reads mapping to chromosome 1 are shown for each method at a scale of 0-50 reads for any given position. A list of AsiSI restriction sites are shown in the top panel. An example of an AsiSI site measured by all three methods is shown by the red highlight.

Comparing the number of AsiSI sites detected with the number of reads sequenced for each experiment further demonstrated the difference between the three methods (**Figure 4.6**). INDUCE-seq detected the presence of breaks at ~230 AsiSI sites despite sequencing 40-fold fewer reads than the corresponding DSBCapture experiment, and 23-fold fewer reads than BLISS (**Figure 4.6A**). This represents an increase over the 214 sites detected by BLISS, and 121 by DSBCapture (**Figure 4.6B**). This demonstrates the vastly enhanced efficiency, and cost-effectiveness of INDUCE-seq.

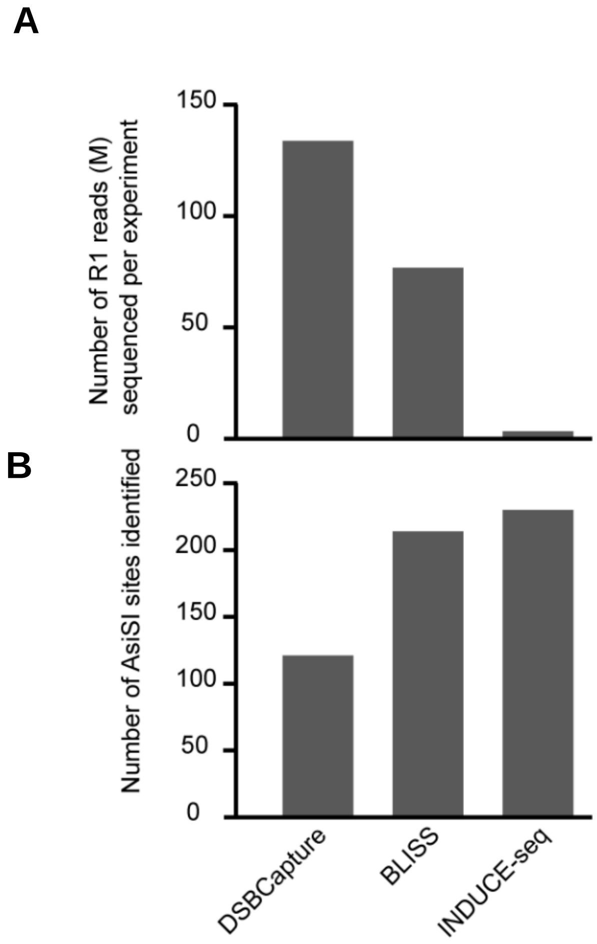


Figure 4.6. Comparison between INDUCE-seq, DSBCapture and BLISS in detecting AsiSI induced breaks in live DiVA cells. The number of reads sequenced (**A**) is compared to the number of AsiSI sites identified for each experiment. (**B**) INDUCE-seq detects the greatest number of AsiSI sites using 40-fold fewer reads than DSBCaptue and 23-fold fewer reads than BLISS.

To determine whether INDUCE-seq had detected the same, or different, AsiSI sites as BLISS and DSBCapture, AsiSI site overlaps were calculated between each of the datasets. **Figure 4.7A** shows the overlaps between a single INDUCE-seq replicate (r1) and BLISS and DSBCapture. INDUCE-seq identifies almost all the DSBCapture set (93.4%, 113/121), as well as most of the BLISS dataset (89.3%, 191/214). All datasets contained unique sites that were not detected by any other method, which might suggest differences in the sample biology or technical differences in the capture method used. To reveal whether increasing the sample size (i.e. increasing starting cell number), would impact the number of individual AsiSI sites detected, a break dataset was generated by combining all four INDUCE-seq technical replicates, and the same analysis was repeated. **Figure 4.7B** shows the number of AsiSI overlaps between the r1-r4 combined INDUCE-seq dataset and BLISS and DSBCapture. Importantly, combining the replicate INDUCE-seq datasets increased the break number from ~2M to ~9.2M (and cell number from ~100,000 to ~400,000). The number of AsiSI sites identified only increased from 230 to 263. As a result, INDUCE-seq now captures a larger set of the total cleavable AsiSI sites, increasing to 95.9% (116/121) of the DSBCapture detected set and 93.9% (201/214) of the set identified by BLISS. These observations indicate that initial cell number does not have a simple linear relationship with the number of AsiSI sites detected. Therefore in the following section, this relationship is examined in detail.

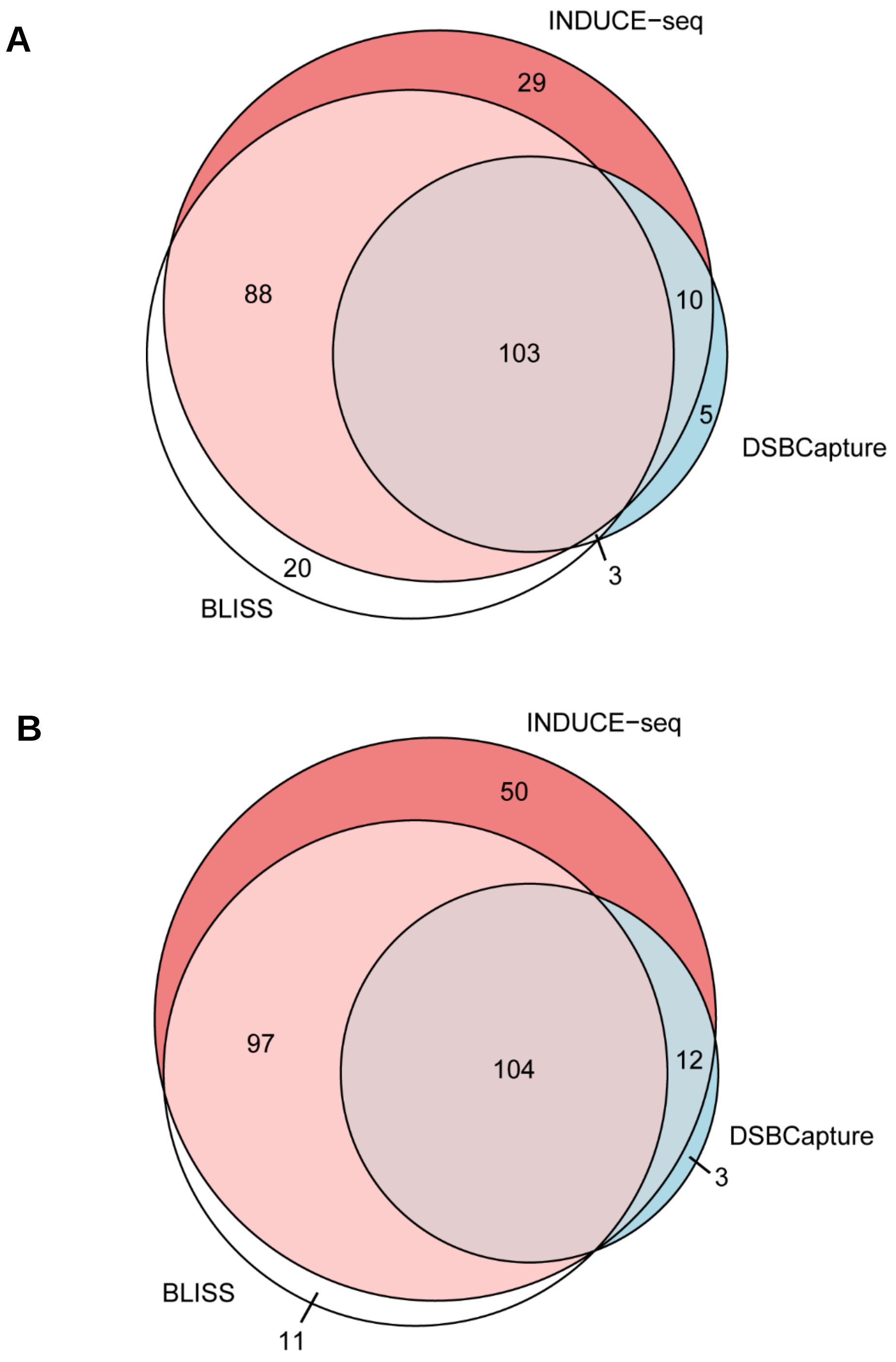


Figure 4.7. Euler diagrams demonstrating AsiSI overlaps detected by each INDUCE-seq, DSBCapture and BLISS. AsiSI sites measured from a single INDUCE-seq technical replicate (**A**, n=230), and a merged dataset comprising all four technical replicates (**B**, n=263) are compared with the sites identified by the other approaches.

4.2.4 Determining the relationship between sample size and AsiSI site break capture in DIvA cells

To investigate the effect of sampling on AsiSI site detection, the number of AsiSI sites detected was plotted against the total number of breaks. This enabled the relationship between the starting cell number and the number of AsiSI sites detected to be determined. In the previous section it was shown that the number of AsiSI sites detected did not increase linearly with increased starting cell number. Furthermore, it was noted that only 263 sites were detected of the possible ~1,200 AsiSI recognition sequences known to exist in the human genome. It has been established previously that fewer than the ~1200 possible AsiSI recognition sequences are cleaved in DIvA cells due to the methylation sensitivity of the enzyme (Iacovoni et al. 2010; Massip et al. 2010; Aymard et al. 2017).

Using the combined INDUCE-seq dataset of ~9.2M breaks, *in silico* subsetting was performed by selecting a desired number of breaks at random, followed by AsiSI site discovery. Subsets ranging several orders of magnitude from 1,000 to 5M breaks were generated, and the total break number was plotted against the number of breaks at AsiSI sites (**Figure 4.8A**), and the number of AsiSI sites detected with breaks (**Figure 4.8B**). For each plot, the separate values from the r1-r4 technical replicates are also plotted for reference (**Figure 4.8, green dots**). As expected, a linear correlation was observed across the subsets between the total number of breaks and the number of breaks identified at AsiSI sites (**Figure 4.8A**), showing a directly proportional relationship between total break number and the number of breaks identified at AsiSI sites. A linear relationship was not observed, however, between the total number of breaks and the number of AsiSI restriction sites detected. Instead, a logarithmic curve fit to the data with an R^2 of 0.9431, showing that as break number increases, new AsiSI sites are identified with diminishing returns. On both plots the original break datasets (replicates r1-r4) fit the distribution, showing that the *in silico* generated subsets accurately represent real break data. This is because breaks measured by INDUCE-seq can act as a proxy for cell number. The following section examines this relationship in further detail.

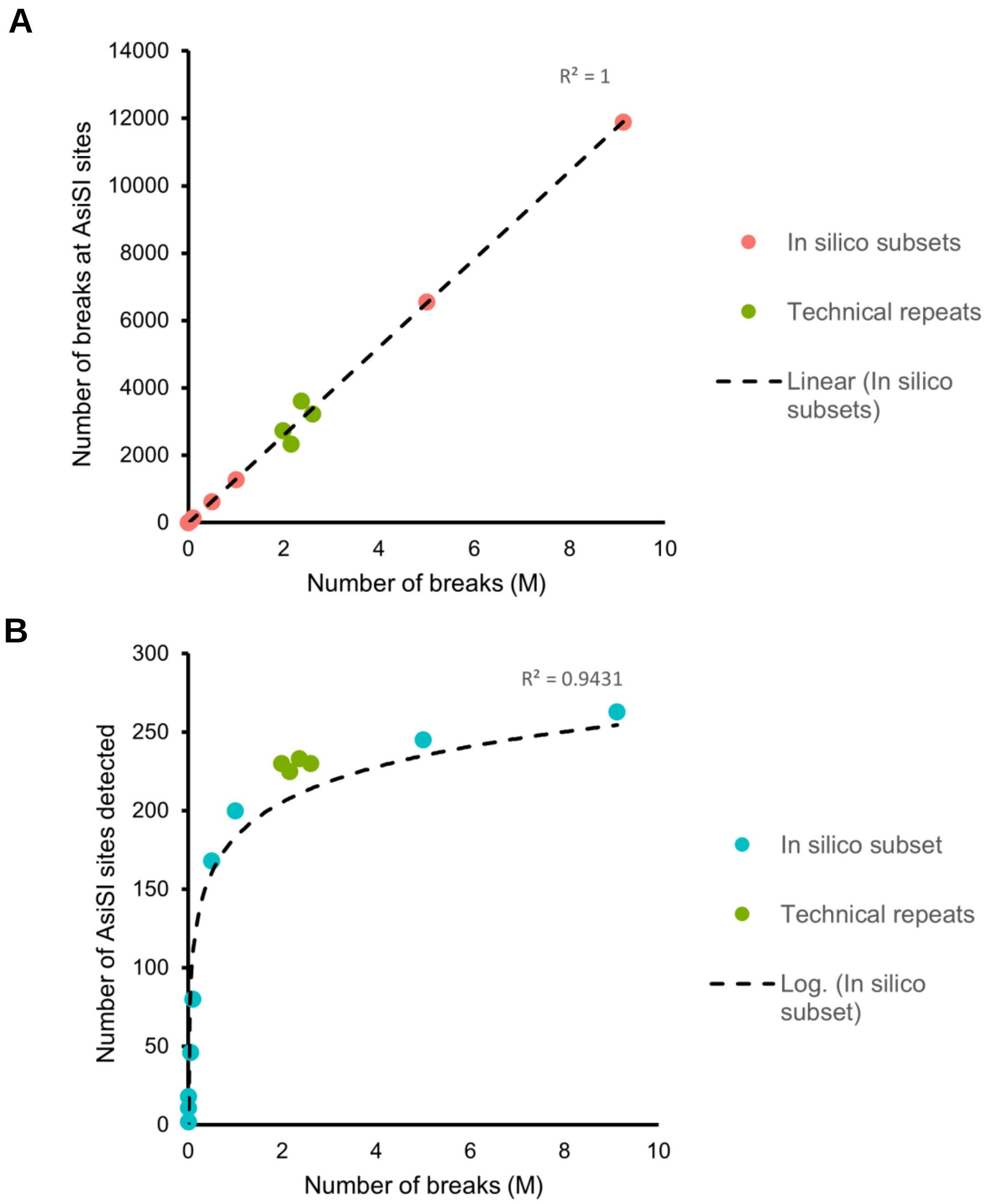


Figure 4.8. Scatterplots showing the relationship between total break number and AsiSI break capture. (**A**) The total number of breaks in a range of *in silico* subsetted 'samples' shows perfect linear correlation with the number of breaks identified at AsiSI sites (R^2 = 1). (**B**) When the total number of breaks is plotted against the number of AsiSI sits detected, a correlation is observed and fits a logarithmic curve (R^2 = 0.9431) that trends towards a plateau in the number of AsiSI sites detected.

4.2.5 Characterising the relationship between of AsiSI break number and site number

Figure 4.8 showed that with increasing total breaks detected, the total number of AsiSI breaks increases linearly, while the detection rate of novel AsiSI sites decreases. Theoretically, if the trend shown in **Figure 4.8B** is extrapolated it will reach a horizontal asymptote, where further increasing starting cell number will no longer yield any new AsiSI sites. The methylation status of AsiSI sites notwithstanding, the relationship between AsiSI site number and break number was therefore explored to better understand why increasing break (starting cell) number did not substantially increase the number of AsiSI sites detected by INDUCE-seq.

The average number of breaks found at each AsiSI site was calculated across all 4 replicates, and ranked (largest to smallest) according to break number. The resulting ranked sites were plotted against the number of breaks per AsiSI site, which shows the range of AsiSI site cleavage (**Figure 4.9A, grey bars**). On the left side of the plot, AsiSI sites that are broken consistently at high levels throughout the population are represented. The average break number found at each site decreases rapidly trending toward a tail, where many sites are found with exactly one DSB. Plotting the break data in this manner reveals that AsiSI cleavage in the DIvA cell system appears to follow a pareto-like distribution, where roughly 80% of observed outcomes come from 20% of the causes. The top quintile of ranked sites (rank 1-52) comprises almost 60% of all breaks identified. This decreases with each following quintile, with the second quintile comprising ~20% of breaks, third quintile ~10%, fourth quintile ~5% and the final comprising only ~2% of AsiSI breaks. A similar pattern can be observed when considering the frequency of each AsiSI site based on the number of breaks found at each site. **Figure 4.9B** shows the average number breaks found per AsiSI site plotted against the site frequency. The most common frequency is that of sites having a single break, which makes up over 12.5% of all sites detected. Around 60% of all sites possess an average number of breaks <10, showing that the majority of sites identified make up a small fraction of the total breaks, which mirrors that observed in **Figure 4.9A**. Similarly, few sites are found with a higher number of breaks: just 3.5% of the AsiSI sites identified possess >50 breaks. Taken together, these figures show that AsiSI site cleavage does not occur consistently between sites. The distribution is comprised simultaneously of a few sites with many breaks, and many sites with very few breaks. This potentially explains the distributions shown in **Figure 4.8**; increasing cell/break number increases the total number of breaks in a linear manner, primarily via a small number of highly recurrent positions. The same phenomenon causes new site discovery to diminish while existing positions are measured with increasing abundance. For the first time, it is now possible to calculate the precise break frequency for each individual induced break site. This is uniquely enabled by INDUCE-seq, and has significant implications for measuring genome editing in the future.

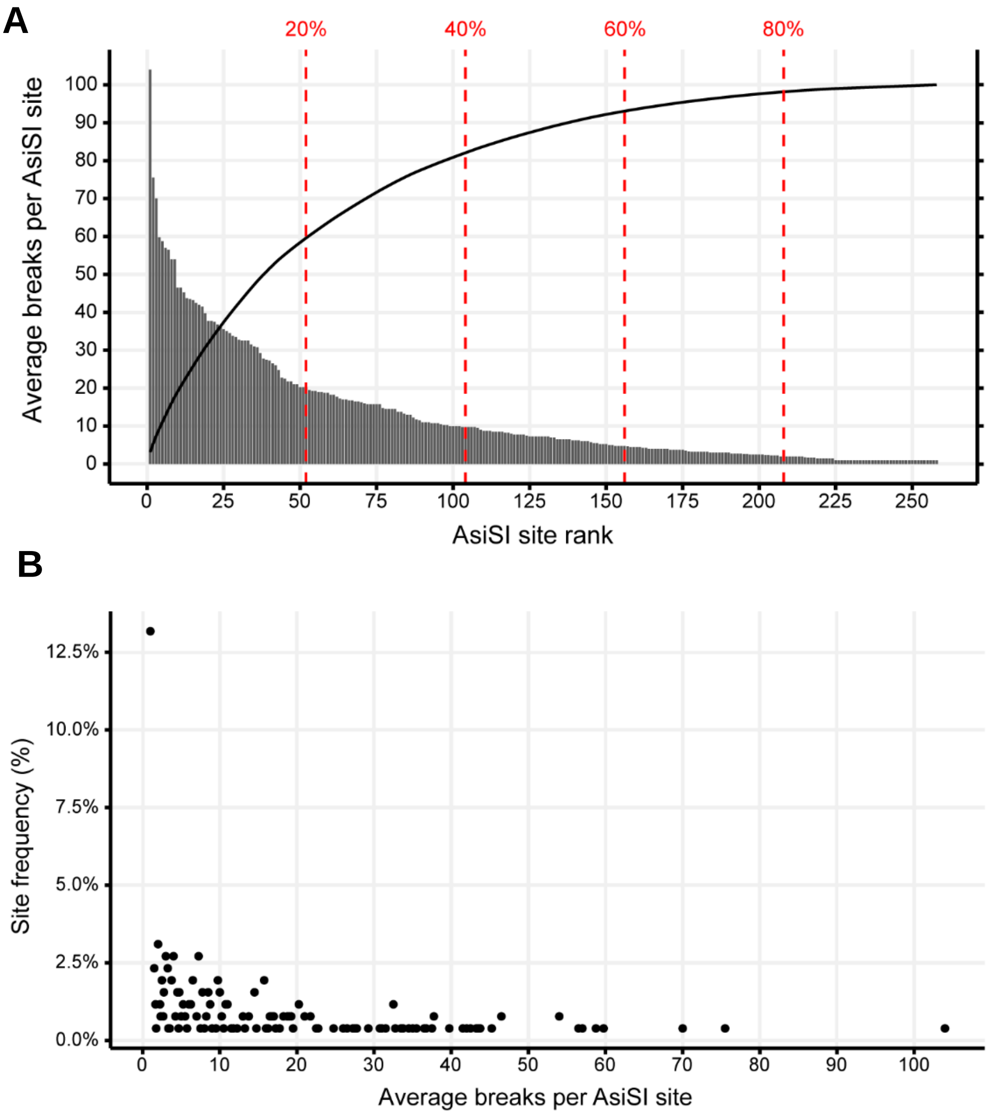


Figure 4.9. The distribution of AsiSI site cleavage. (**A**) AsiSI sites are ranked by break number, and the rank (largest to smallest) is plotted against the breaks found at each site. AsiSI sites breakage follows a pareto distribution, where the top 20% of ranked sites accounts for ~60% of the total AsiSI breaks measured. (**B**) The frequency of each site according to average number of breaks per site is plotted, showing that there are many sites with few breaks and few sites with many breaks.

4.2.6 Characterising the endogenous DSB landscape measured in a variety of different cell types

In addition to validating the measurement of live-cell, enzymatically-induced breaks using INDUCE-seq, the work in this chapter also sought to characterise endogenous break formation in different cell types. Following the development of INDUCE-seq, the technology was applied in several collaborative projects that examined break formation in different cellular contexts and conditions (**2.1.5**). Each of these experiments incorporated negative control samples that were measured by INDUCE-seq, thus producing a map of the endogenous break landscape that was present in these different cell types. These samples provided a unique opportunity to compare break data generated from several different types of cell that cover a range of cell lineages. The cell types used for this analysis include commonly used cell lines (HEK293T, N=6), (U2OS, N=3), (RPE1, N=2), induced pluripotent stem cells (iPS cells, N=3), *in vitro* differentiated neural progenitor cells (NPC, N=3) and mature neurons (MN, N=3), and lastly prostate cancer cells (C42, N=4).

First, the total number of breaks was compared between each of the different cell types (**Figure 4.10A and 4.10C**). As per previous experiments, each sample originated from roughly the same number of starting cells (~100,000) and should be considered equivalent. Strikingly, the total number of breaks measured across the various cell types differed considerably. Breaks varied from an average of ~26 million for the NPC samples to just ~300,000 for the RPE1 samples, representing a range of three orders of magnitude. Across the different cell types the relationship between read number sequenced (**Figure 4.10A and 4.10C, pink bars**) and breaks defined (**Figure 4.10A and 4.10C, blue bars**) remained constant, apart from the U2OS cells and RPE1 cells, which show only ~50% of reads resulting in breaks (**Figure 4.10C**).

Although cells were seeded in equivalent numbers for each of the samples, it was noted that for some samples, a greater number of cells were lost during the *in situ* stages of the INDUCE-seq procedure. To determine whether the number of breaks measured was impacted by differing cell number, the breaks were normalised to the DNA used for the INDUCE-seq library preparation (breaks per ng) as a proxy for cell number (**Figure 4.10B and 4.10D**). The normalised number of breaks also spanned three orders of magnitude between the cell types, demonstrating that endogenous breaks in different cell types remained highly variable and cell type specific, and could not be accounted for simply by different numbers of cells at the start of the experiment. Interestingly, the number of breaks captured in iPS cells appeared to be impacted the most following normalisation, showing 5-fold more breaks per ng than the next highest cell types (**Figure 4.10D**).

Similarly, U2OS cells and C42 cells showed nearly identical breaks per ng of DNA, despite the U2OS cell samples possessing ~2-fold more breaks on a per sample basis.

To better understand the differences in INDUCE-seq reads and break number between the cell types, total reads generated per sample was plotted against total breaks per sample for each dataset (**Figure 4.11A**). As expected, the total number of reads and total number of breaks show a near perfect linear correlation ($R^2 = 0.99854$) across three orders of magnitude. The U2OS and RPE1 cell data, as identified in **Figure 4.10C**, slightly deviate from the general trend, and are skewed towards higher reads compared to breaks. These results confirm that INDUCE-seq generates a readout where one read is equivalent to one break across multiple cell types and orders of magnitude.

Given that some samples were affected more by normalisation than others, the relationship between DNA amount and break number was plotted to determine if using DNA quantity was an appropriate strategy for normalisation (**Figure 4.11B**). Interestingly, when plotted in this manner, each cell type appears to cluster based on both sample DNA amount and on break number. In particular, the total amount of DNA extracted appears to vary greatly between different cell types, but remains constant within replicates of the same cell type. Furthermore, samples with a lower yield of DNA (iPS and C42 cells) were affected more by normalisation. These results suggest that normalisation using DNA yield is not appropriate and does not serve as an adequate proxy for cell number, as each cell type varies considerably in terms of DNA amount. Consequently, for all subsequent analysis in this chapter, no sample normalisation was applied, and cell samples are compared on a per sample basis.

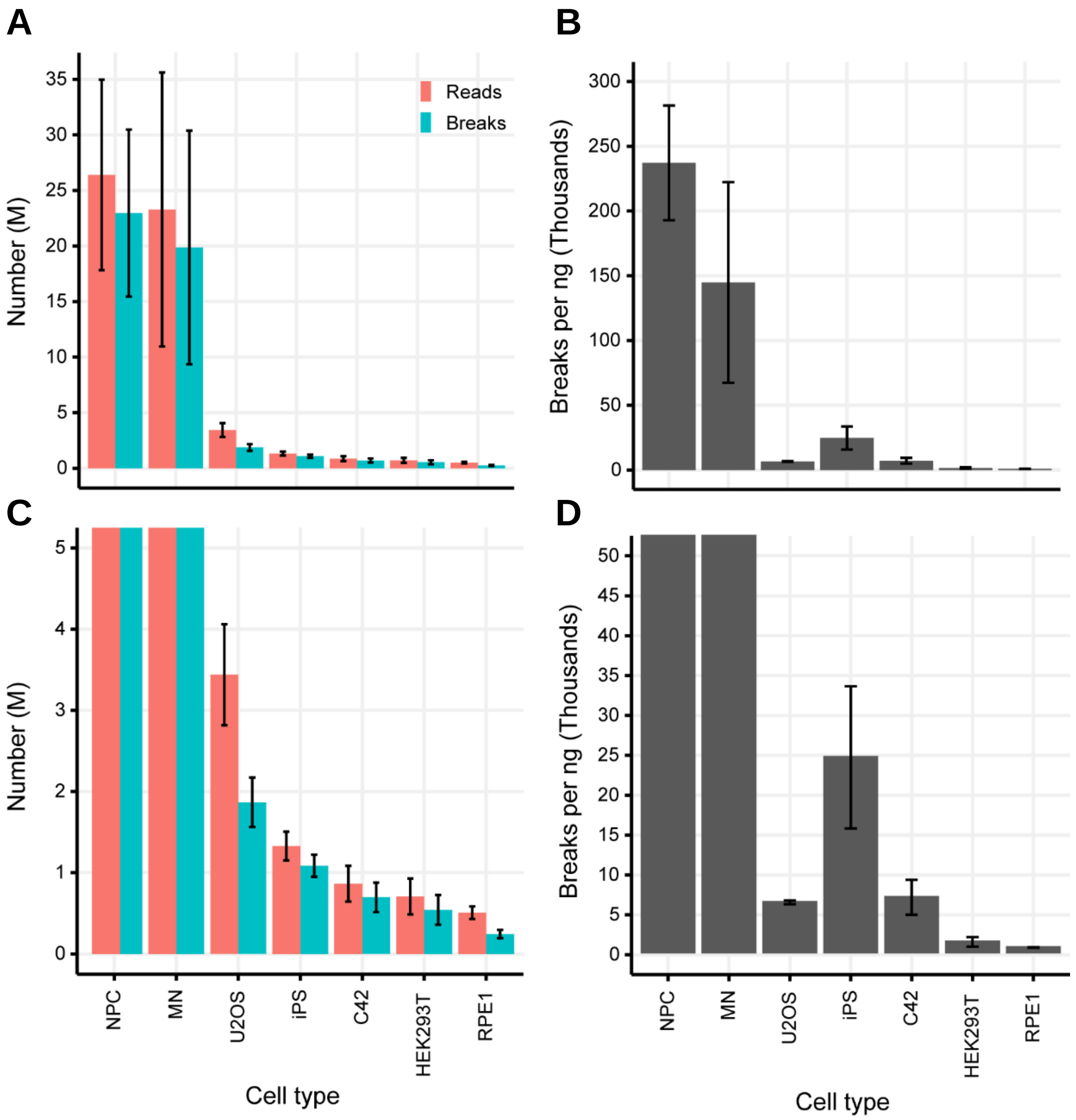


Figure 4.10. INDUCE-seq reveals that endogenous break formation is highly variable and cell-type specific. (**A**) The number of breaks identified throughout 7 different cell types spans three orders of magnitude. (**B**) The normalised number of endogenous breaks per ng of DNA used for each INDUCE-seq sample. (**C** and **D**) Y-axis zoomed-in versions of plot **A** and **B**, respectively**.** HEK293T (n=6), U2OS (n=3), RPE1 (n=2), iPS (n=3), NPC (n=3), MN (n=3), C42 (n=4), error bars as SD.

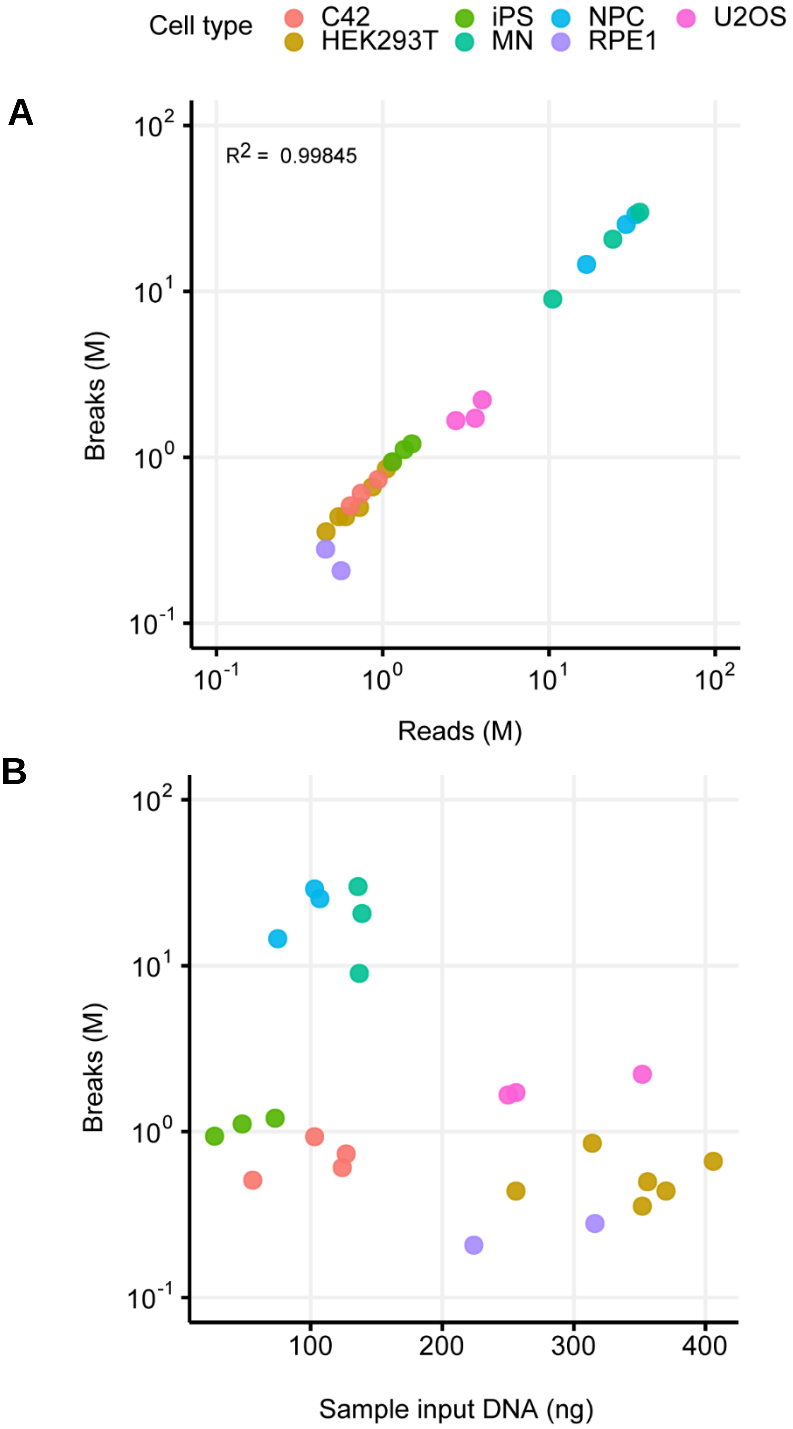


Figure 4.11. Characterising the relationship between DNA amount, INDUCE-seq reads, and breaks defined for each of the cell types. Each data point represents a separate INDUCE-seq sample, which are colour coded according to cell type. (**A**) INDUCE-seq reads strongly correlate with breaks defined across the different cell types (R^2 = 0.99845). (**B**) Replicates of each cell type cluster based on the sample DNA amount in ng and the breaks measured.

4.2.7 Examining the sites of single and recurrent endogenous breaks

Having compared the high-level differences in break number between the different cell types, the number of single and recurrent endogenous breaks was compared to better understand the distribution of break accumulation in the genome. Accurately defining recurrent endogenous breaks is challenging, however, because breaks induced by the same genomic process do not strictly occur at the same genomic position. For example, compared with the strict and clearly defined recurrent breaks that are induced at precisely the same genomic coordinate by a nuclease, transcription associated break induction might span the promoter region, first exon, or even the entire gene. To accommodate a degree of positional variation, recurrent breaks were first defined based on a proximity of 10 bp. Break sites that occurred within +/- 10 bp from another break position were defined as a recurrent break site, and the number of breaks recorded for each occurrence (**2.2.5**). If no other break is positioned within +/- 10 bp of a given break, it is defined as a single break position. **Figure 4.12A** and **4.12B** shows the average number of single break positions (pink bars) and recurrent break positions (blue bars) for each of the different cell types. Interestingly, recurrent break positions make up a small fraction of the total number of sites, as defined by a 10 bp window. For example, for NPC samples, 2.5M out of 19M (13.2%) total positions are recurrent (**Figure 4.12A and 4.12B, far-left**). Correspondingly, single breaks made up almost 90% of all break positions, confirming that despite using a 10 bp window to define recurrent sites, the majority of endogenous breaks identified were single events. Comparing between samples, the ratio of single: recurrent breaks appears relatively consistent across the cell types. This suggests that the number of recurrent positions could be a simple function of total the number of break positions. This is examined in the following section.

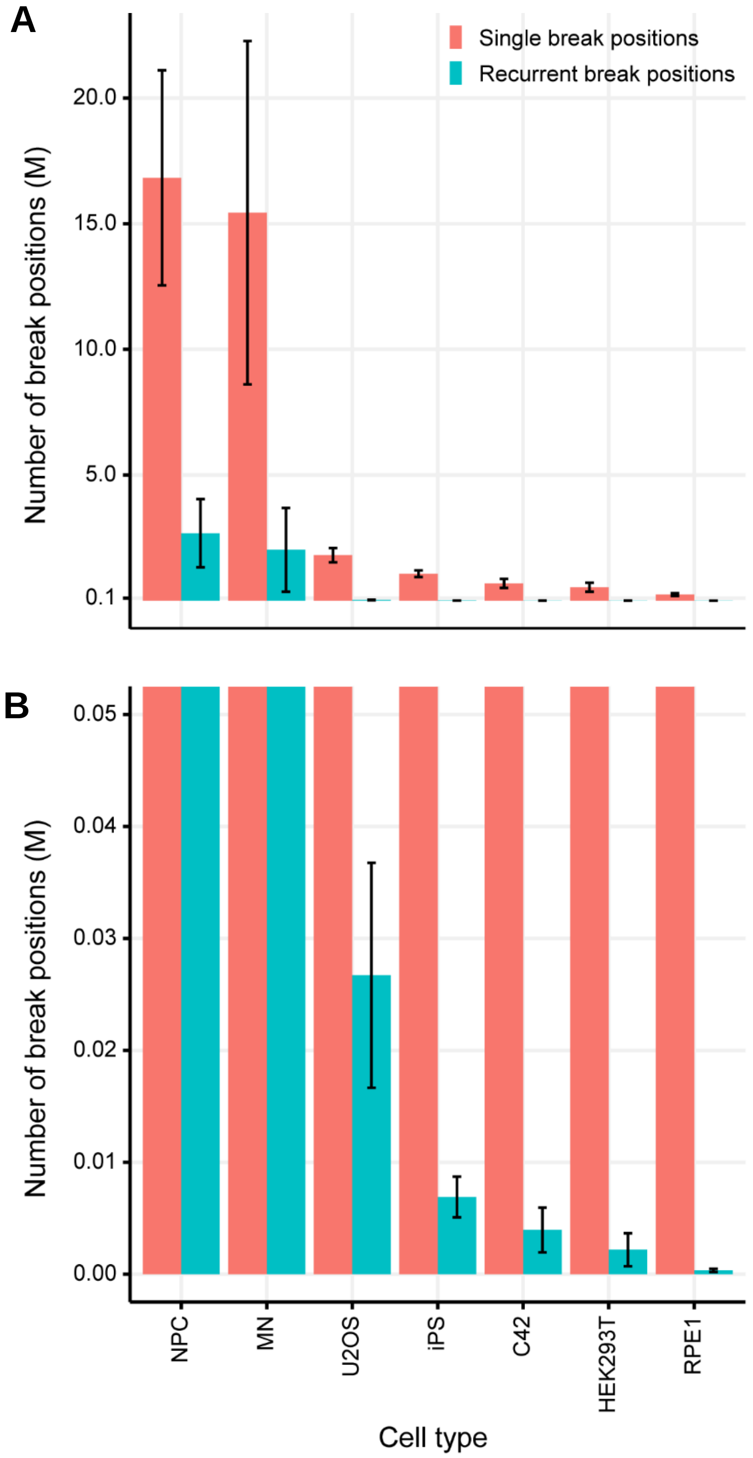


Figure 4.12. The number of single and recurrent break positions for each of the cell types when defined using a 10 bp window. (A) The ratio of single break positions and recurrent break positions appears consistent between the cell types, regardless of the total number of breaks measured. (**B**) **Y**-axis zoomed in version of plot **A.** HEK293T (n=6), U2OS (n=3), RPE1 (n=2), iPS (n=3), NPC (n=3), MN (n=3), C42 (n=4), error bars as SD.

4.2.8 Determining the relationship between total number of endogenous breaks and break recurrence

To determine whether the number of recurrent break positions was simply a function of the total number of breaks measured, the number of total breaks was plotted against the number of recurrent break sites for each sample (**Figure 4.13A**). The total break number and number of recurrent break sites shows a strong linear correlation ($R^2 = 0.96866$), confirming that as break number increases, the number of recurrent positions increases at the same rate. Next, the relationship between the recurrence of each recurrent break site (the number of breaks per recurrent site) and total number of breaks was determined. The average number of breaks per recurrent site for each sample was calculated and plotted against the total number of breaks (**Figure 4.13B**). In order to visualise the data, total breaks are represented on a logarithmic scale, and average number of breaks per site on a linear scale, which revealed a linear trend ($R^2 = 0.9061$). This demonstrates that as break number increases, the number of breaks found per recurrent break site increases at a proportional rate. Taken together, these plots show that both the number of recurrent break sites, and the number of breaks at recurrent sites, increase linearly with the total number of endogenous breaks.

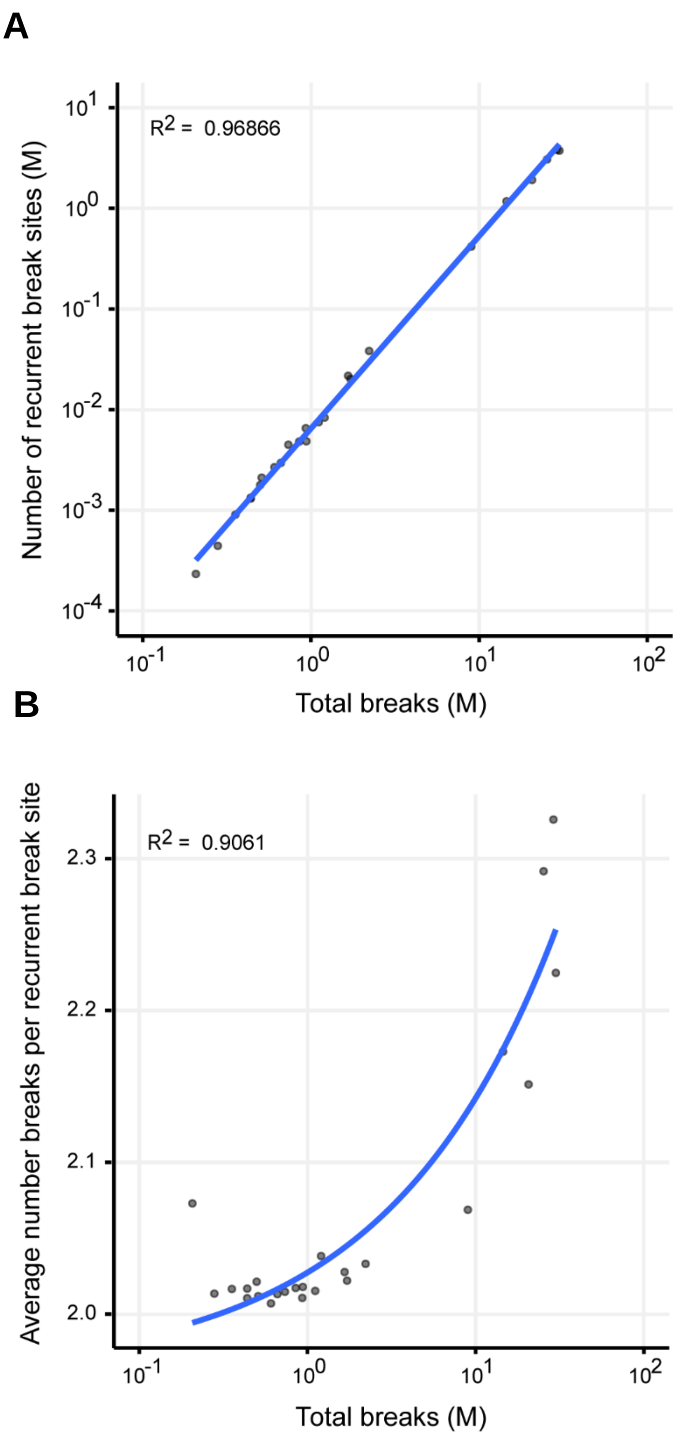


Figure 4.13. Within a 10 bp window, sample break recurrence is determined by the total number of breaks. The total number of breaks measured across the different cell samples shows strong linear correlation with the number of recurrent sites identified (**A**) (R^2 = 0.96866) and the average number of breaks per recurrent site (**B**) (R^2 = 0.9061).

4.2.9 Understanding how endogenous breaks are distributed in the genome

To better understand the nature of the distribution of endogenous breaks in the genome, the effect of break density on recurrent break site calling is examined here. This element is important to understand, because as more breaks are measured throughout the genome there is the potential that more break sites could be located closely together simply by chance, and thus become erroneously classed as recurrent break sites. If breaks are formed stochastically and are randomly positioned throughout the genome, by definition, increasing their density would increase the probability of recurrent break positions in a manner unrelated to the underlying genome biology. A series of random 'break' datasets were generated *in silico* to address this. Each of these sets correspond to an experimentally generated endogenous break dataset. The number of *in silico* generated 'breaks' were chosen to match their experimentally generated counterparts, and were created by randomly shuffling the experimental break positions throughout the genome (**2.2.5**). Furthermore, to assess the effect of break density on recurrent break site calling, recurrent break sites were defined using a range of window sizes (1, 5, 10 and 50 bp) for each of the experimental and random *in silico* break datasets.

Figure 4.14 shows the relationship between both the experimental and random break datasets between the total number of breaks and the number of recurrent break sites. As was observed previously for the recurrent breaks number defined using the experimental break data and a 10bp window (**Figure 4.13A**), a linear relationship was also observed between total break number and number of recurrent break sites using a 1, 5 and 50 bp window (**Figure 4.14, left column, pink datapoints**). Each of the corresponding randomly generated datasets also strongly correlate, showing a similar trend with high R^2 values across the window sizes (**Figure 4.14, left column, blue datapoints**). Interestingly, at the 1 bp window size the experimental and random break datapoints appear to form parallel and non-overlapping distributions, with the experimental break data demonstrating consistently higher numbers of recurrent break sites than the corresponding random datasets. As the window size is increased form 1 bp to 50 bp, this difference appears to diminish as the distributions converge at the 50 bp window size. These findings suggest that as the window size increases to define recurrent break sites, more sites are defined by chance. To quantify the difference between each pair of distributions, the log2 ratio (fold-change) was calculated between each corresponding experimental and random dataset, and plotted against the total read number (**Figure 4.14, right column**). The fold-change is greatest when using the 1 bp window size, revealing that on ~2.4-fold more recurrent break sites are found in the experimental break data than the random break data. This difference decreases as the window size increases, reducing to a ~1.3-fold increase when using a 50 bp window (**Figure 4.15A**). These graphs show that although experimental break data follows a similar distribution to randomly generated positions, considerably more recurrent break sites (1.3-fold to 2.4-fold) are identified for an equivalent number of genomic positions, depending on the window size used for recurrent site identification. These findings show that experimentally defined endogenous breaks are not positioned randomly throughout the genome.

Although a consistently high linear correlation was observed across all window sizes for both the experimental and random datasets ($R^2 > 0.9$), a second trend was also observed across the R^2 values depending on the window size that was used to define recurrent break positions. As the recurrent break defining window size increases from 1 bp to 50 bp, the correlation between the total break number and the number of recurrent sites becomes stronger for both the experimental and random datasets. To determine the relationship between the correlation (R^2 value) of total break number and break recurrence, and the window size used to define the recurrent breaks, the R^2 value from each of the trendlines shown in **Figure 4.14** was plotted against each of the different window sizes (**Figure 4.15B**). Strikingly, the correlation increases with increasing window size. However, the correlation for the experimental and random break datasets increases with a different rate. As expected, the correlation between the total number of random breaks and recurrent sites increases linearly with the increasing window size (**Figure 4.15B, blue line**). The experimental break data, however, shows a non-linear distribution (**Figure 4.15B, pink line**). At small window sizes (1, 2 and 5 bp) the correlation between total break number and number of recurrent sites is weaker than is expected by chance. In contrast, at larger window sizes (10, 20 and 50 bp) the correlation is stronger than is expected by chance. Collectively, these observations confirm that using a tuneable window size it is possible to distinguish recurrent break sites above the rate that would be discovered by chance. The ability to do this has implications for the characterisation of different classes of recurrent break sites, and the selection of the appropriate window size should be considered depending on the different on the application.

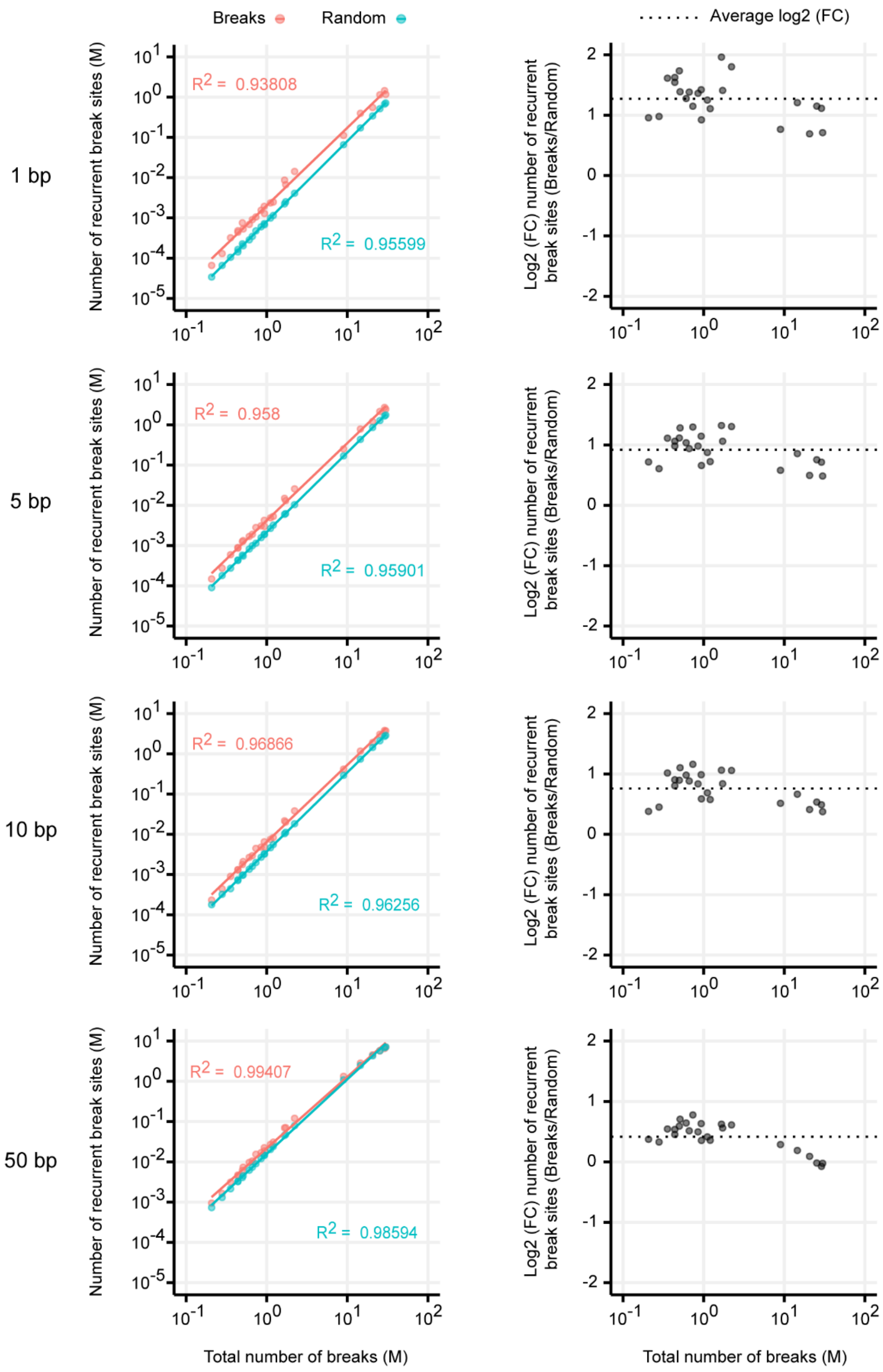


Figure 4.14. Comparing the relationship between break number and recurrent site number for real and randomised break datasets. Each row corresponds to recurrent break values when defined using 1, 5, 10 and 50 bp windows. **Left panel**: Scatterplots showing correlation between total breaks and number of recurrent sites for the the real (pink) and randomised (blue) break datasets. **Right panel**: The log2 fold-change between number of recurrent sites identified using the real breaks and random data.

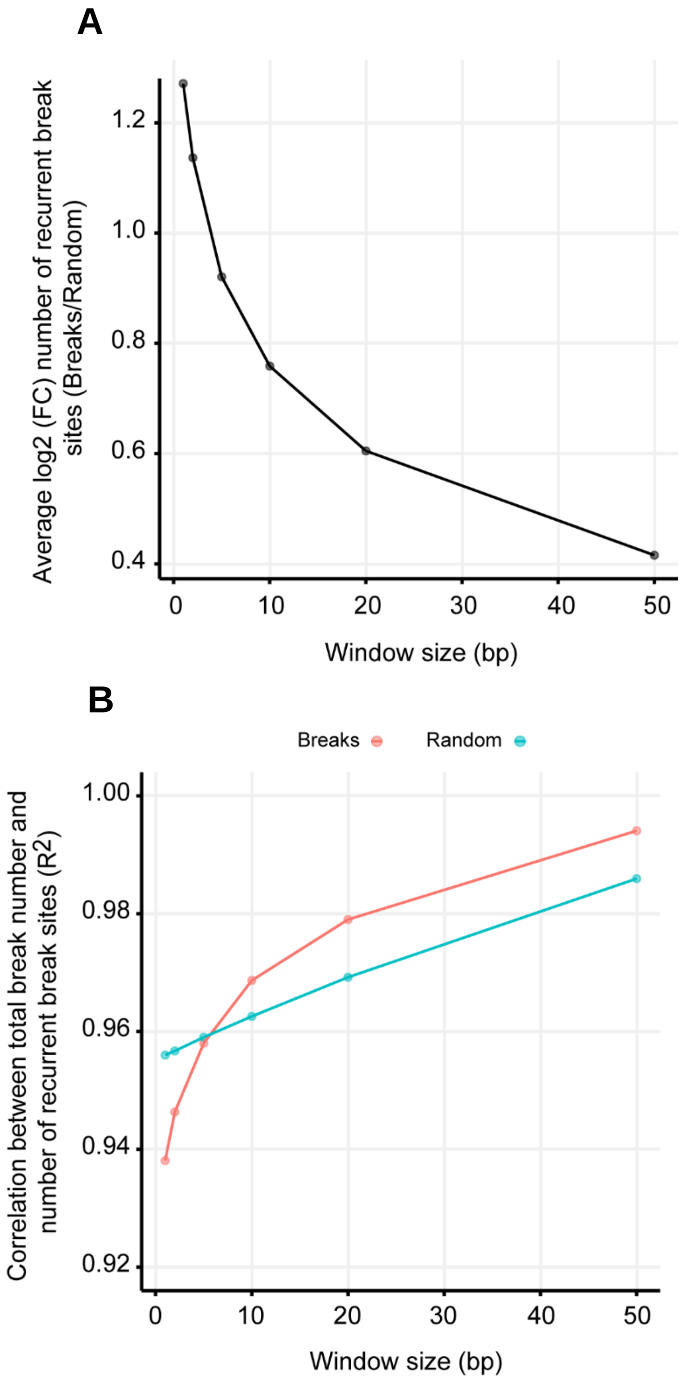


Figure 4.15. Quantifying the effect of the window size used to define recurrent breaks (**A**) The log2 fold-change between the number of recurrent defined in each real and random dataset decreases with increasing window size in bp**.** (**B**) Recurrent break window size impacts how well total break number and number of recurrent break sites correlate. Window sizes of 1, 2, 5, 10, 20 and 50 bp were applied.

To further characterise the differences between experimental and random break datasets, the average number of breaks per recurrent site was calculated and plotted against the total number of breaks for each window size, as shown previously for **Figure 4.13B**. The average number of breaks per recurrent break site for the experimental and random break datasets diverge as the window size increases from 1 to 50 bp (**Figure 4.16**). This contrasts with the distributions shown in **Figure 4.14**, which converged with increasing window size. Additionally, the average number of breaks per recurrent break site increases more rapidly in experimental samples with increasing window size, than the equivalent random datasets. The difference is more pronounced for datasets of >1M breaks and cannot be used reliably to identify biological recurrency in datasets with fewer features. Indeed, this difference exists primarily in large datasets, showing the greatest difference for datasets with >1M positions, suggesting that as more breaks are measured, they position closer together than can be expected from an equivalent number of random positions.

Important differences between the experimental and random break datasets can also be seen at the lower window sizes. Using a 1 bp window to define recurrent breaks reveals no linear correlation between total break number and average breaks per recurrent site. This is in stark contrast to that seen for the random datasets, which show a strong linear correlation, regardless of the size of the recurrent break defining window. To better understand how window size effects the correlation between total break number and average number of recurrent breaks per site, the R^2 value from each of the trendlines in **Figure 4.16** was plotted against the window size used to define recurrent breaks (**Figure 4.17**). As shown by the distributions in **Figure 4.16**, for the random break datasets (**Figure 4.17, blue line**), as expected, there is a consistently high correlation between the total break number and the average number of breaks per recurrent break site across all window sizes examined. This shows that when positions are selected at random, the total number of positions determines the average break recurrence regardless of window size. Experimental break data, on the other hand, does not share this relationship. Below a threshold of 20bp, the recurrence window fails to capture break recurrence and does not correlate with total break number. This further demonstrates that experimentally derived endogenous breaks are not formed randomly throughout the genome.

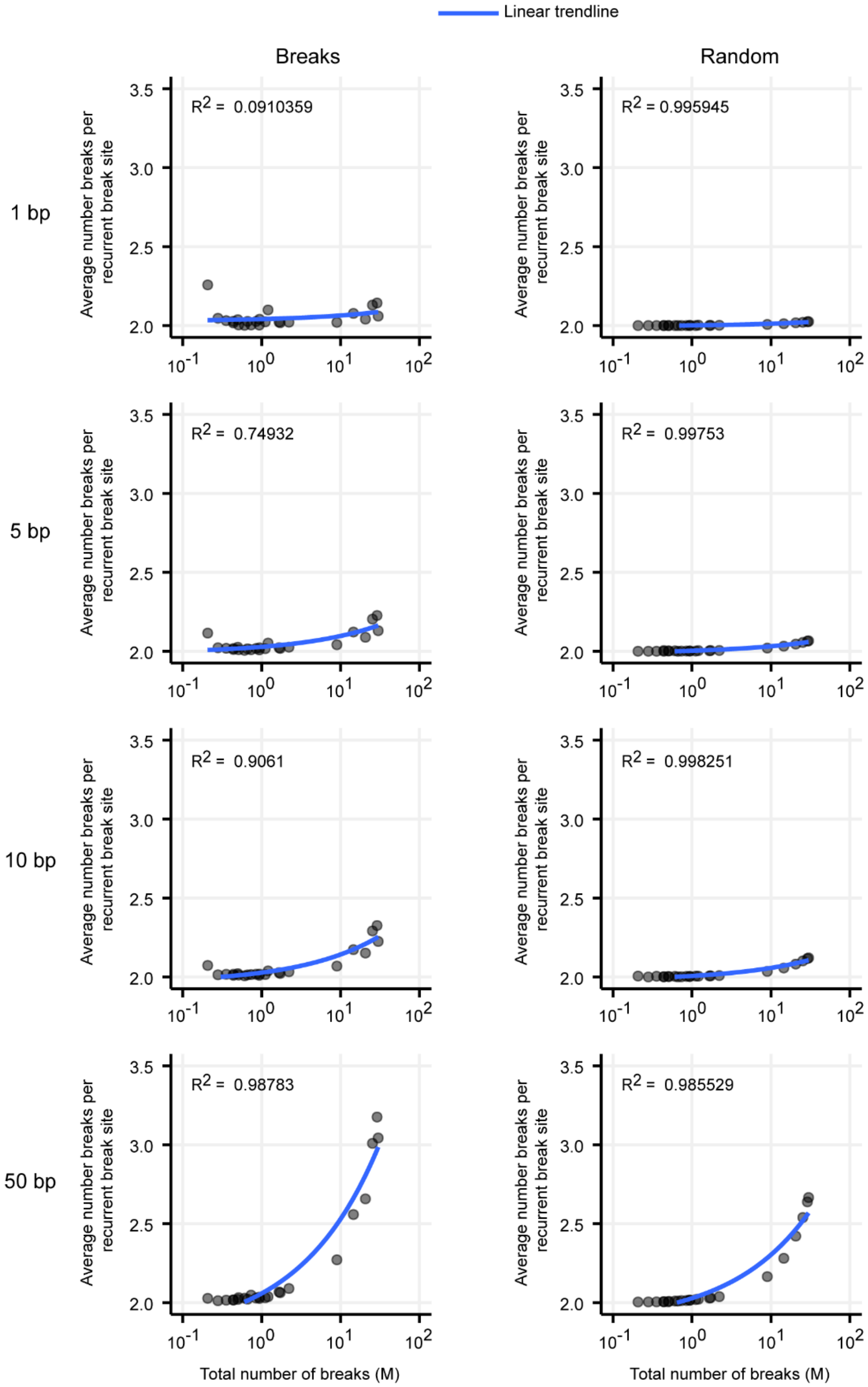


Figure 4.16. Comparing the relationship between break number and the average number of breaks per recurrent site real (left panel) and randomised (right panel) break datasets. Each row corresponds to recurrent break values when defined using 1, 5, 10 and 50 bp windows.

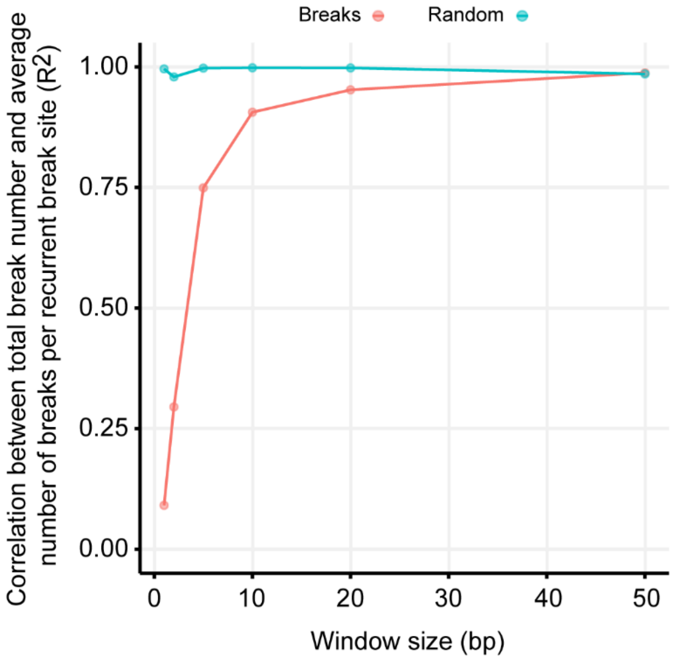


Figure 4.17. The window size used to define recurrent break sites alters the correlation between break number and the average number of breaks per recurrent break site at each site. Random datasets show uniformly high correlation regardless of the window size used (Blue line). For real break data shows no correlation between break number and average breaks per recurrent site when a small window size is used (1-5 bp), but high correlation when a larger window size is used (10-50 bp) (Pink line). Window sizes of 1, 2, 5, 10, 20 and 50 bp were applied.

The data presented in this section reveals several important characteristics of the endogenous breaks. Firstly, recurrent break sites are identified more frequently using the experimentally defined endogenous break datasets than the random break datasets, across all window sizes used. This observation holds despite the fact that both classes show strong linear correlation between the total number of breaks and the number of recurrent break sites. Secondly, for the experimental datasets, the total number of breaks does not correlate with the average break number at recurrent break sites when using a small window size (<5 bp) to define recurrent break positions. This is in contrast to the random datasets, which show strong correlation regardless of the window size used. Furthermore, the difference in the average break number at each recurrent site increases between the experimental and the random datasets as the window size increases. This shows that despite the potential for defining more recurrent positions by chance, the number of breaks at recurrent positions in the experimental data is higher and asymmetric compared to the equivalent number of random events. Collectively, these findings show that although the number of recurrent breaks increases linearly with increasing total break number, unlike random positions, endogenous break accumulation is not distributed evenly throughout the genome. This insight can now be substantiated quantitatively with this data, showing that rare and seemingly random break events in a population of cells are not random due to the underlying genomic structure.

4.3 Discussion

In this chapter, INDUCE-seq's ability to detect induced breaks in live cells was investigated using the AsiSI inducible DIvA cell system. In each of the replicates tested INDUCE-seq reproducibly detected breaks at ~230 of the ~1,200 AsiSI recognition sequences that exist in the human genome. The reproducibility of break detection between replicates was further confirmed by the closely correlated distributions of breaks measured at each individual AsiSI site. This revealed that INDUCE-seq measured induced breaks in the genome in accurate and undistorted proportions regardless of the total number of breaks measured. Furthermore, compared with the HindIII experiment, where 96.7% of breaks were positioned at HindIII restriction sites, only 0.11-0.16% of breaks measured in each DIvA sample were located at AsiSI sites. This demonstrates that in the live DIvA cell system, low-level endogenous breaks make up the majority of all breaks measured despite four hours of break induction via a restriction endonuclease. This is a particularly important observation considering the future application of INDUCE-seq in experiments involving CRISPR-Cas9, which typically has a single on-target site and is not likely to cleave with as great an efficiency.

When compared with the alternative methods BLISS and DSBCapture, INDUCE-seq demonstrated enhanced efficiency and cost-effectiveness, requiring 23-40 fold fewer sequencing reads to detect the highest number of AsiSI restriction sites. INDUCE-seq utilises the same break labelling procedure as the BLISS method, meaning that any differences in AsiSI site detection occur as a direct result of the different library preparation and sequencing strategies that each of the methods employ. The improved sensitivity of detection possible with INDUCE-seq can be attributed directly to the lack of PCR amplification used during sample preparation. It is often thought that to detect low-frequency events the signal must be amplified sufficiently for accurate detection and to mitigate the risk of signal loss (break events in this case). These experiments demonstrate that the PCR-free library preparation and flow cell enrichment strategy employed by INDUCE-seq is sufficiently specific and powerful that the lack of amplification improves sensitivity and outweighs the negative side effect of signal loss during sample preparation.

Interestingly, despite INDUCE-seq's enhanced sensitivity, a maximum of 263 out of the 1,211 total AsiSI restriction sites could be identified, suggesting that not all AsiSI sites in the genome can be targeted by the enzyme. Indeed, it has been shown previously that ƴH2AX is not present at all AsiSI sites following induction, possibly because AsiSI cleavage is blocked by CpG methylation and the structure of surrounding chromatin (Iacovoni et al. 2010). This feature of the DIvA cell system may explain the distributions observed in **Figure 4.8**, where total break number at AsiSI sites increases linearly with total breaks (used as a proxy here for cell number), but new AsiSI sites are identified with diminishing returns. Further examining the ranked distribution of breaks at each AsiSI site revealed a pareto distribution, where the top 20% of sites accounted for ~60% of all AsiSI breaks. Simultaneously, a long tail of low frequency AsiSI sites was also observed, showing many AsiSI sites with just a single break. This shows that even at cleavable AsiSI sites, not all are cleaved with equal efficiency in a population of cells. Certain sites show much greater propensity for cleavage, which suggests differences in genomic context that positively or negatively regulate DSB formation and repair. Further work using the DIvA cell data should focus on annotating these positions to better understand which genomic factors most greatly impact nuclease-induced DSBs in the genome. Importantly, these features could then be taken into consideration during the design of CRISPR guide RNAs, as they have the potential to enhance off-target activity or supress on-target activity, both of which should be avoided.

In addition to demonstrating INDUCE-seq's ability to measure induced breaks in live cells, the endogenous break distributions from a variety of different cell types were also compared in this chapter. Part of the experimental design presented cell types of different genomic stability ranging from cells that are generally regarded as genetically stable in the case of RPE1 (Bodnar et al. 1998; Scott et al. 2020), to cells with high levels of genetic plasticity in case of the neural progenitor cells (Alt and Schwer 2018; Tully 2020). In addition to accurate profiling of DSBs genome-wide, the PCR-free approach taken by INDUCE-seq enables the quantitative measurement of total break number from a sample of cells. Strikingly, when simply comparing the total number of breaks in each of the cell types, a range spanning three orders of magnitude was observed, from ~26 million breaks in a sample of neural progenitor cells to just, ~300,000 in a sample of RPE1 cells. These experiments demonstrate cell-type specific endogenous break characteristics that are in line with certain known aspects of their genome stability, but that have not previously been observed using a genome-wide break detection method. Indeed, alternative approaches lose much of this quantitative information following break sequence amplification.

The number of recurrent breaks was also defined to better characterize the break distribution for each cell type, which is important to understand the factors that contribute to the stability of the genome. Recurrent breaks were defined using an approach which uses the proximity (a window size set in bp) of each break to the next, rather than defining genomic bins and counting breaks per bin. This strategy allows for variable width around recurrent break sites, and is thus better represents of how breaks would cluster in regions of the genome that are more prone to frequent breakage. Starting with a 10 bp window, the number of recurrent break sites (defined as a position of ≥2 breaks) was calculated for each cell type, revealing that a similar proportion of recurrent breaks was present for each sample. Furthermore, at least when defining recurrent breaks using a 10 bp overlap window, the total number of breaks correlates strongly with recurrent site number and the average number of breaks identified at each recurrent site. Simply put, when more breaks are observed in a sample, more recurrent breaks are defined.

To better understand if the recurrent breaks that were defined using the was simply a reflection of the stochastic positioning of an increasing number of breaks, an equivalent number of random positions were generated, and the number of recurrent breaks was compared between experimental and random break datasets, across window sizes ranging from 1 – 50 bp. This analysis revealed that it is more likely that a proximal break occurs in another cell than would be expected by chance alone. This confirms that break positioning is not entirely stochastic and is therefore being influenced by some aspect of the underlying biology of the genome. Furthermore, the second part of the analysis compared the average break number found at each recurrent break site with the total number of breaks. This revealed that when using a small window size, no correlation existed between the two factors. This is in stark contrast to a very strong linear correlation shown by the random datasets, further confirming the non-random nature of the break positioning.

In summary, these findings suggest that different sub-types of recurrent break sites may exist within the experimental break data. Some recurrent breaks appear to be highly recurrent and precisely positioned to within a single bp, whereas others only appear when a larger window size is applied, which represent regions that are generally enriched for DSBs, but are not precisely overlapping. This phenomenon is probably due to differences in genome biology that give rise to alternate recurrent break types. For example, Type II DNA topoisomerase enzymes (TOP2) or retroviral DNA insertion events may lead to more precisely positioned recurrent breaks. In contrast, transcription associated DSBs may occur with variable positioning immediately downstream of the promoter region. There is considerable scope to characterise the various types of recurrent break that occur endogenously in the genome, and in future work should focus on characterizing the break distribution across a greater range of cell types and lineages. This should include defining the factors that give rise to the cell-type specific differences in DSB number and genomic distribution. Similarly, directed research into specific DSB repair mutants or chemical break induction will help unravel the spatio-temporal induction of endogenous breaks and reveal the underlying mechanisms that result in the different recurrent breaks described here.

Finally, the results in this chapter confirm that INDUCE-seq is capable of accurately measuring both restriction endonuclease induced breaks and endogenous breaks in a variety of cell types. In the following chapter, INDUCE-seq is applied to discovery of CRISPR-Cas9 off-targets in the genome.

5 Chapter V - The discovery and characterisation of CRISPR-Cas9 off-target editing using INDUCE-seq

5.1 Introduction

Targeted genome editing nucleases such as CRISPR-Cas9 now offer remarkable promise to transform human health via the direct modification of DNA. Due to their mechanism of action however, they have the potential to generate substantial DNA damage in the form of DSBs at locations in the genome other than those targeted for editing (Tsai et al. 2015; Zhang et al. 2015). These so-called off-target events have been observed for all classes of nuclease-based genome editors, including meganucleases, ZFNs, TALENS, and CRISPR-Cas, and have the potential to introduce serious mutagenic consequences for the cell or organism being edited (Cheng and Tsai 2018). CRISPR-Cas9, which employs an RNA-guided targeting mechanism, is particularly prone to off-target formation as RNA-DNA interactions are susceptible to mispairing with sites of high sequence similarity. Understanding the determinants of these off-targets is of fundamental importance, and is a major hurdle that must be overcome for the translation of CRISPR-Cas9 into the clinic as a therapeutic modality (Tsai and Joung 2016).

5.1.1 The causes of CRISPR-Cas9 off-target editing

Considerable progress has now been made to understand the rules that determine the specificity of CRISPR-Cas9, which has revealed a range of factors that promote off-target editing (**Figure 5.1**). The first and most obvious determinant of off-target editing is the number, position, and distribution of mismatches at genomic sites with a similar sequence to the desired target (**Figure 5.1, bottom**). Numerous studies have shown that CRISPR-Cas9 is able to cleave with high efficiency, sometimes higher than at the on-target site, at sites with as many as six mismatches in the protospacer sequence in experiments conducted in live cells (Tsai et al. 2015), and up to ten mismatches in *in vitro* experiments (Pattanayak et al. 2013; Newton et al. 2019). Furthermore, off-target frequency is greatly impacted by the position of mismatches in the protospacer region. More than two mismatches in the seed region of the protospacer, as defined as the first 8-12 nucleotides immediately upstream of the PAM, has been shown to abolish Cas9 activity *in vitro,* whereas 5' PAM-distal mismatches are often tolerated (Cong et al. 2013). Finally, a range of non-canonical PAM sequences (NGG) have been observed at Cas9 off-targets, including NAG, NGA, NAA, NGT, NGC and NCG, demonstrating that the DNA-protein interface of Cas9 is also susceptible to erroneous binding (Hsu et al. 2013; Tsai et al. 2015).

In addition to target sequence, a range of other factors are known to impact off-target editing (**Figure 5.1, top**). Various modifications to the guide RNA have been shown to improve target specificity, including truncated guide RNAs (Fu et al. 2014), longer guide RNAs, which incorporate a 5’ hairpin structure (Kocak et al. 2019), and chemically modified guide RNAs that incorporate locked or bridged nucleic acid bases (Cromwell et al. 2018). Similarly, a range of modified high-fidelity CRISPR-Cas9 enzymes are now available that have been engineered for increased specificity. Genomic factors are also known to contribute to CRISPR off-target formation; heterochromatin has been shown to obstruct Cas9 cleavage and repair at target sites, thus indicating that open chromatin is more likely to harbour off-targets (**Figure 5.1, middle**) (Fujita et al. 2016). Other than simply being at more accessible regions of the genome, the process of active transcription can directly stimulate DNA cleavage by influencing Cas9 release in a strand-specific manner. Cytosine-5 methylation at CpG dinucleotides does not directly block Cas9 binding *in vitro*, but does show marginal inhibition in a cellular context (Verkuijl and Rots 2019).

Finally, delivery of Cas9 into cells, which directly impacts Cas9/sgRNA abundance, can significantly alter off-target editing. Numerous studies have now shown that directly delivering CRISPR-Cas9 in the form of ribonucleoprotein complexes (RNPs) results in much lower levels of off-target editing than plasmid-based, viral or mRNA delivery approaches, while maintaining high levels of on-target editing (Kim et al. 2014).

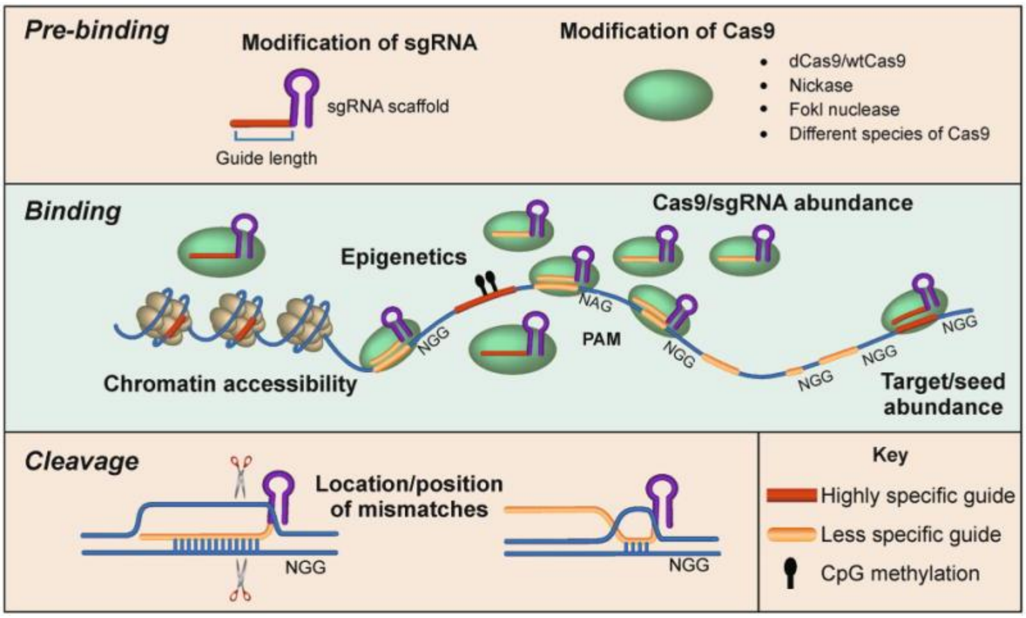


Figure 5.1. Factors that impact Cas9 specificity. Prior to binding, Cas9 and sgRNA architecture can be modified to alter specificity. At the DNA level, chromatin context, epigenetic factors, and Cas9/sgRNA abundance, alter specificity at a given locus. Once bound to DNA, the degree of sequence complementarity to the sgRNA spacer defines whether the interaction is transient, or cleavage is initiated. Taken from Wu et al. 2014.

5.1.2 Current approaches for the prediction and measurement of CRISPR-Cas9 off-targets in the genome

Given the range of factors that influence off-target site selection, binding and cleavage, and the potential hazards associated with the introduction of unintended genomic DSBs, it has become paramount to assess the off-target profile for any therapeutic CRISPR guide RNA. Recently, a variety of NGS-based methods have been developed for this purpose, and these are outlined in section **1.7**. In addition to empirically determining off-targets using NGS, the use of *in silico* tools for predicting CRISPR-induced off-targets has become one solution for assessing the safety of CRISPR genome editing. These tools have the potential to remove the need for performing the experimental procedures described previously. *In silico* methods vary substantially in approach; from using a simple sequence based algorithm (e.g. Cas-Offinder), to using machine learning and/or deep learning to predict off-targets based on criteria defined from an array of published off-target and epigenetic datasets (CRISTA, Elevation, deepCRISPR) (Bae et al. 2014; Abadi et al. 2017; Chuai et al. 2018; Listgarten et al. 2018).

Despite these advances, the accurate prediction of CRISPR off-targets is extremely difficult given the scale of the genome search space, which predicts very large numbers of off-targets, while the number of true off-targets determined experimentally remains very low. This discrepancy in off-target prediction and detection marks a significant problem for the field, highlighting the need to better define the criteria that determine where and how the Cas9/sgRNA RNP interacts within the genome. Furthermore, artificial intelligence-based (AI) approaches are limited by the availability of real-world training and validation datasets that are accurate and at a sufficiently large scale (Cheng and Tsai 2018). As discussed previously (**3.1.1**), many of the cell-based approaches such as GUIDE-seq, which are typically used for training AI-based *in silico* prediction tools, suffer from significant biases because of the PCR-based DSB-enrichment and library preparation techniques that they employ. Training AI-based prediction tools using such biased data, therefore has the potential to lead to significant inaccuracies in off-target prediction. Until these methods have been determined to be reproducible and accurate, experimental tools should be prioritised for the safety profiling of CRISPR genome editing.

5.1.3 Chapter aims

In the previous chapter, INDUCE-seq was used to capture nuclease induced-breaks in live DIvA cells, thus demonstrating its applicability for measuring CRISPR induced on- and off-target breaks. The work in this chapter aims to demonstrate the use of INDUCE-seq for the discovery of CRISPR off-targets in the genome using a well-studied guide

RNA (EMX1), which has been tested by a variety of other methods (Frock et al. 2015; Kim et al. 2015; Tsai et al. 2015; Tsai et al. 2017; Yan et al. 2017; Wienert et al. 2019). Because INDUCE-seq measures a snapshot of DSBs genome-wide at a given moment in time, we will measure DSBs over a time course following treatment. In doing this, we will evaluate the ability of INDUCE-seq to discover known and novel off-targets induced over time, which in turn will enable the determination of the kinetics of on- and off-target DSB formation and repair by the guide RNA EMX1. Furthermore, the development of a novel analysis pipeline for the discovery of CRISPR off-targets derived from INDUCE-seq datasets is also described in this chapter.

5.2 Results

5.2.1 Validation of EMX1 on-target editing

To capture the genome-wide off-targets induced by the guide RNA EMX1 over time, HEK293 cells were treated with Cas9/EMX1RNP and sampled at 0, 7, 12, 24 and 30 hours post nucleofection as described in section **2.1.1**. To account for technical differences during nucleofection, a full time course experiment was performed in duplicate, referred to as r1 and r2. Prior to running the INDUCE-seq experiment, editing at the EMX1 target site was quantified to confirm editing activity and successful cell nucleofection. CRISPR editing typically forms indels (insertions and deletions), hence mutation induction was measured across regions spanning the EMX1 target site from the final 30h treated and control samples. To this end, single-target PCR amplicons that span the EMX1 target site was analysed by Sanger sequencing.

Using the Synthego Inference of CRISPR edits (ICE) analysis tool (Hsiau et al. 2019), Sanger sequencing traces were aligned to the EMX1 target spacer sequence and PAM (**Appendix D, Figure A1**) and were compared for the treated and control sample. This analysis revealed that the edited fraction for both 30h samples was greater than the unedited fraction, with r1 showing a 68% indel frequency and r2 59%, confirming the presence of indels and CRISPR editing (**Appendix D, Figure A2**).

5.2.2 Measuring genomic DSBs in CRISPR-edited cells

Having confirmed highly efficient and reproducible EMX1 target editing for both r1 and r2 repeats, INDUCE-seq was performed on samples grown to 0, 7, 12, 24 and 30 hours following CRISPR treatment to measure genome-wide DSB induction over time. As expected, the relationship between the total number of reads sequenced and the number breaks defined was consistent for each of the treated and control samples (**Figure 5.2A**). This showed a near perfect linear corelation (**Figure 5.2B, R^2 = 0.99779**). Similarly, as observed for DIvA cells in the previous chapter, the number of reads and breaks measured between the treated and control samples at each time point was relatively consistent. This suggests that CRISPR treatment at this scale did not induce large-scale increases in total break number. In general, fewer breaks were measured in the treated samples, which averaged 3.15 million, compared to 3.90 million breaks measured for the control samples (**Figure 5.2C**). No significant difference in break number was observed between the treated and the control group (P = 0.2113, Mann Whitney test). It is noteworthy that a single outlier was identified as one of the 12h control samples. Taking account of the outlier reduced the average number of breaks measured in the control to

3.29 million, indicating that it accounted for the increase in average break number across the control set.

Interestingly, across the time course most breaks were observed at 7h and 12h following nucleofection (**Figure 5.2A**). For the CRISPR treated samples, 3.4-fold more breaks were identified at 7h than 0h, and 2.9-fold more breaks were identified at 12h than 0h. This trend appeared to diminish for the 24h and 30h samples which show a 1.1 and 1.5-fold increase, respectively. A similar pattern was observed for the control samples, suggesting that the time point of sampling following nucleofection has a greater effect on total break number than the CRISPR treatment.

Taken together these observations give a high-level impression of break detection of CRISPR treated cells using INDUCE-seq that can now be used for further downstream secondary analysis and off-target discovery.

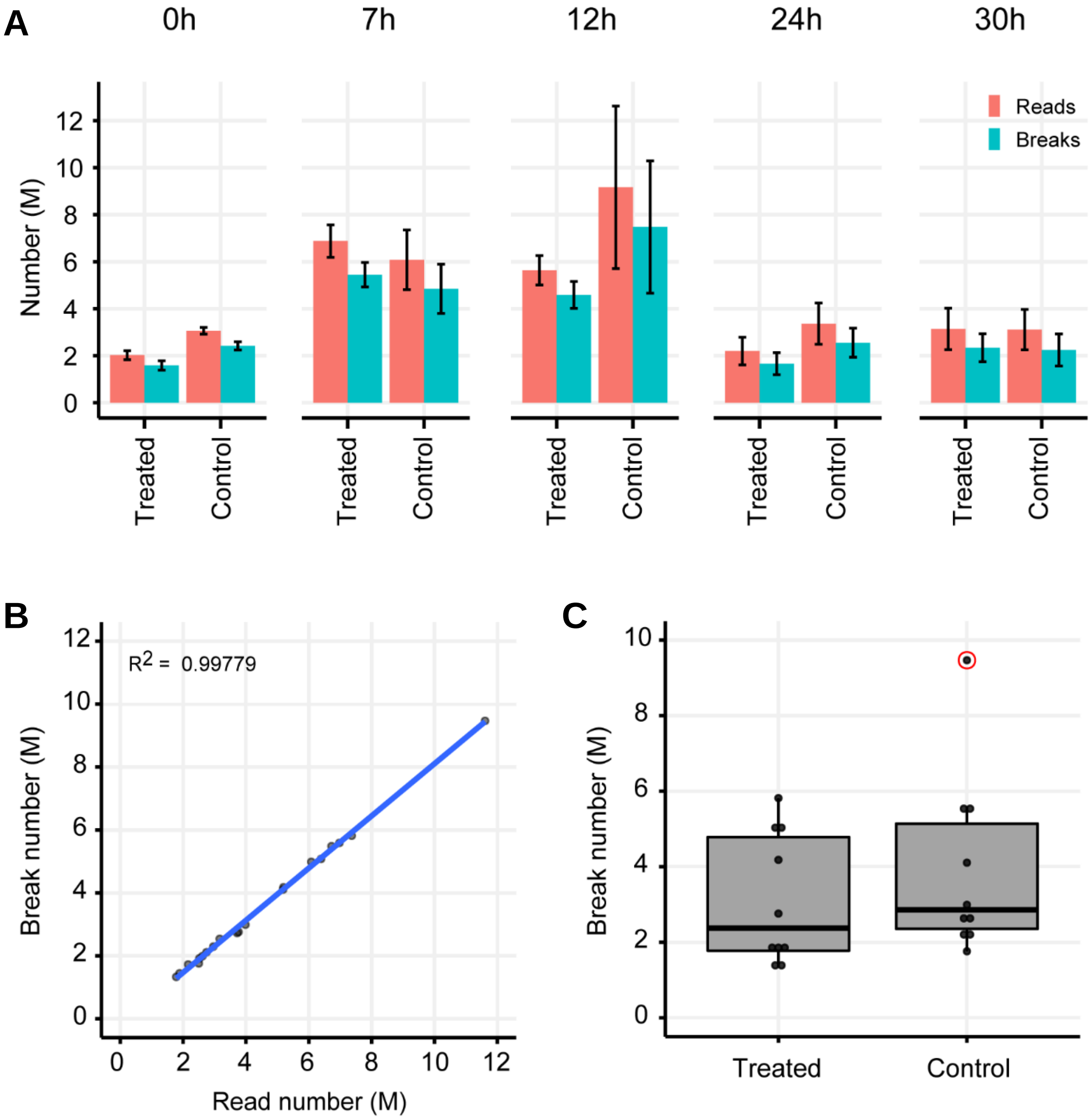


Figure 5.2. The number of sequencing reads and genomic breaks detected for INDUCE-seq for EXM1-treated and control samples. (**A**) The number of reads sequenced, and breaks defined for each INDUCE-seq sample. (**B**) Scatter plot of the number of INDUCE-seq reads sequenced and the number of breaks defined from individual INDUCE-seq experiments. (**C**) Boxplot showing the distribution of the number of breaks detected from the EMX1 treated and control samples. The outlier datapoint in the control set is marked with a red circle. Treated and control samples (n=2), error bars as SD.

5.2.3 Visualisation of DSBs induced at the EMX1 target site following editing

A simple way to visualise CRISPR-Cas9 induced DSBs detected by INDUCE-seq, is to examine reads which mapped at the EMX1 on-target site using a genome browser. Given that indel formation was detected at these sites at 30h following CRISPR treatment, significant break activity is expected at this site throughout the time course. Visualisation of a 5 kb region around the EMX1 target site (Chr2:73,158,860 - 73,163,860) was performed using the treated sample where the greatest number of breaks were observed at the EMX1 on-target site (7h, r2) (**Figure 5.3**). Around 270 reads were mapped around the target site (**Figure 5.3, treated sample, coverage track shown in grey**), showing significant breakage similar to the level observed previously for AsiSI-induced breaks in DIvA cells (**4.2.2**). Furthermore, similarly to the pattern observed at HindIII and AsiSI-induced break sites (**4.2.1**), adjacent stacks of plus strand and minus strand mapped reads were observed demonstrating that both sides of the CRISPR-Cas9 induced DSBs are captured by INDUCE-seq. In contrast to this, breaks mapped to the regions flanking the EMX1 target site, and across the entire region for the control sample, are representative of low-level endogenous breaks which are sparse and rarely recurrent within the sample population. Displaying the on-target site in this manner confirmed the presence of CRISPR-induced breaks at the EMX1 target site in the treated samples, and lack of CRISPR-induced breaks in the control samples.

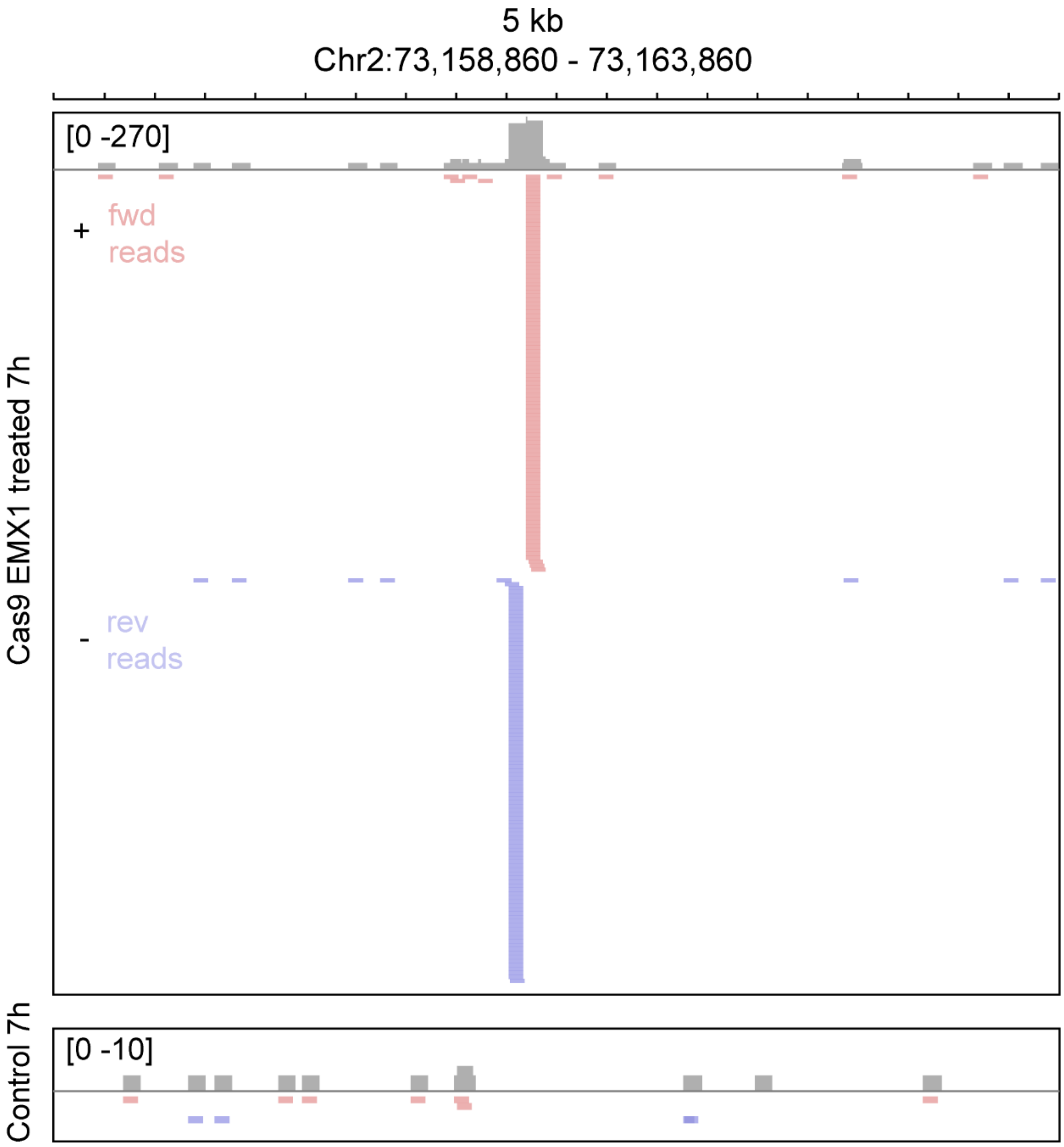


Figure 5.3. Genome browser view of EMX1 on-target DSB induction. A 5kb section of chromosome 2 shows the on-target cleavage pattern induced by EMX1. INDUCE-seq reads mapped to the plus strand (purple), and minus strand (blue) are positioned in recurrent stacks either side of the DSB position. Single endogenous breaks are present in the region surround the EMX1 target and across the entire region of the untreated control sample**.**

Closer inspection of the EMX1 target site in the treated sample revealed the precise positioning of the reads mapping at the CRISPR induced DSBs. **Figure 5.4** shows the read coverage track observed in a 180bp window around the target site, and a closer view of the reads positioned relative to the underlying target sequence. Interestingly, although a similar number of reads are observed on the right (plus strand) and left (minus strand) side of the DSB as expected, the narrow peak at the centre of the coverage track indicates the reads are overlapping rather than being positioned adjacently. Indeed, the read position in the 40bp window shows that although the reverse (rev) reads mapping to the minus strand are positioned at the expected break position at 3bp upstream of the EMX1 PAM site, the forward (fwd) reads on the plus strand are positioned 1 bp upstream of the expected position (**Figure 5.4, bottom panel dotted line**). This overlapping pattern of strand breakage has been reported previously for the same EMX1 guide RNA by Wienert *et al.* 2019. Using DISCOVER-seq, they reported that EMX1 exclusively induced a 1bp-overhanging DSB instead of the canonical blunt-ended DSB generally associated with CRISPR-induced strand breakage (Shi et al. 2019). In addition to showing that INDUCE-seq agrees with the published literature for this guide, these findings confirm that INDUCE-seq can be used to infer the end structure at CRISPR-Cas9 induced DSBs at single nucleotide resolution. This aspect will be explored in further detail in the next chapter.

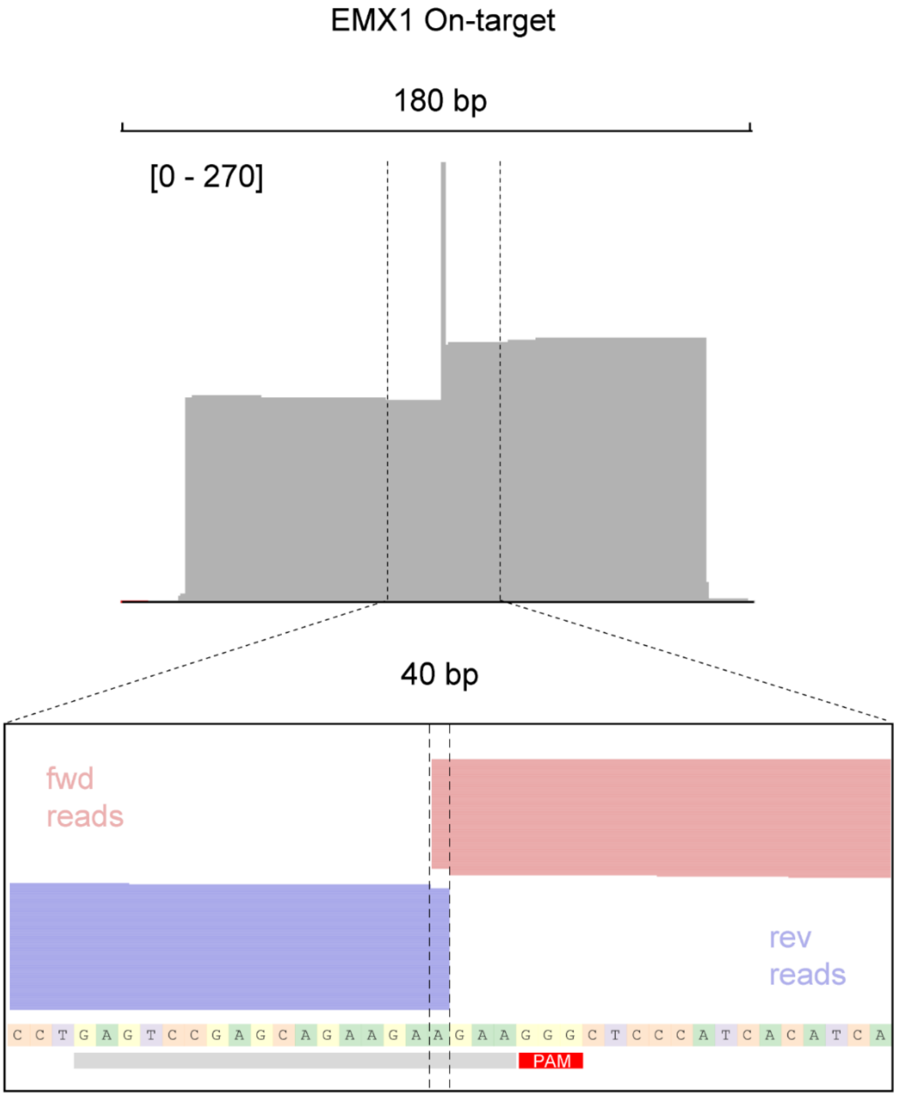


Figure 5.4. Base pair resolution genome browser view of the INDUCE-seq reads mapped to the EMX1 on-target site. Plus strand (pink) and minus strand (blue) reads are positioned either side of the expected break site (dashed line) 3bp upstream of the CRISPR PAM (red box). Reads are positioned with a 1bp overhang, suggesting that a staggered break end was induced rather than a blunt-end.

5.2.4 The development of a bioinformatics discovery pipeline for CRISPR induced off-target breaks

As stated in the chapter aims, the primary goal of these experiments was to demonstrate the use of INDUCE-seq to measure CRISPR-induced off-targets throughout the genome. Despite the fact that several different genomic methods have defined a range off-targets for the EMX1 guide, none of the available off-target data analysis pipelines are compatible with the data output generated by INDUCE-seq. To address this, a novel bioinformatics off-target discovery pipeline was developed that could extract CRISPR-induced breaks from the background distribution of endogenous breaks within the genome. Briefly, this involved scanning the genome for potential off-target sites based on sequence, filtering the sites according to several parameters, and calculating the number of break overlaps at the expected off-target break site. To determine robust off-target discovery in the treated samples, a range of filter conditions were tested in order to reduce the false-discovery rate using the control samples. **Figure 5.5** outlines the details of the off-target discovery pipeline for filter condition 1, and **Figure 5.6** shows the range of filtering conditions tested and the effect on off-target discovery. For simplicity, the discovery pipeline will be described using filter condition 1, however the same main steps apply for any of the filtering parameters used.

The first stage of the of the off-target discovery pipeline predicts the full spectrum of EMX1 induced off-target positions using the homology search tool Cas-OFFinder (Bae et al. 2014). This tool scans the human reference genome with an input guide RNA sequence and returns every matching position up to N bp mismatches. Using filter condition 1 as an example, up to 6 mismatches were allowed across the whole EMX1 guide RNA sequence including the NGG PAM region, which identified 35,046 possible genomic positions. Next, these positions were filtered to remove sites based on the number of mismatches in the seed region of the guide spacer sequence, defined as the 12 nucleotides proximal to the PAM site. Previous studies have demonstrated that mismatches in the seed region substantially reduce Cas9 activity *in vitro* (Hsu et al. 2013), however this number is not clearly defined and is likely to differ for all guide RNA sequences. For filter condition 1, up to 2 mismatches were allowed in the seed region, which reduced the number of potential sites from 35,056 to 16,683.

Next, each of the off-target positions were reduced from full 23bp genomic intervals to the 2bp expected break position, 3-4bp upstream of the PAM site. A 2bp window was used to account for the positioning of breaks depending on whether they are on the right (plus strand) and left (minus strand) side of the DSB (**Figure 5.5, pink and blue triangles**). Furthermore, for guides such as EMX1 that seem to cleave with a 1bp overhang rather than a blunt-ended DSB (**Figure 5.4**), the 2bp window would still capture the offset break-end. For each of the treated and control samples, overlaps were calculated between the INDUCE-seq 1bp break positions and the 16,683, 2bp-wide, potential off-target positions, thus leaving only sites where breaks are positioned precisely at predicted off-target sites.

The final stage of the discovery pipeline filters the sites based on the mismatch number and number of breaks observed for each off-target. When predicting sites based on mismatches, the number of potential sites increases exponentially with increasing mismatches from the target site. This therefore increases the probability of endogenous breaks unrelated to CRISPR activity overlapping by chance at EMX1-like sequences. We have established previously (**Chapter IV**) that recurrent endogenous breaks are rare and single endogenous breaks make up the vast majority of breaks in the genome. Therefore, off-target sites identified with a *single* break at an EMX1-like sequence are the most likely to be called by *chance*. To address this and remove these potential false-positive off-targets, the final filter uses both mismatch number and break number to remove sites with (i) a low break occurrence and (ii) a high mismatch number. Thus for filter condition 1, off-targets identified with a single break were rejected if they contained more than 2 total mismatches across the whole off-target site including the PAM.

The specific parameters used for filter condition 1 detect EMX1 off-target break sites if: (i) ≤6 mismatches were present in the spacer and PAM, (ii) ≤2 mismatches were present in the seed sequence, and (iii) if only a single break is present at the off-target site, and contained no more than 2 mismatches. Using these conditions, 81 EMX1 off-target sites were identified across the full set of CRISPR-treated samples, while the full set of corresponding control samples only resulted in 3 (false-positive) off-target sites.

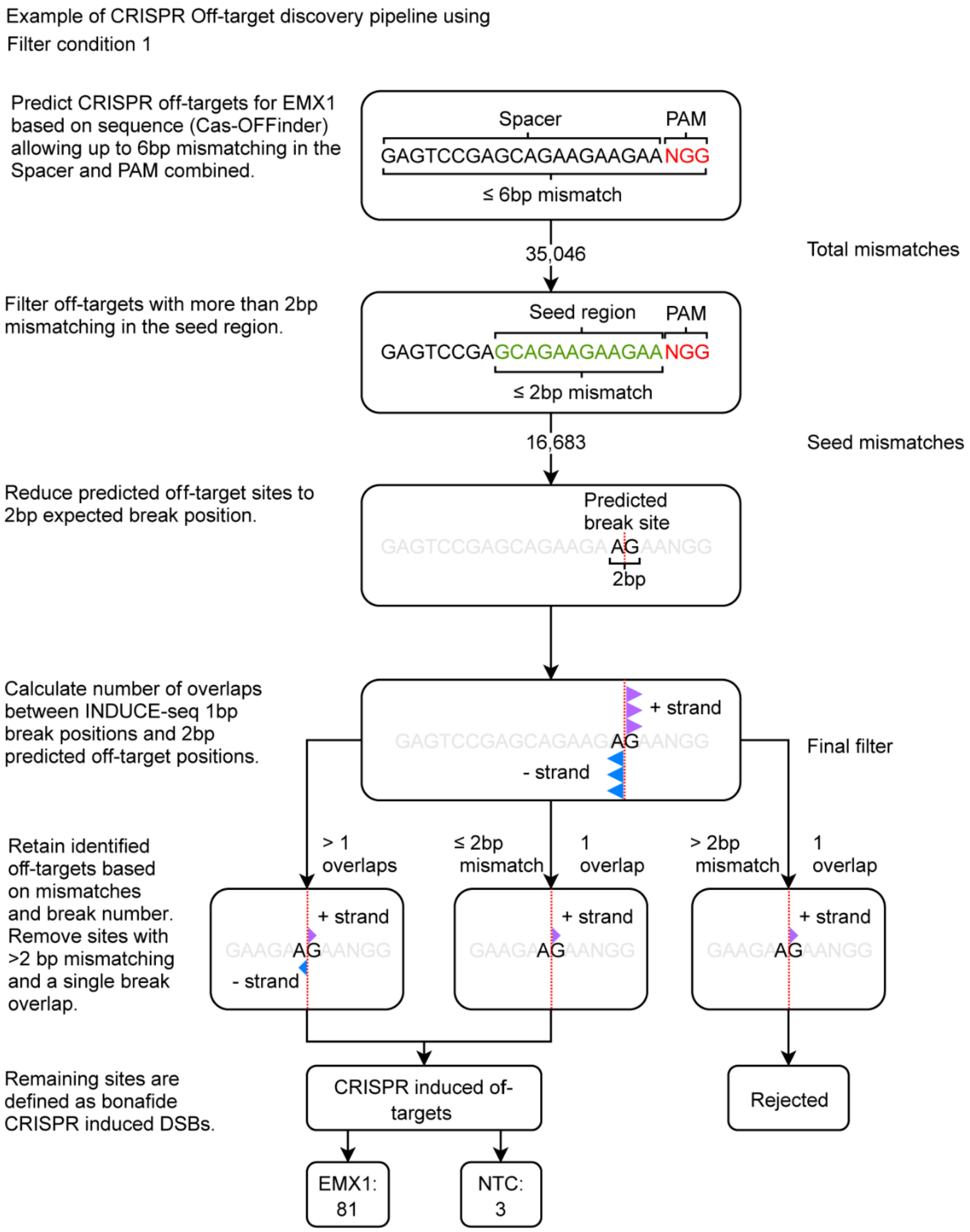

Example of CRISPR Off-target discovery pipeline using
Filter condition 1
Predict CRISPR off-targets for EMX1 based on sequence (Cas-OFFinder) allowing up to 6bp mismatching in the Spacer and PAM combined.
Spacer
PAM
GAGTCCGAGCAGAAGAAGAANGG
≤ 6bp mismatch
35,046
Total mismatches
Filter off-targets with more than 2bp mismatching in the seed region.
Seed region
PAM
GAGTCCGAGCAGAAGAAGAANGG
≤ 2bp mismatch
16,683
Seed mismatches
Reduce predicted off-target sites to 2bp expected break position.
Predicted break site
GAGTCCGAGCAGAAGAAGAANGG
2bp
Calculate number of overlaps between INDUCE-seq 1bp break positions and 2bp predicted off-target positions.
+ strand
GAGTCCGAGCAGAAGAAGAANGG
- strand
Final filter
Retain identified off-targets based on mismatches and break number. Remove sites with >2 bp mismatching and a single break overlap.
> 1 overlaps
≤ 2bp mismatch
1 overlap
> 2bp mismatch
1 overlap
+ strand
GAAGAAGAANGG
- strand
+ strand
GAAGAAGAANGG
+ strand
GAAGAAGAANGG
Remaining sites are defined as bonafide CRISPR induced DSBs.
CRISPR induced of-targets
Rejected
EMX1: 81
NTC: 3

Figure 5.5. CRISPR off-target discovery pipeline. Example of the procedure for the identification of off-target CRISPR-induced DSBs. The numbers shown refer to filter condition 1.

To assess how different filtering parameters alter off-target discovery, 32 different filter conditions were tested and compared to determine the right balance between stringency and discovery rate for the EMX1 guide (**Figure 5.6**). This included (i) ≤6 or ≤7 mismatches across the seed and PAM region in the first off-target homology search, (ii) filtering seed mismatches with decreasing stringency of ≤2, ≤3, ≤4, and ≤5, and (iii) final filtering based on rejecting single breaks at sites with >2, >3, >4 and >5 mismatches. For simplicity, each of these parameters was tested by assessing off-target discovery across the full set of 10 treated and control samples. The effect of the varying stringency of filter conditions can be seen in **Figure 5.6**. This shows the total number of off-targets detected across the treated and control samples (**Figure 5.6, pink and blue bars**), in addition to the number of off-targets detected per sample (**Figure 5.6, pink and blue boxplots**). The most stringent combination of filter parameters used was for filter condition 1, which identified 81 off-targets in the EMX1 treated samples and 3 off-targets in the control samples. Correspondingly, the least stringent combination of parameters was for filter condition 32, which identified 259 off-targets in the EMX1 treated samples and 105 off-target in the control samples. Visualising the spectrum of filter conditions in this way demonstrates exactly how the different levels of filtering behave in relation to the number of off-targets identified in the treated and control datasets. By way of example, filter condition 1 and filter condition 17, which differ only by the total number of mismatches allowed in the first off-target homology search, identified an identical number of off-targets in the treated samples (81), with just 2 additional off-targets identified in the control sample for filter condition 17 (3 and 5, respectively). Furthermore, comparing between filter conditions 1 and 2: simply reducing the stringency of the final filter - rejecting off-targets with a single break if they possess >2 to >3 mismatches - substantially increased the number of off-targets identified in the treated datasets, from 81 to 107 with no additional increase in the number identified in the control samples.

To determine which of the filter conditions was the most appropriate for off-target discovery, the false discovery rate was calculated for each of the filter conditions. True positive (TP) off-targets were defined as the sites identified across the treated sample group (**Figure 5.6, pink bars**), and false positive (FP) off-targets were defined as those identified across the control sample group (**Figure 5.6, blue bars**). Next, the false discovery rate (FDR = FP/(FP+TP)) for each condition was plotted against off-target yield, defined as the total number of off-targets identified in the treated set minus the control set (**Figure 5.7**). Interestingly, when plotted in this manner the various filter conditions appeared to cluster most prominently based on the final filter condition used. Rejecting off-targets with a single break and more than 4 mismatches appeared to generally produce the greatest off-target yield with a low false discovery rate of <0.15 (**Figure 5.7, blue circles**). Given that a filter condition comprising the perfect set of parameters would position in the bottom right corner of the graph (high off-target discovery, with low false-discovery rate), condition 7 (**Figure 5.7, red highlight)** was selected as the best performing condition for efficient off-target discovery for EMX1. Therefore, all subsequent downstream analysis was conducted using the off-targets discovered using filter condition 7.

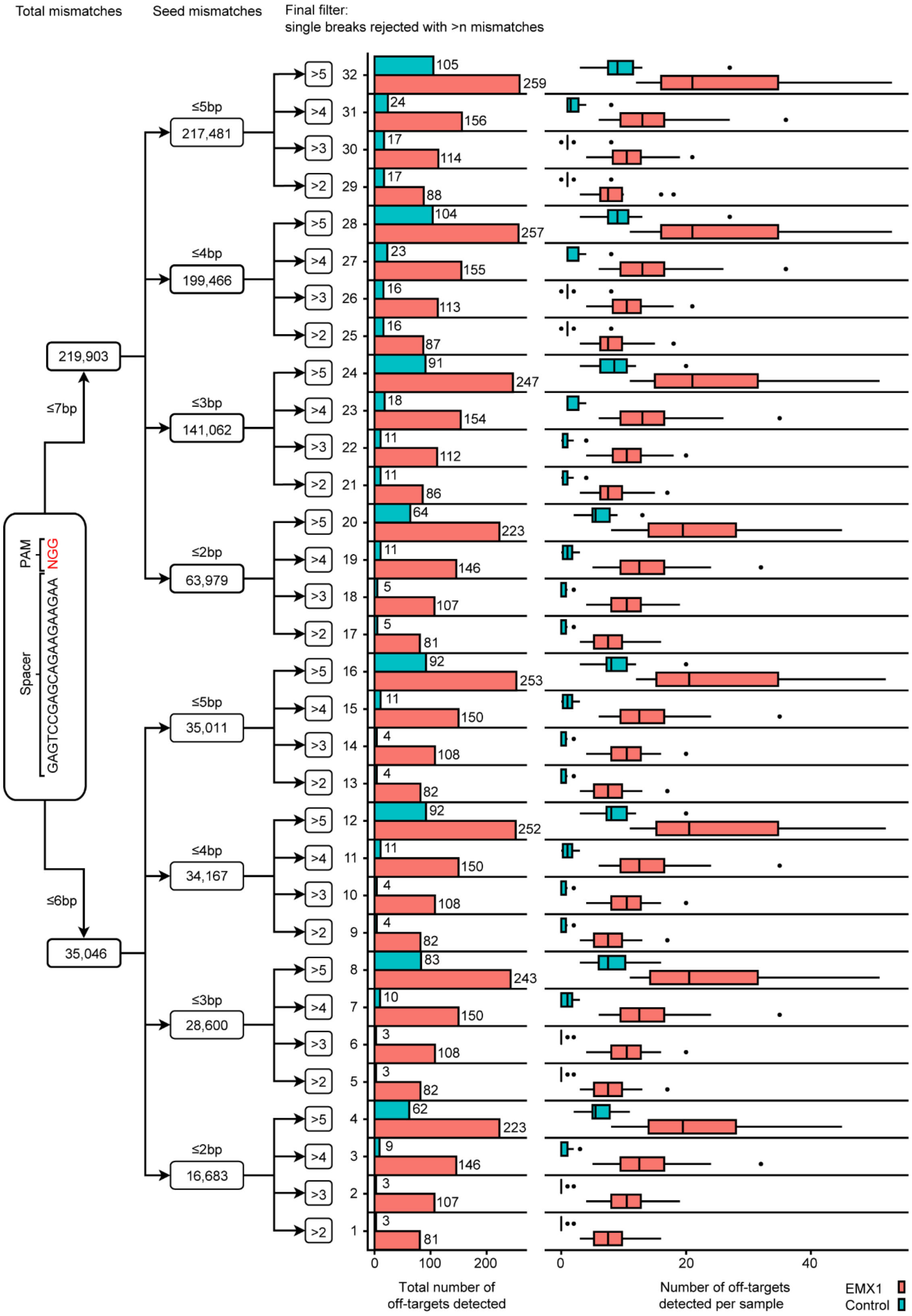

Total mismatches
Seed mismatches
Final filter:
single breaks rejected with >n mismatches
PAM
NGG
Spacer
GAGTCCGAGCAGAAGAAGAA
≤7bp
219,903
≤6bp
35,046
≤5bp
217,481
≤4bp
199,466
≤3bp
141,062
≤2bp
63,979
≤5bp
35,011
≤4bp
34,167
≤3bp
28,600
≤2bp
16,683
>5
>4
>3
>2
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
105
259
24
156
17
114
17
88
104
257
23
155
16
113
16
87
91
247
18
154
11
112
11
86
64
223
11
146
5
107
5
81
92
253
11
150
4
108
4
82
92
252
11
150
4
108
4
82
83
243
10
150
3
108
3
82
62
223
9
146
3
107
3
81
0
100
200
Total number of
off-targets detected
0
20
40
Number of off-targets
detected per sample
EMX1
Control

Figure 5.6. Overview of all filter conditions used for off-target discovery. (**Left side**) The three levels of filtering employed as described in the discovery pipeline. A variable number of mismatches was tested at each level of filtering, resulting in 32 combinations of filter conditions. (**Right side**) For each filter condition a set of CRISPR-induced DSBs was determined for EMX1 treated and control samples. The total number of off-target sites for all EMX1 and control samples combined is shown by the horizontal bars. The boxplots show the number off-targets detected for each individual dataset.

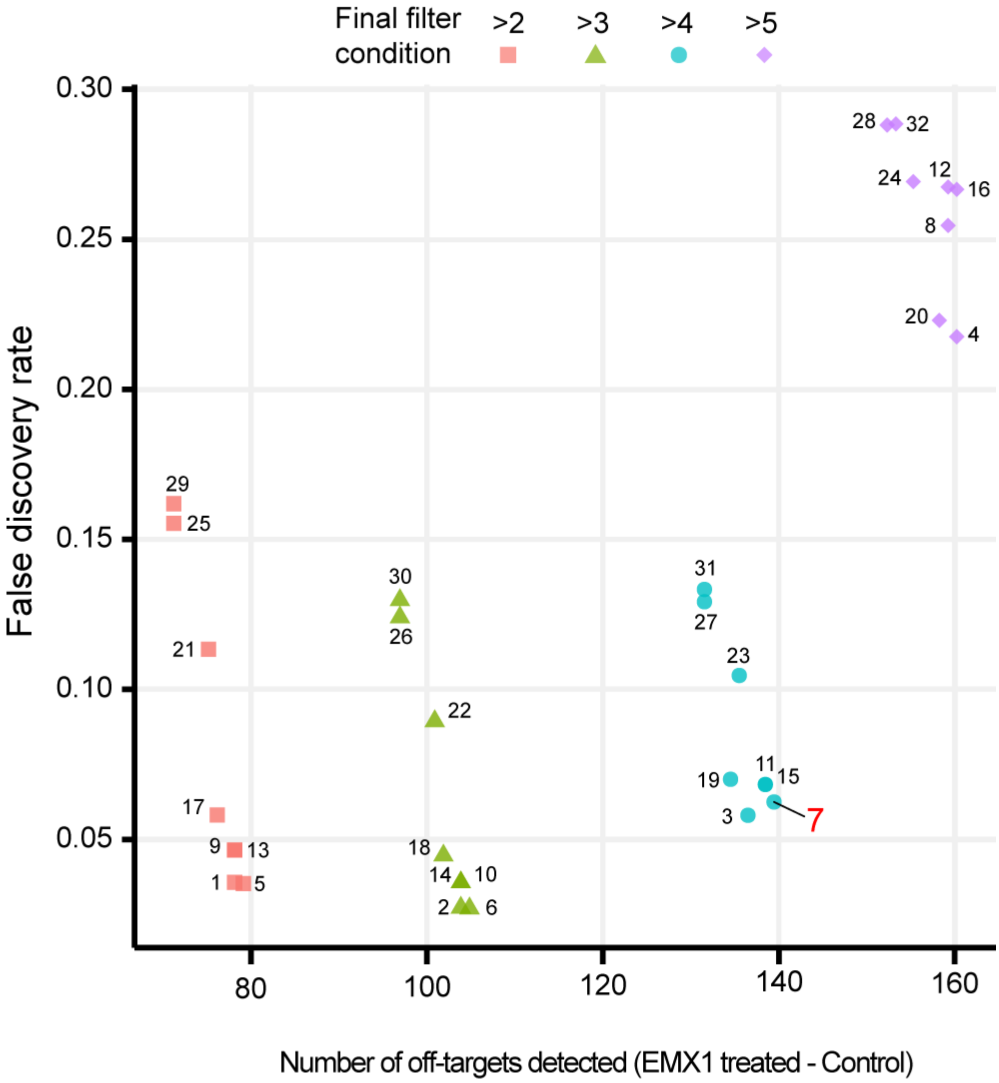


Figure 5.7. Selection of parameters for determining CRISPR off-target discovery. Scatterplot showing the relationship between EMX1 off-target site yield and the false discovery rate for each filter condition (1-32). Condition 7 (red highlight) was selected for all subsequent analysis.

5.2.5 Quantification of the on- and off-target editing by EMX1 using INDUCE-seq

Having determined a set of EMX1 off-targets from the INDUCE-seq samples, it is now possible to quantitatively compare on-target and off-target break induction to better understand the kinetics of on- and off-target editing throughout the genome. **Figure 5.8A** shows the number of breaks identified at the on-target site (pink bars), and the number of breaks identified across all the off-targets identified at each time point (blue bars). On-target break activity can be seen immediately at 0h, which then peaks at 7h, and decreases towards the end of the time course at 30 hours. Aggregate off-target activity generally appeared to follow the same trend. Strikingly, immediately following cell treatment at 0h, more breaks were found at off-targets than at the on-target site. Similarly, although off-target break number was lower than at the on-target at 7h, 12h and 24h, approximately the same number of breaks were found at off-target sites as the on-target at the 30h time point. In summary these observations reveal similar kinetics between on-target and aggregate off-target break induction across all sites.

In addition to comparing the total number of breaks across all off-target sites, it is also important to consider how many distinct off-target sites are identified for each of the time points (**Figure 5.8B**). As with the total number of breaks identified at off-targets, the number of off-target sites peaked at 7h. This confirms that off-target activity was also greatest at this time point. Interestingly, the next highest time point for number of off-target sites was at 0h. This is despite the fact that ~2-fold more breaks were measured at off-targets at 12h. These findings suggest that at 0h, many off-targets are identified with a low number of breaks, and that at 12 hours many breaks are measured across a smaller number of different off-target locations. Collectively, these results show that on- and off-target break activity changes dramatically at the gross level during the time course following cell treatment with the Cas9-EMX1 RNP. Furthermore, it also shows that differences in off-target site abundance can be observed independently of total off-target break activity.

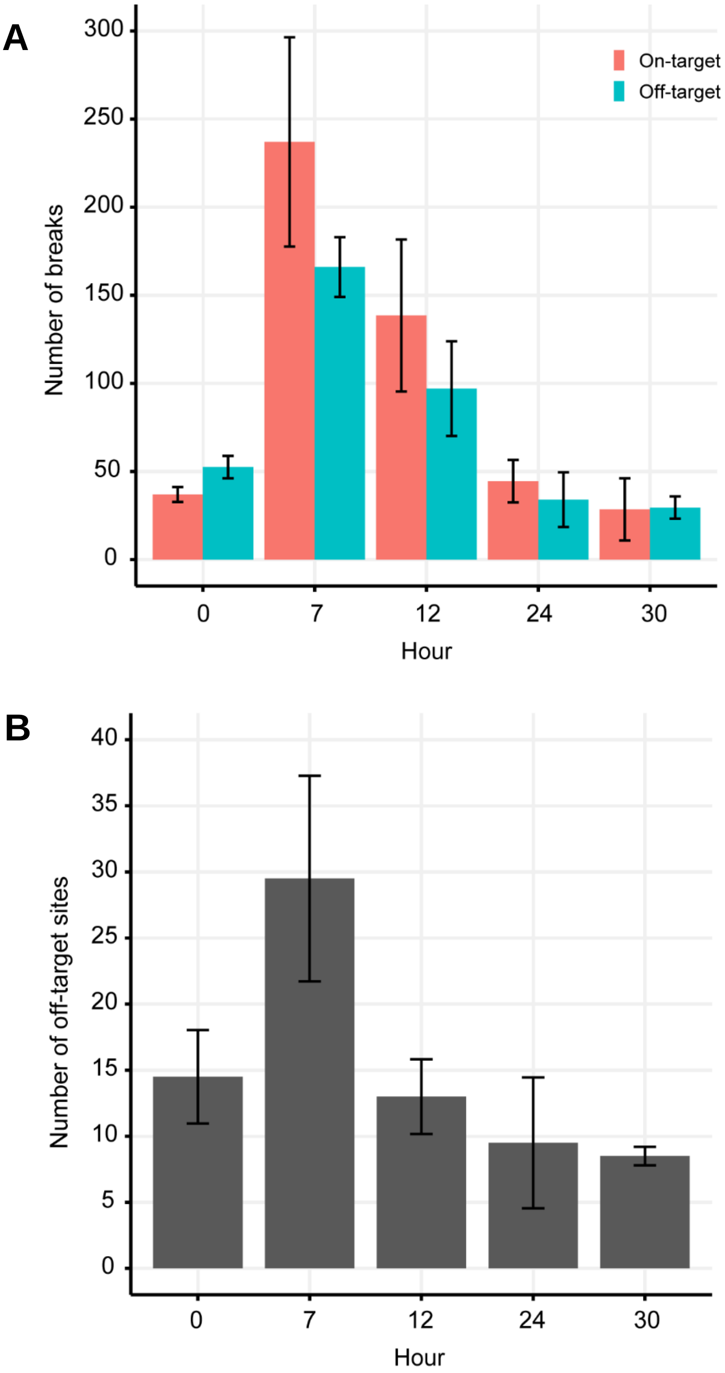


Figure 5.8. On and off-target break activity measured by INDUCE-seq for EMX1 at different time points. (**A**) The average number of breaks measured at the on-target (pink bars) compared to total number of breaks at all the off-target sites (blue bars) for time course r1 and r2. (**B**) The average number of off-target sites detected for each of the time points. Error bars as SD, n=2.

In order to understand the kinetics of CRISPR break induction, it is important to consider the cell number measured during the editing time course. This is because differences in cell number could contribute to the number of total and CRISPR-induced breaks measured. Therefore, the number of total and CRISPR-induced breaks was compared with the amount of DNA used for each INDUCE-seq sample library preparation (**Figure 5.9A and 5.9B**). However, no correlation was observed between the DNA used for each INDUCE-seq library and the total sample break number (R^2 = 0.0997791) (**Figure 5.9A**). Similarly, no correlation was observed between the number of on-target breaks with the amount of INDUCE-seq library DNA used for each sample (R^2 = 0.114) (**Figure 5.9B**). This lack of correlation confirms that for the set of CRISPR treated samples, the total number of breaks and the number of on-target CRISPR breaks present in each sample were not determined by the amount of INDUCE-seq library DNA, and is therefore independent of the number of cells processed for each sample. This means that the number of total and CRISPR-induced breaks measured is not simply determined by the number of cells in the sample. Therefore, in this setting the cell number (library DNA amount) is not appropriate for normalisation between samples measured within the time course.

It was noted that the pattern of total break induction, during the time course, was similar to the pattern of CRISPR-induced breaks (**Figure 5.2A and Figure 5.8A**). Therefore, it was considered whether the total number of breaks could be used to normalise between samples during the time course. **Figure 5.9C** and **5.9D** show that the number of on- and off-target breaks was strongly correlated with the total break number (R^2 = 0.87237 and R^2 = 0.82242, respectively). Similarly, as expected, the total number of off-target sites also correlate with the total sample break number (R^2 = 0.77155) (**Figure 5.9E**). Therefore, normalisation of CRISPR-induced breaks as a proportion of the total breaks measured, enables comparisons between different sampling times during the editing process to be evaluated quantitatively. This approach is therefore applied for normalisation for the remainder of this thesis when kinetics of editing is being investigated.

Finally, when comparing the number of on- and off-target breaks for each sample during the time course, a very strong linear correlation is observed (R^2 = 0.95322) (**Figure 5.9F**). This demonstrates that off-target break activity is proportional to the on-target break activity, irrespective of the time point examined. The implications of this for developing safe genome editing strategies are significant, as the proportion of on- and total off-target editing occurs independently of time.

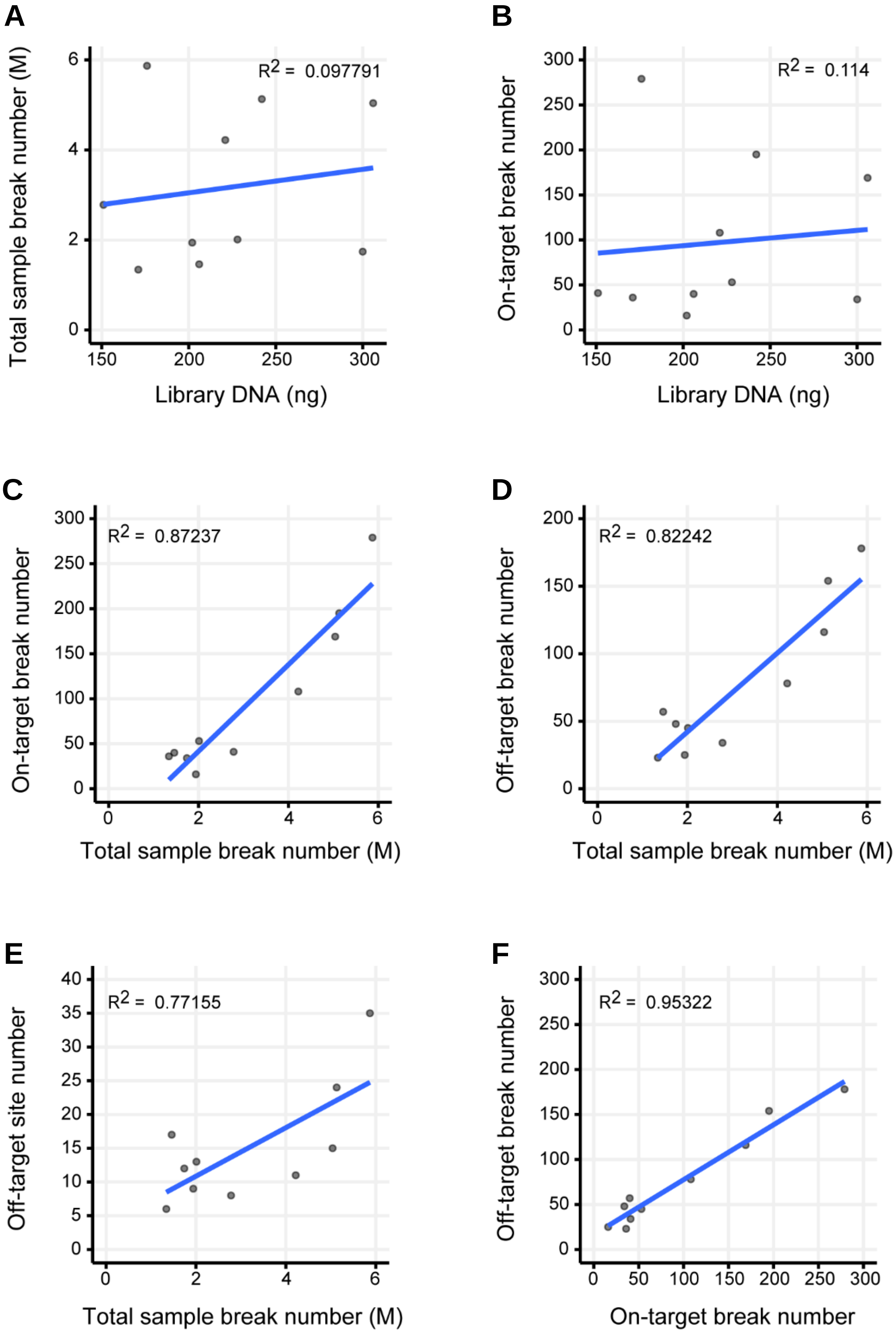

A
Total sample break number (M)
6
4
2
0
R2 = 0.097791
150
200
250
300
Library DNA (ng)
B
On-target break number
300
250
200
150
100
50
0
R2 = 0.114
150
200
250
300
Library DNA (ng)
C
On-target break number
300
250
200
150
100
50
0
R2 = 0.87237
0
2
4
6
Total sample break number (M)
D
Off-target break number
200
150
100
50
0
R2 = 0.82242
0
2
4
6
Total sample break number (M)
E
Off-target site number
40
35
30
25
20
15
10
5
0
R2 = 0.77155
0
2
4
6
Total sample break number (M)
F
Off-target break number
300
250
200
150
100
50
0
R2 = 0.95322
0
50
100
150
200
250
300
On-target break number

Figure 5.9. Characterising and comparing CRISPR activity. (**A**) and (**B**) No correlation was observed between the amount of INDUCE-seq library DNA and the total number of breaks or the number of EMX1 on-target breaks. (**C**) and (**D**) Total break number correlates strongly with the number of on- (**C**) and off-target breaks (**D**) and throughout the time course. (**E**) Similarly, total break number correlates with the number of different off-target sites. (**F**) The number of breaks at the on-target and the number of breaks at off-targets proportional across all samples of the time course**.**

5.2.6 Demonstrating the reproducibility of off-target discovery using INDUCE-seq

In section **5.2.5** it was demonstrated that on-target and total off-target break number was consistent between replicate sample time points. Importantly, this analysis does not provide individual site information for the off-target sites, which is paramount in understanding how reproducible INDUCE-seq is for off-target site discovery. Furthermore, this information is also crucial to characterise the pattern of off-target editing for any given guide RNA. To examine the reproducibility of INDUCE-seq, the off-target sites identified within all r1 and r2 replicates were compared (**Figure 5.10A**). Across the r1 time course a total of 34 different off-target sites were identified, while across the r2 time course 47 different sites were identified. Between the two replicates 21 sites were shared, representing 61.8% of the r1 set, and 44.7% of the larger r2 set. Assuming that EMX1 RNP CRISPR activity is consistent between two cell populations, the number of breaks found at each of the shared sites between the replicate datasets was compared to determine if DSB induction was consistent on a per site basis (**Figure 5.10B**)**.** This analysis revealed a very strong correlation between the r1 and r2 time course break count at shared off-targets (R^2 =0.9965). Moreover, break induction varies at the off-targets depending on the number of mismatches present in the site. Each of the data points in **Figure 5.10B** are coloured based on the mismatch number present in the off-target site, showing reproducible measurement of the different classes of sites from the on-target at nearly 1,000 breaks, to very low frequency off-targets with 5 mismatches measured just once in each time course. These results demonstrate that CRISPR induced DSBs and INDUCE-seq break measurement is highly reproducible between experimental replicates.

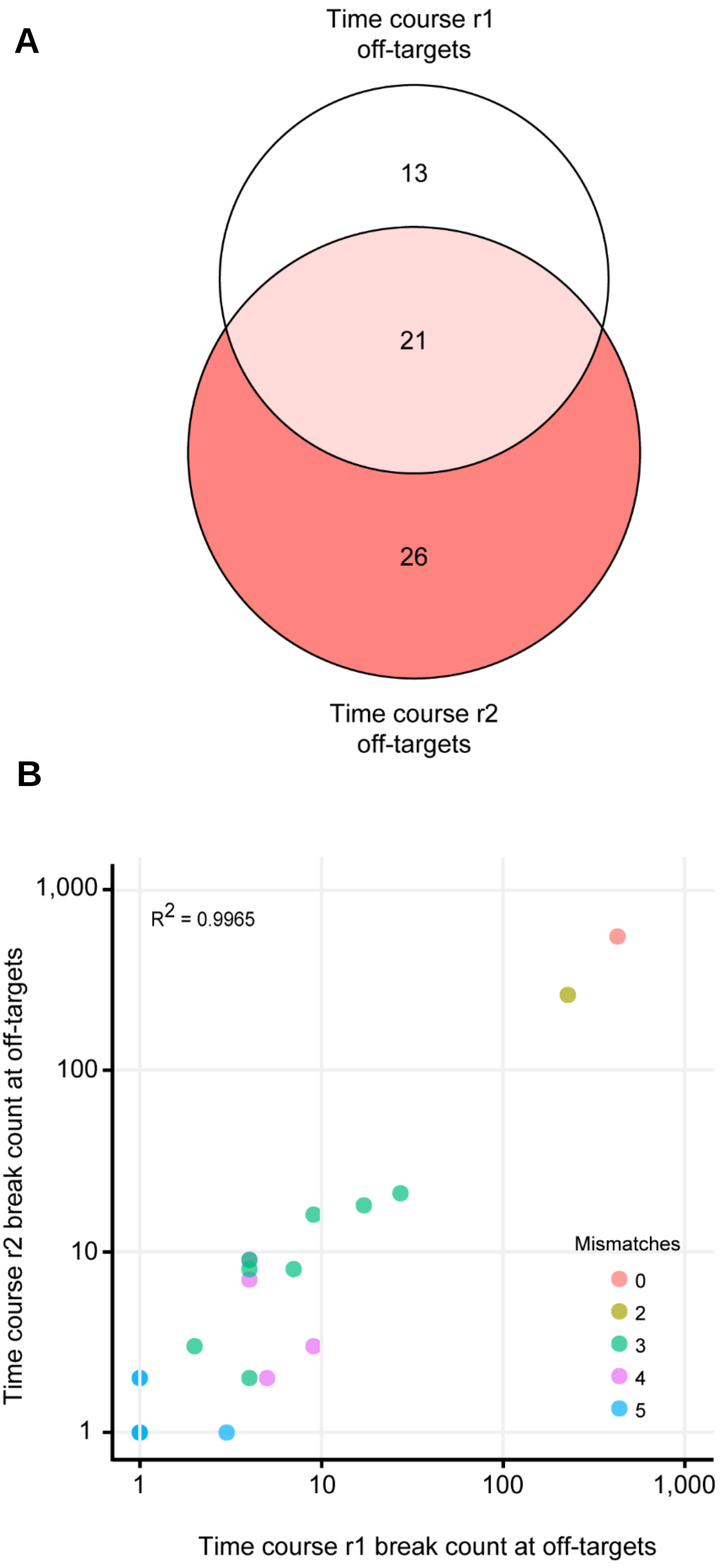


Figure 5.10. Comparing the off-targets identified in time course r1 and r2 experiments. (A) Euler diagram comparing the off-target sites identified in time course r1 (n=34) and r2 (n=47**). (B)** Scatterplot showing the break number found at CRISPR off-target sites identified in both independent experiments.

5.2.7 Assessing the kinetics of EMX1 off-target formation and comparison with alternative methods

Given the high reproducibility between the replicate time courses, the off-target datasets from both time course experiments (r1 and r2, 0h-30h, 10 samples total) were combined to generate a single pooled dataset for off-target characterisation (**Appendix D, Table A5**). The combined set of r1 and r2 off-targets totalled 60 sites and was plotted as a CRISPR off-target mismatch plot to present the profile of mismatches and ranked by the number of breaks identified at each site (**Figure 5.11A**). With reference to the EMX1 target sequence, the position of mismatches in the off-target sites are highlighted based on the mismatch type. As reported previously for EMX1, many off-target sites possess up to 6 base pairs mismatching from the target sequence, with the majority of the protospacer mismatches present at the 5' distal end of the target (Tsai et al. 2015). Mismatches were also observed in the seed region of the target, although to a lesser extent, and with a more homogeneous mismatch pattern. Interestingly, off-targets with a range of noncanonical PAM sequences were also observed, including NAG, NGA, NCG, NTG and NAT. Next, the number of breaks detected at each of the individual off-target sites was quantified separately for each of the time points to better understand the kinetics of editing throughout the time course (**Figure 5.11B**). For this analysis, the off-target break numbers from the replicate sample time points were combined and divided by the total number of reads sequenced in millions to normalise between samples. As was shown in **Figure 5.9**, total breaks (and therefore reads) per sample provides a more appropriate means for sample normalisation compared to using the total DNA of each INDUCE-seq sample. Strikingly, in addition to the on-target site, substantial activity can be observed at the most frequently broken off-target (OT1). OT1 was broken at roughly 50% of the frequency of the on-target site. This demonstrates sustained and highly active cleavage from 0 to 30 hours post-cell treatment. Furthermore, several other off-targets (OT2, 4, 5 and 6) were also detected at every time point, although at much lower frequency (between 1.34%-4.94%) of that of the on-target. Although the greatest number of off-target sites was detected at 7 hours post nucleofection, off-target activity, as measured by breaks per million was higher for many off-targets at 0 hours in line with that observed in **Figure 5.8**. Specifically, many of the top 12 off-target sites showed higher breaks per million reads at 0 hours, which represents repeated cleavage directly following cell nucleofection. These observations show that due to the quantitative nature of INDUCE-seq it possible to assess the kinetics of individual off-target sites accurately and simultaneously during CRISPR-Cas9 editing.

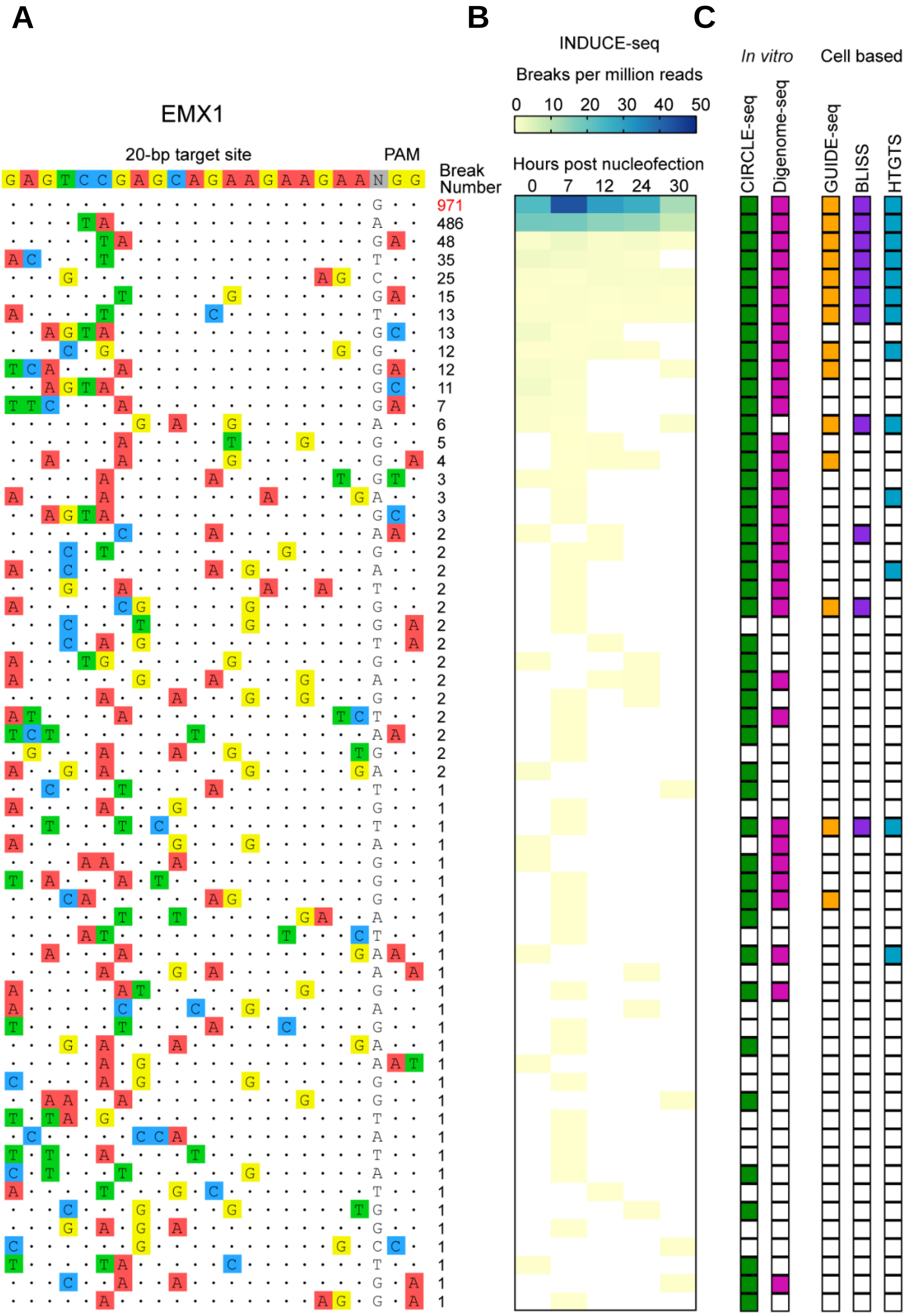

A
B
C
EMX1
20-bp target site
PAM
G A G T C C G A G C A G A A G A A G A A N G G
Break Number
971
486
48
35
25
15
13
13
12
12
11
7
6
5
4
3
3
3
2
2
2
2
2
2
2
2
2
2
2
2
2
2
1
INDUCE-seq
Breaks per million reads
0 10 20 30 40 50
Hours post nucleofection
0 7 12 24 30
In vitro
Cell based
CIRCLE-seq
Digenome-seq
GUIDE-seq
BLISS
HTGTS

Figure 5.11. INDUCE-seq sensitively discovers and quantifies CRISPR-Cas9 induced on- and off-target DSBs. (**A**) Off-target sequences and the number of breaks identified using INDUCE-seq for the EMX1 sgRNA. (**B**) INDUCE-seq reveals the kinetics of EMX1-induced DSB formation in a cell population. Quantification of the number of breaks detected per million reads for each sample revealed high Cas9 activity both on- and off-target immediately following cell nucleofection. (**C**) The comparison between off-targets identified by INDUCE-seq with established *in vitro* methods CIRCLE-seq and Digenome-seq, in addition to cell-based methods GUIDE-seq, BLISS, and HTGTS. INDUCE-seq detects many off-targets that were previously only identifiable by *in vitro* approaches. Substantially more off-target sites were identified than by any of the current cell-based methods. INDUCE-seq also identifies multiple off-targets not detected by any other method.

The final part of **Figure 5.11** compares the profile of off-targets measured for EMX1 using INDUCE-seq to the off-targets discovered using a range of alternative methods (**Figure 5.11C**) (Frock et al. 2015; Kim et al. 2015; Tsai et al. 2015; Tsai et al. 2017; Yan et al. 2017; Wienert et al. 2019). These methods are separated in to two categories based on whether they measure Cas9 cleavage of purified genomic DNA *in vitro*, or whether they are cell-based methods and measure Cas9 activity directly in cultured cells. Of the 60 off-targets identified by INDUCE-seq for the EMX1 guide, only 13 were detected by GUIDE-seq, 10 were detected using BLISS, and 12 were detected using HTGTS. Although this overlap represented 13/15 of the off-targets reported using GUIDE-seq, all of the off-targets detected using BLISS and HGTGS were contained within the 60 sites detected using INDUCE-seq. Remarkably, INDUCE-seq also identified multiple off-targets for EMX1 in live cells that have previously only been observed using *in vitro* off-target discovery assays such as CIRCLE-seq and Digenome-seq. Furthermore, 65% (39/60) of the INDUCE-seq set was identified by CIRCLE-seq and 52% (31/60) was identified by Digenome-seq. Interestingly, INDUCE-seq also identified numerous off-targets that were not detected by CIRCLE-seq (15) or Digenome-seq (29), suggesting that *in vitro* cleavage may not accurately represent Cas9 activity in a cellular environment.

To better understand the impact of the time of sampling on off-target detection, the number of off-target overlaps between each individual timepoint sample and the sets generated by CIRCLE-seq, GUIDE-seq, and BLISS were calculated (**Figure 5.12**). **Figure 5.12A** and **5.12B** display Euler diagrams that quantify the number of overlaps between each sample for each of the methods mentioned above. At eight of the ten replicate time points, INDUCE-seq identified sites outside of those identified by CIRCLE-seq, GUIDE-seq and BLISS. This is further demonstrated when comparing the plots generated from all r1 samples combined (**Figure 5.12C**), and all r2 samples combined (**Figure 5.12D**). Combining all r1 and r2 replicates datasets (equivalent off-target set shown in **Figure 5.11**) revealed a total of 46 off-targets that were identified by INDUCE-seq that were not detected by GUIDE-seq and BLISS, and 21 off-targets that were not detected by any other method (**Figure 5.12E**). Across all comparison plots, CIRCLE-seq identified a substantial number of off-target sites not identified by any of the cell-based methods. However, it has been recognised that CIRCLE-seq and other *in vitro* approaches discover many false-positive sites that are not detected in cell-based experiments or *in vivo* (Wienert et al. 2019). In conclusion, these studies show that when compared to existing technologies, INDUCE-seq significantly outperforms the alternative CRISPR off-target detection methods, CIRCLE-seq, GUIDE-seq, and BLISS.

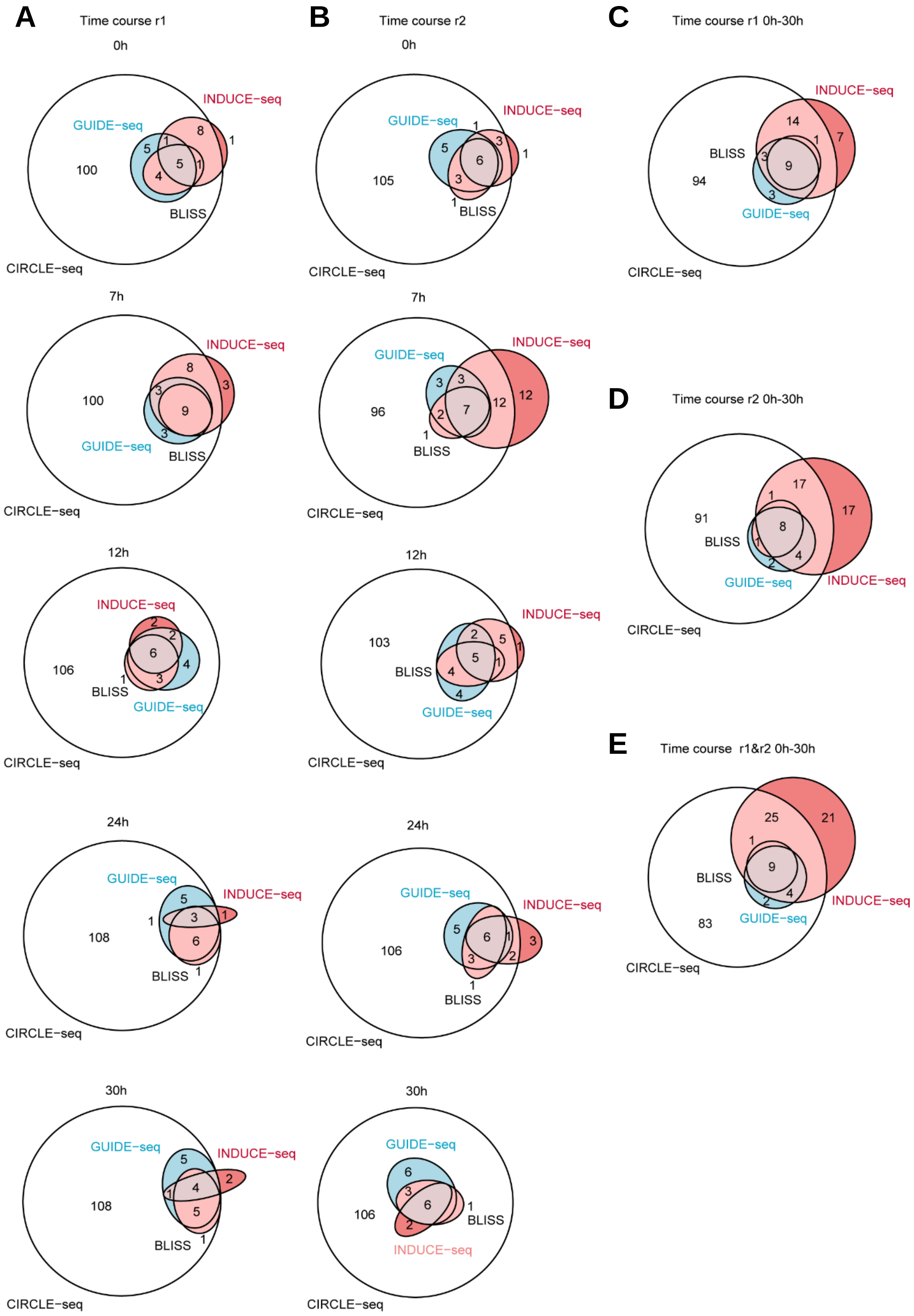

A
Time course r1
0h
INDUCE−seq
GUIDE−seq
8
1
5
1
100
5
1
4
BLISS
CIRCLE−seq
7h
INDUCE−seq
8
3
100
3
9
3
GUIDE−seq
BLISS
CIRCLE−seq
12h
INDUCE−seq
2
2
6
106
4
1
3
BLISS
GUIDE−seq
CIRCLE−seq
24h
GUIDE−seq
5
INDUCE−seq
1
3
1
108
6
BLISS
1
CIRCLE−seq
30h
GUIDE−seq
INDUCE−seq
5
2
4
108
1
5
BLISS
1
CIRCLE−seq
B
Time course r2
0h
INDUCE−seq
GUIDE−seq
1
5
3
1
6
105
3
1
BLISS
CIRCLE−seq
7h
INDUCE−seq
GUIDE−seq
3
3
12
12
96
2
7
1
BLISS
CIRCLE−seq
12h
INDUCE−seq
103
2
5
5
1
BLISS
4
1
4
GUIDE−seq
CIRCLE−seq
24h
GUIDE−seq
INDUCE−seq
5
6
1
3
106
3
2
1
BLISS
CIRCLE−seq
30h
GUIDE−seq
6
3
6
1
106
2
BLISS
INDUCE−seq
CIRCLE−seq
C
Time course r1 0h-30h
INDUCE−seq
14
7
BLISS
1
3
9
94
3
GUIDE−seq
CIRCLE−seq
D
Time course r2 0h-30h
17
1
17
91
8
BLISS
1
4
2
GUIDE−seq
INDUCE−seq
CIRCLE−seq
E
Time course r1&r2 0h-30h
25
21
1
9
BLISS
2
4
INDUCE−seq
83
GUIDE−seq
CIRCLE−seq

Figure 5.12. Euler diagrams showing intersection of the off-targets identified by INDUCE-seq, CIRCLE-seq, GUIDE-seq and BLISS. (**A and B**) Overlaps calculated for samples 0h to 30h in isolation from the independent experiments r1 (**A**) and r2 (**B**). (**C and D**) The combined overlaps from all time points for set r1 (**C**) and r2 (**D**). (**E**) Overlaps calculated between methods when all INDUCE-seq samples are combined.

5.2.8 Characterising the mismatches at EMX1 off-targets

Although the off-target mismatch plot (**Figure 5.11A**) in the previous section gave an overview of the general mismatch profile at EMX1 off-targets, quantitative analysis is difficult with the data presented in that format. For example, while it is clear that more mismatches appear at the 5' distal region of the guide sequence compared to the seed region, the precise difference in mismatch type and frequency cannot be inferred. To remedy this, and to better understand how the mismatch position and type affect off-target cleavage, the aggregate profile of off-target mismatching nucleotides was plotted for time course r1 and r2 as a frequency at each position of the guide sequence (**Figure 5.13**). The individual mismatch frequency profiles for each separated sample time point are shown in the **Appendix D, Figure A3**. For reference, **Figure 5.13A** shows the structure of the EMX1 target sequence including the position of each nucleotide, the 20 bp target site, the 12 bp seed region, and the PAM.

At first glance the off-target mismatch profiles for time course r1 (**Figure 5.13B**) and r2 (**Figure 5.13C**) are remarkably similar, showing the same general trend of mismatch frequency and types across the 23 positions. This is despite the fact that the off-targets identified within the time course replicates r1 and r2 do not fully overlap: 38.2% of the off-target sites were identified exclusively in the r1 set and 55.3% were identified exclusively in the r2 set. The similar mismatch profile suggests that although the same sites are not identified in each of the replicate sets, they represent the same class of off-target sequences based on mismatch position and frequency. The two most frequent mismatches that were observed were a T at position 5, and an A at position 6, each comprising around 26% of the total mismatch dataset, as seen in **Figure 5.13A**. These findings correspond directly to the off-target with the highest frequency, OT1, which makes up the majority of the off-target break dataset. Interestingly, the next most frequent mismatching position is within the PAM at position 22, accounting for nearly 8% of the total mismatches. Outside of the PAM, in general, most of the mismatches were observed within the distal region (defined as positions 1-8) of the protospacer target sequence. One notable exception is at position 8, which possesses a mismatch frequency that is lower than several of the seed region mismatches (positions 12, 18, and 19). This is despite the fact that seed region mismatches were restricted to ≤3 as part of the off-target discovery filter conditions used. This suggests that the strict definitions of distal and seed regions may not be appropriate for every guide RNA in the context of off-target discovery.

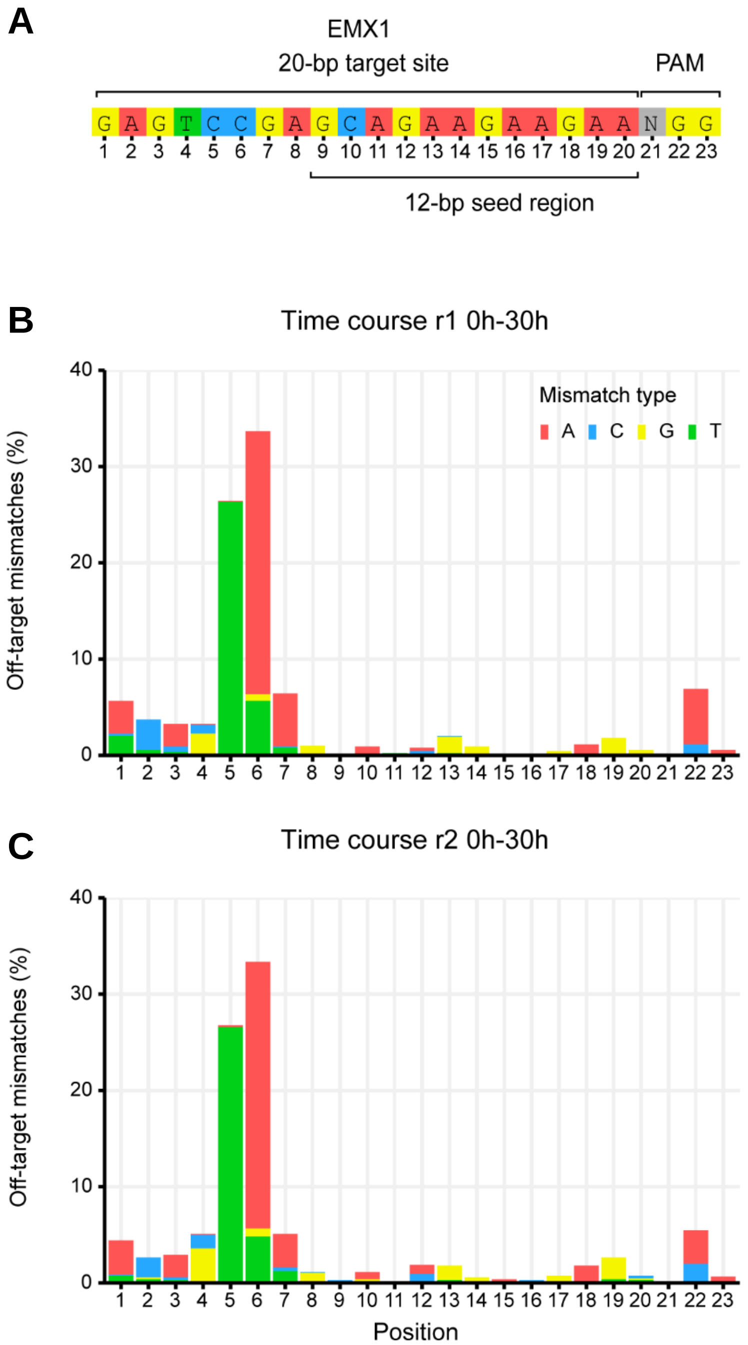


Figure 5.13. The mismatch profile at EMX1 off-targets. (**A**) Structure of the EMX1 target sequence showing the target site, seed region, and PAM. (**B**) and (**C**) Frequency of mismatches found at off-target sites across the EMX1 target sequence for time course r1 (**B**) and r2 (**C**).

While useful as a means of evaluating mismatch type and position, these profiles are greatly skewed in favour of more frequent off-targets. In particular, OT1 contributes most significantly to the mismatch distribution, which is demonstrated by the highly frequent mismatches at position 5 and 6. Because of this, it is difficult to assess the full variety of mismatches at different off-target sites, irrespective of break number. Furthermore, it also makes low frequency sites difficult to analyse as the mismatch frequency of each position becomes insignificant. To better understand the distribution of mismatches at off-targets, the mismatches were categorised based on position, and plotted against the *break frequency* across all off-target sites (**Figure 5.14A)**. In addition, the *off-target site frequency,* which is calculated independent of break number, was also plotted (**Figure 5.14B**)**.** For simplicity, the pooled full r1 and r2 time course were used for figure generation. However, the full spectrum of separated time points is shown in **Appendix D, Figures A4-A7**. When looking at the full target sequence (**Figure 5.14, pink bars)**, OT1, the only off-target with 2 mismatches, can be easily distinguished in this context. OT1 comprised >60% of the *breaks* identified at off-target sites in each time course (**Figure 5.14A, pink bars, 2 mismatches**), but only accounted for 1/34 *sites* identified in time course r1, and 1/47 sites in r2, averaging a site frequency of just 2.5% (**Figure 5.14B, pink bars, 2 mismatches**). In this manner, the relationship between the *break frequency* at off-targets, off-target *site frequency*, and mismatch number can be assessed for each of the regions of the EMX1 target sequence. These INDUCE-seq-derived off-target mismatch profiles provide a unique opportunity for the quantitative assessment of guide RNA accuracy, and could also enable improved guide RNA design in the future. A full description of the characteristics of the EMX1 off-target mismatch profile, in addition to the analysis performed on each separated time point sample, is shown in **Appendix D**.

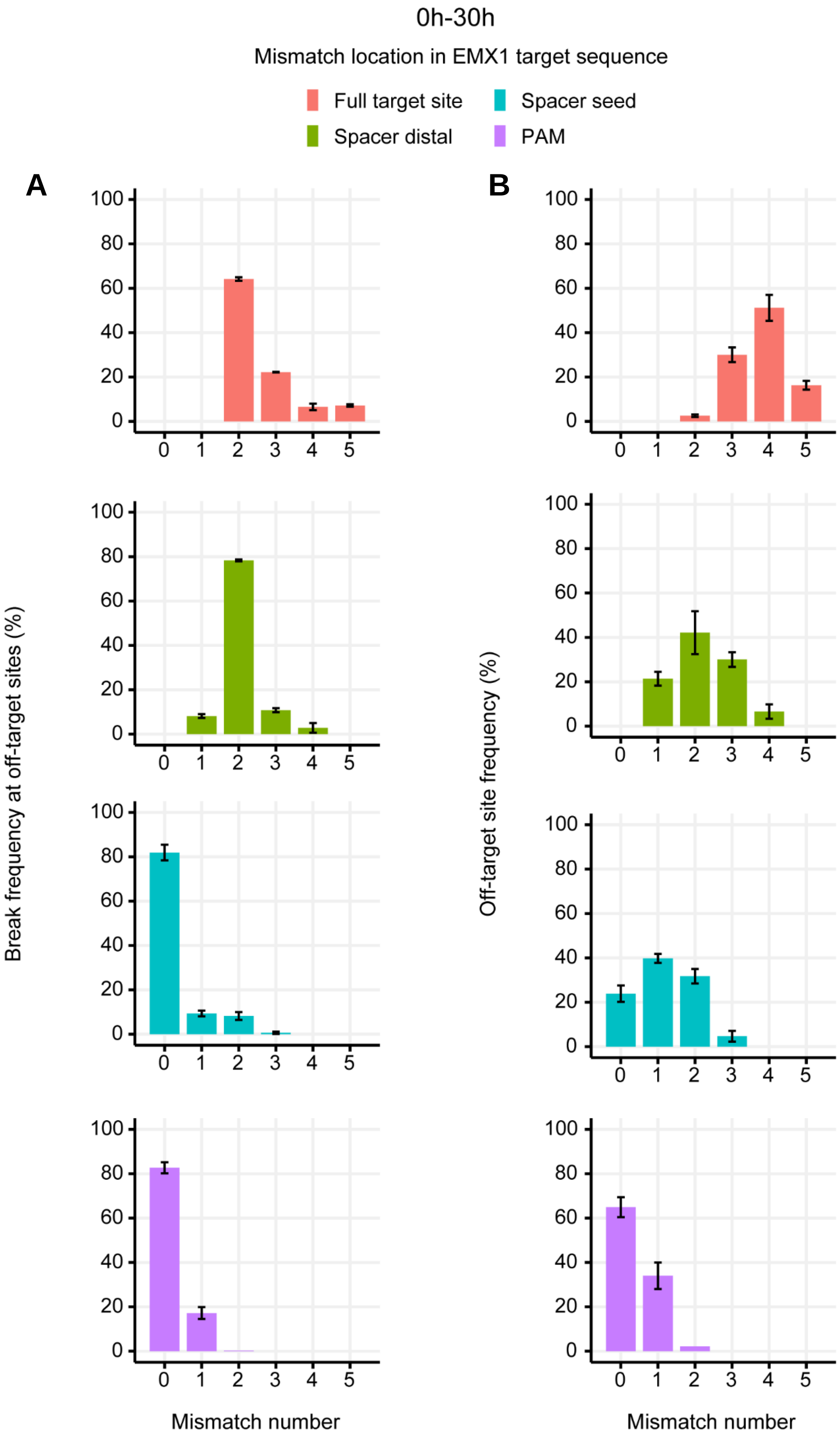

0h-30h
Mismatch location in EMX1 target sequence
🟥 Full target site
🟦 Spacer seed
🟩 Spacer distal
🟪 PAM
A
B
100
80
60
40
20
0
0 1 2 3 4 5
Break frequency at off-target sites (%)
Off-target site frequency (%)
Mismatch number
Mismatch number

Figure 5.14. Quantification and classification of at EMX1 off-targets mismatches. Across the off-target sites mismatch position is further sub classified based on the location within the target sequence. For each of the regions of the off-target sites, the break frequency at off-target sites (**A**), and the site frequency (**B**), are shown for each mismatch number. Error bars as SD, n=2.

5.3 Discussion

In this chapter, the use of INDUCE-seq for the detection of CRISPR-Cas9 induced on- and off-target DSBs in the genome has been demonstrated. These experiments revealed several similarities for CRISPR genome editing to that observed in the previous chapter for AsiSI activity in DIvA cells. First, following cell treatment of the Cas9 RNP complex, no significant difference in DSB number was observed between the treated and control samples, confirming that cellular nuclease treatment does not affect general genome stability by causing a gross increase in DSB formation. Furthermore, the level of DSB formation observed at the EMX1 on-target site at 7 hours (~250 breaks), was in line with that observed in DIvA cells at AsiSI sites following 4 hours of induction (≤125 breaks), confirming that highly efficient enzymatic cleavage was occurring within the CRISPR treated cells.

Interestingly, cellular nucleofection appeared to result in greater levels of DSB formation across both the treated and mock-nucleofected control samples. This suggests that the process itself may cause elevated DNA damage. During INDUCE-seq sample preparation, cell viability was not used as a selection criterion because the entire sample was used for the DSB capture procedure. This is because measuring only live cells when examining CRISPR off-targets may bias towards surviving cells, which would underestimate the risk associated with the guide RNA being tested. Given that nucleofection is known to result in significant cell death (Sherba et al. 2020), there is the potential that the increased break number observed from the 7h and 12h samples was due to labelling DSBs in a dead cell sub-population.

INDUCE-seq measures CRISPR-Cas9 activity at single nucleotide resolution, revealing the quantity and structure of CRISPR-Cas9 induced breaks, as shown in previous chapters for the restriction endonucleases HindIII and AsiSI. For the EMX1 target site, INDUCE-seq revealed a non-canonical 1 bp overhanging DSB rather than a blunt-ended DSB that is expected for Cas9 mediated DSB formation as observed in previous *in vitro* and genome-wide studies (Shou et al. 2018; Wienert et al. 2019). This confirms that Cas9 is able to cleave DSBs with an alternative structure, which has implications for repair of the induced DSB and the resulting indel editing outcome. Furthermore, given that an overhang is observed at the on-target, there is also the possibility that the same occurs at off-target sites. In the following chapter, the relationship between induced DSB structure and the editing outcome is explored further at both the on- and off-target sties.

In order to reveal the profile of off-target DSBs induced by the EMX1 guide RNA, a novel off-target discovery pipeline was developed that would best utilise the data output produced by INDUCE-seq. This pipeline was based around calculating the overlaps between DSBs measured by INDUCE-seq and off-target sites that were predicted using Cas-OFFinder, a homology search algorithm (Bae et al. 2014). This approach uses mismatches to filter off-target sites, and is highly efficient for calling high frequency off-target breaks, which occur well above the background rate of break formation. However, this strategy becomes less effective when interrogating low frequency sites due to the high number of endogenous break sites (noise) that are unrelated to CRISPR editing. To address this, a range of filtering parameters were calculated and compared to determine the most appropriate for revealing off-targets breaks in the treated samples, with minimal false-positive off-target breaks in the control samples. An approach using the false discovery rate and off-target yield was used to define the best parameters for filtering. This differs from the standard method of plotting a receiver operating characteristic curve (ROC curve), which compares the true positive rate (sensitivity) and false positive rate of a test at various thresholds of stringency. The false discovery rate method was chosen because it is not possible to determine false negative off-targets without using data generated by orthogonal approaches, or subsequent indel analysis as a reference, thus making the sensitivity calculation impossible. Comparing the false discovery rate and off-target yield results in robust off-target discovery that can (i) identify new off-targets that have not yet been measured by other approaches and (ii) performs independently without the need to confirm the observations by another method (Tsai et al. 2015; Tsai et al. 2017; Yan et al. 2017).

Quantification of the subsequent off-targets revealed that that relative amount of editing, as measured by the total number of breaks identified at off-targets, and the number of off-target sites identified, was reproducible across replicates of the same time points. Notably, the greatest number of off-targets were identified at the early stages of the time course, in particular at 7h, suggesting that Cas9 activity peaks rapidly following nucleofection, reducing by 30h. Reproducibility was good when comparing the full set of off-target sites identified across the r1 and r2 replicate time courses, with 21 sites shared between the datasets. Significantly, when editing activity was measured as the number of breaks identified per off-target site, a very strong linear correlation was observed ($R^2 = 0.9965$) between experiments. This observation confirmed that the relative amount of break induction by Cas9/EMX1 at on- and off-targets was consistent between independent experiments.

Given these findings, all samples from the r1 and r2 replicate time courses were pooled to provide a single larger set of off-targets for subsequent mismatch analysis and comparison with other methods. A total of 60 off-targets were identified for EMX1 across all replicate samples, ranging from the highest frequency off target (OT1), with over 400 breaks, to many sites represented by just a single break. In agreement with previous studies, off-target mismatches ranged from 2 to 6 bp from the target site with the majority observed at the distal region of the spacer (Hsu et al. 2013; Tsai et al. 2015). Several PAM mismatches were also observed, which have previously been reported, including NAG and NGA. Off-targets with the PAM sequences of NCG and NTG were also observed, highlighting a variety of PAM-interactable sequences that can occur at off-targets. Interestingly, these less common PAM mismatches have not previously been observed using the other cell-based methods GUIDE-seq, BLISS and HTGTS (Frock et al. 2015; Tsai et al. 2015; Yan et al. 2017), but were identified by *in vitro* approaches CIRCLE-seq and Digenome-seq (Kim et al. 2015; Tsai et al. 2017). Furthermore, when looking at the total number of off-targets identified, INDUCE-seq demonstrates remarkable sensitivity compared to the other cell-based approaches. INDUCE-seq identified many off-target sites in live cells that have only been observed previously using *in vitro* approaches. Importantly, INDUCE-seq also discovered several off-target sites that had not been identified previously by any other method. These findings are significant as they demonstrate that *in vitro* approaches can miss genuine CRISPR off-targets (false negatives). It is recognised that *in vitro* approaches in general significantly overestimate the number of off-targets for a given sgRNA (Wienert et al. 2019). This is likely because of the vastly different conditions of isolated genomic DNA compared to the dynamic chromatin environment in living cells. These findings suggest that *in vitro* approaches are not sufficiently accurate and are therefore of limited value for the safety profiling of CRISPR guides. Consequently, future safety profiling should be performed with highly sensitive cell-based methodologies such as INDUCE-seq.

Analysis of the frequency of mismatches at off-target sites revealed that generally most mismatches were present at the distal region of the spacer, and few mismatches were present in the seed region. It is worth noting that this analysis will be impacted by the upstream filtering steps used for off-target discovery, as the seed region was limited in mismatch number (≤3) for filter condition 7. In future, it would be worth comparing the off-target profile from a dataset of sites generated without any restriction on the seed region to confirm whether the pattern observed was caused by CRISPR biology or was affected by upstream filtering steps. Given that no difference in off-target number was observed between filter condition 7 using ≤3 seed mismatches and condition 15 using ≤5 seed mismatches, however, it seems very unlikely that over-filtering alone is accounting for the fewer mismatches observed in the seed region.

Aside from the mismatches occurring as result of the highest frequency off-target OT1, surprisingly, the second nucleotide in the PAM (position 22) was the highest frequency mismatching position. Furthermore, of all the off-target breaks identified outside of OT1, ~50% of the breaks occur at sites with a PAM mismatch. These findings highlight the importance of PAM mismatches for off-target site selection, which are often overlooked when designing the spacer element of CRISPR guides. Modification of the PAM-interacting domain for greater stringency could therefore be an effective strategy for reducing off-target editing.

Finally, the results in this chapter show, for the first time, quantitative mapping of CRISPR off-target breaks in the genome. The data generated by INDUCE-seq reveals both the position and frequency of off-targets without distortion, enabling the accurate analysis of the kinetics of CRISPR break induction, and revealing the profile of mismatches occurring at off-target sites. Using a single well-characterised guide RNA, EMX1, described in this chapter, provides a paradigm that can be applied as a basis for the rational design and safety profiling of other guide RNA and/or CRISPR proteins during the development of cell and gene therapy. In the next chapter, the pattern of CRISPR break induction measured by INDUCE-seq is further examined and compared to mutagenic outcome of the editing process.

6 Chapter VI - Characterising the repair outcome at CRISPR induced on- and off-target sites

6.1 Introduction

In the previous chapter, the genome-wide profile of off-targets induced by the guide RNA targeting EMX1 were determined using INDUCE-seq. As discussed in section **1.6**, off-target discovery is the important first step to test for CRISPR-associated genotoxic outcome and oncogenesis during the development cell and gene therapies. Following off-target site discovery, the next step involves the assessment of the mutational consequences of on- and off-target genome editing. Analysis of the position and frequency of CRISPR-induced mutagenic outcomes is particularly important given the systemic, non-random activity of programmable nucleases. One threshold that has been suggested to define off-target edits as safe, is if they fall below the background mutation rate of ~1.6 x 10^{-8} in dividing cells (Cheng and Tsai 2018). This figure, however, does not take in to account the recurrent nature of CRISPR activity; RNA-guided DSB induction is targeted, even at low frequencies. During cell-based experiments any given off-target may appear below or at the frequency of spontaneous mutations, but prolonged and variable exposure during a cellular and/or patient treatment regimen could lead to the accumulation of off-target-derived mutations. Furthermore, as the location of off-target mutations are non-random, some sites in the genome should be considered more dangerous that others when risk is being assessed. Mutations at off-target sites that inhibit tumour suppressor genes or activate proto-oncogenes could have extremely deleterious consequences even at frequencies well below the background mutation rate. Correspondingly, it is conceivable that mutations at off-targets in non-coding regions of the genome could be considered safe even at higher frequencies. Finally, in patient populations, individual single nucleotide polymorphisms (SNP) can result in off-target activity at novel sites, resulting in unique, patient-specific editing outcomes (Fellows 2016). A nuanced risk-benefit analysis is therefore required for the safe and sustainable development of any CRISPR-based cell and gene therapy.

Several different technologies now exist for measuring off-target editing outcomes, with the most popular approaches including the mismatch cleavage T7E1 assay, sanger sequencing, followed by sequence trace decomposition analysis (TIDE/ICE analysis), and whole genome/ targeted NGS (Bennett et al. 2020). Deep sequencing at >1000x coverage of amplicons by NGS, known as amplicon sequencing or amplicon-seq, has become the gold-standard approach for off-target mutation detection due to the quantity of information, sensitivity, and accuracy provided by a single measurement. Amplicon-seq can detect off-target mutations down to ~0.1% frequency, in addition to revealing the full spectrum of insertion, deletion, and substitution mutation events. This provides valuable information about the risks of genome editing. For example, at off-targets in actively transcribing genes, frame-shift mutations and premature nonsense mutations can result in loss of gene function. Both of these can be determined from the indel sizes and substitution types produced following editing. Furthermore, several groups have used the indel profile at on-target sites to infer the relative contributions of c-NHEJ and MMEJ during DSB repair at different guide RNA target sequences. This aiding in the prediction of editing outcomes for untested guide RNAs (van Overbeek et al. 2016; Shen et al. 2018). Indeed, the accurate prediction of indel formation at both on- and off-target sites in different cell types and genetic backgrounds would greatly reduce the variation of outcome of DSB-mediated gene knockouts, a major limitation of classical genome editing via targeted nucleases such as Cas9. It is evident that determining how the guide RNA target sequence, chromatin context, and characteristics of DSB formation, affect the editing outcome at off-target sites, would significantly enhance the robust risk assessment of therapeutic CRISPR guide RNAs.

6.1.1 Chapter aims

Given the importance of understanding the relationship between off-target break formation and editing outcome, the work in this chapter aims to characterise the mutagenic outcomes at the on-target and off-targets measured for EMX1 by INDUCE-seq.

6.2 Results

6.2.1 Measuring the indel editing outcome at off-targets identified by INDUCE-seq

To measure the on- and off-target editing outcome following treatment with the EMX1 guide RNA, simultaneous deep sequencing of all off-target sites identified by INDUCE-seq was performed using amplicon-seq (**2.1.10**). A custom target enrichment panel was designed for this purpose, directly informed by the findings described in the previous chapter. This enabled the parallel amplification of PCR-products corresponding to the 60 off-target sites identified throughout each sample from the r1 and r2 replicate time courses. Deep sequencing to >1000x coverage of each amplicon, enables the quantification of the overall frequency of edited versus unedited cells and reveals the precise proportion of each indel editing outcome at the 60 off-target sites.

To determine the relationship between break induction and the frequency of indel editing outcome, the average indel frequency identified at each off-target for replicates r1 and r2 (**Figure 6.1C**) was first plotted against the off-target mismatch and break profiles shown in the previous chapter (**Figure 6.1A and B**) (**Appendix D, Table A5**). Strikingly, of the 60 total off-target sites that were identified by INDUCE-seq, only four were identified with indels when measured by amplicon-seq. These findings mirror that observed previously by Vakulskas *et al.* 2019, who demonstrated that following EMX1 RNP nucleofection into HEK293 cells, only six off-targets were identified by amplicon-seq at 48 hours following cell treatment. Of the six, the frequencies of five that were also identified by INDUCE-seq are shown (**Figure 6.1C, far right**), which demonstrates a very similar indel frequency to that identified at 30 hours. Given the known detection limit of amplicon-seq for mutation detection of ~0.1% (i.e. 1 in 1000), this shows that INDUCE-seq is substantially more sensitive at detecting off-target DSBs than amplicon-seq is for detecting CRISPR off-target editing outcome. These findings demonstrate that many genuine off-targets likely fall well-below the detection limit of amplicon-seq.

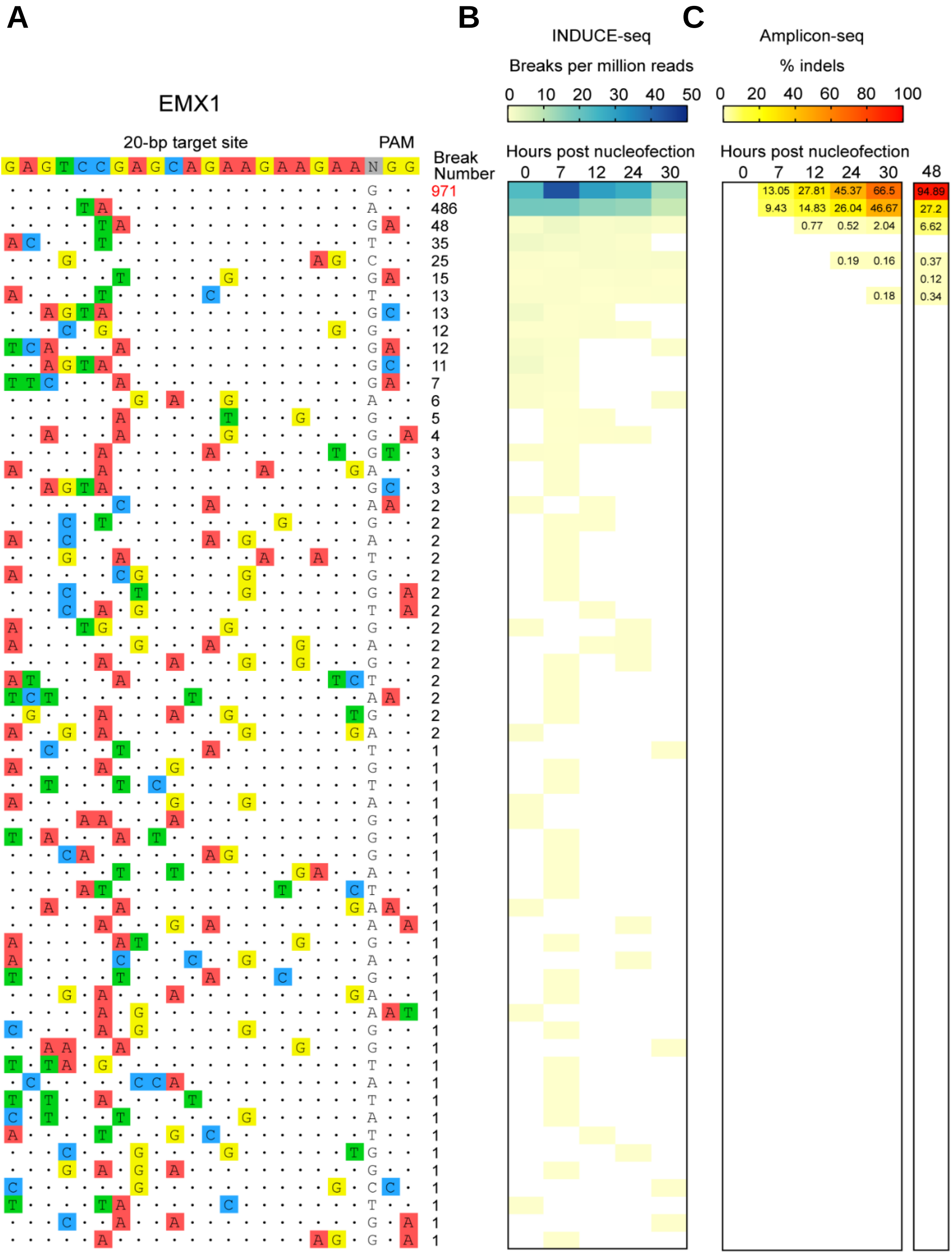

A
B
C
EMX1
20-bp target site
PAM
G A G T C C G A G C A G A A G A A G A A N G G
Break Number
971
486
48
35
25
15
13
13
12
12
11
7
6
5
4
3
3
3
2
2
2
2
2
2
2
2
2
2
2
2
2
1
INDUCE-seq
Breaks per million reads
0 10 20 30 40 50
Hours post nucleofection
0 7 12 24 30
Amplicon-seq
% indels
0 20 40 60 80 100
Hours post nucleofection
0 7 12 24 30
48
13.05 27.81 45.37 66.5
94.89
9.43 14.83 26.04 46.67
27.2
0.77 0.52 2.04
6.62
0.19 0.16
0.37
0.12
0.18
0.34

Figure 6.1. Comparison of the frequency of break induction and indel editing outcome at EMX1 off-targets identified by INDUCE-seq. (**A**) Off-target sequences and the number of breaks identified using INDUCE-seq for the EMX1 sgRNA. (**B**) Normalised break number identified at each off-target for each time point sample. (**C**) The indel frequency measured by targeted amplicon sequencing for each of the off-targets each time point sample. Amplicon sequencing identifies only 4 of the 60 off-targets discovered using INDUCE-seq and is limited by the background indel false-discovery rate of 0.1%. (**C, far-right**) Indel frequency reported previously for EMX1 48 hours post RNP nucleofection (Vakulskas *et al.* 2019).

6.2.2 Comparing the kinetics of editing at on- and off-target sites

The kinetics of editing was further characterised by comparing the profiles of break induction and the accumulation of indels throughout time. For this analysis, the on-target and two most frequent off-targets (OT1 and OT2) were selected, as they were the only sites with detectable indels at ≥3 time points of the series. **Figure 6.2A and B** show the normalised break number identified at each of the sites from 0 to 30 hours, which confirms the previous observation that break on-target activity is greatest immediately following cell treatment (0-7 hours), and declines towards 30 hours. Interestingly, the same peak in activity was not observed for OT1 or OT2, which show more constant activity with a slower decline. Next, the cumulative number of breaks at each site was compared to understand how breaks accumulated throughout the time course (**Figure 6.2C and D**). The trend for all three sites show a similar pattern, with the steepest rise in break number earlier in the time course (0h-12h). This then plateaus between the 12- and 30-hour samples. The accumulation of indels appears to follow a similar trend to the accumulation of breaks (**Figure 6.2E and F**). Indels are undetectable at 0 hours and accumulate throughout the time course, peaking at 30 hours. The difference in indel frequency and breaks accumulated at 0h is expected and can be directly attributed to requirement of DNA repair to form indels following DSB induction by Cas9. Apart from this, the relative proportion of break accumulation and indel frequency appears consistent for each of the sites. This suggests that break and indel accumulation may be correlated.

To determine whether the editing outcome was proportional to the break number, the breaks accumulated at each time point (r1 and r2, 10 total) were plotted against the indel frequency detected (**Figure 6.3**). Both the on-target (**Figure 6.3A**) and OT1 (**Figure 6.3B**), show strong correlation between the breaks accumulated and indel frequency ($R^2 = 0.87766$ and $R^2 = 0.77945$, respectively), confirming that the frequency of editing is generally proportional to the break activity at these sites. However, no correlation is observed for OT2 (**Figure 6.3C**) ($R^2 = 0.310123$), which contains many samples with breaks measured, but an indel frequency of 0%. Collectively, these results suggest that for sites broken actively with a high frequency indel outcome, break measurement by INDUCE-seq may be predictive of the frequency of the editing outcome.

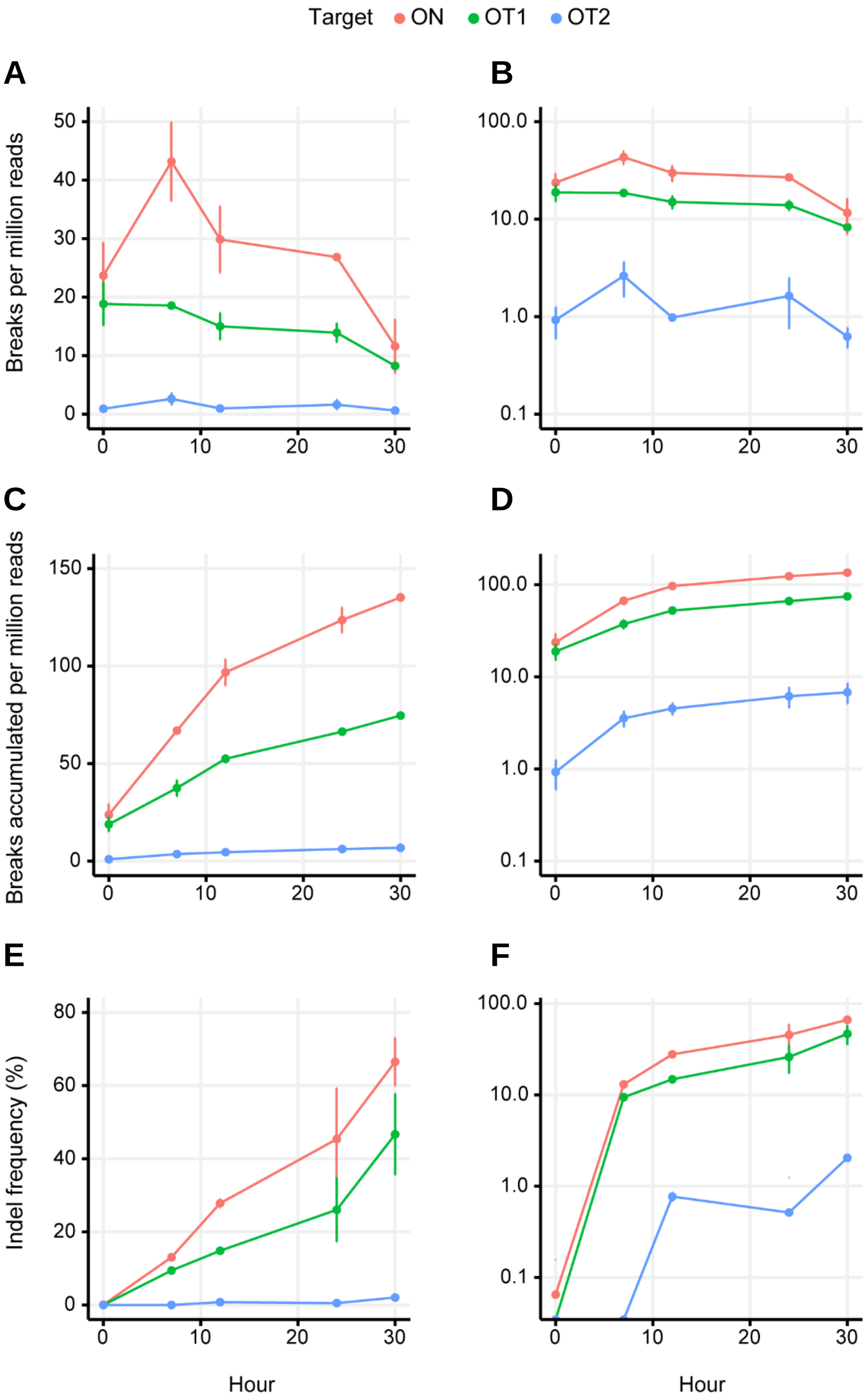


Figure 6.2. The kinetics of break formation and repair the EMX1 on-target, OT1 and OT2. (**A** and **B**) The normalised break number measured across the time points shows peak activity at 0 and 7 hours which then drops off towards 30 hours. The relative proportion of accumulated breaks (**C** and **D**) and indels (**E** and **F**) follow a similar trend and appear consistent for each of the sites.

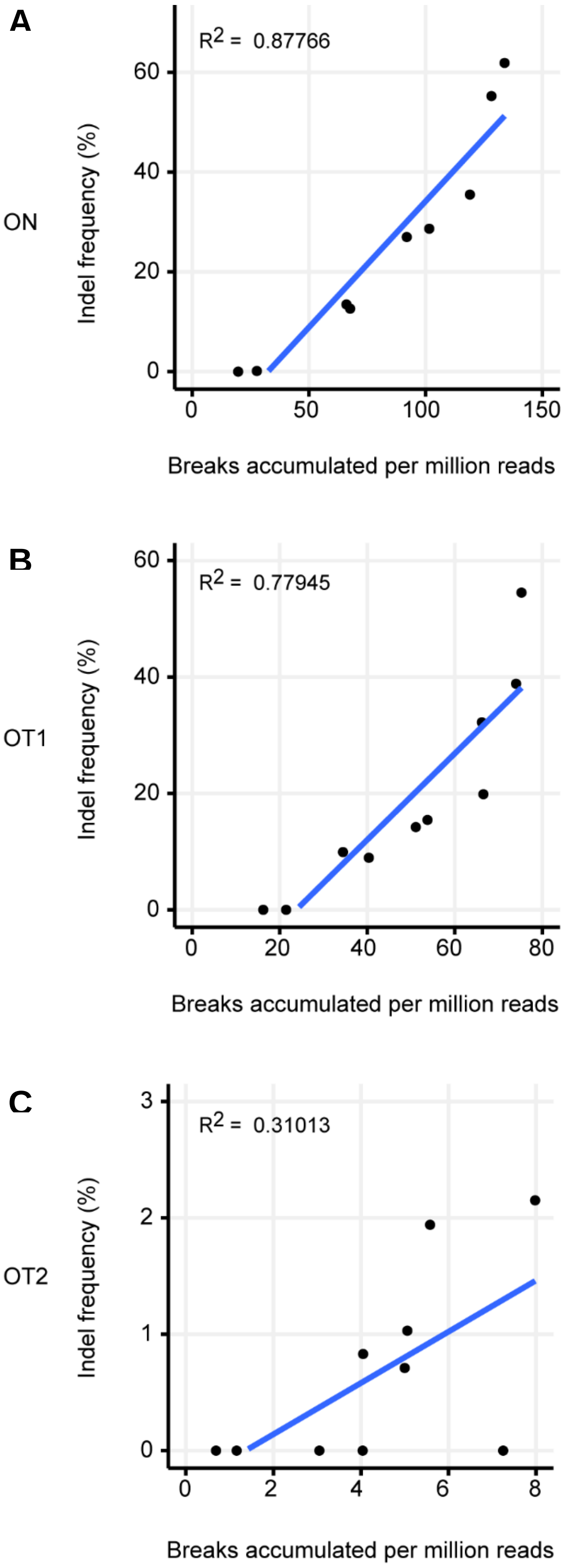

A
ON
R2 = 0.87766
Indel frequency (%)
0
20
40
60
0
50
100
150
Breaks accumulated per million reads
B
OT1
R2 = 0.77945
Indel frequency (%)
0
20
40
60
0
20
40
60
80
Breaks accumulated per million reads
C
OT2
R2 = 0.31013
Indel frequency (%)
0
1
2
3
0
2
4
6
8
Breaks accumulated per million reads

Figure 6.3. Scatterplots comparing the number of breaks accumulated and the indel frequency at the EMX1 on-target, OT1 and OT2. The number of accumulated breaks are strongly correlated with the indel frequency at the on-target (R^2 = 0.87766) (**A**) and OT1 (R^2 = 0.77945) (**B**), but show weak correlation at OT2 (R^2 = 0.31031) (**C**).

6.2.3 Characterising the indel profile at on- and off-target sites

In addition to providing a sensitive measurement of the overall editing frequency, the main benefit of performing amplicon-seq is that it produces a detailed readout of the full spectrum of insertion and deletion events present at a CRISPR-edited sites. From this information, it is possible to characterise the diversity of editing outcomes of a given guide RNA, which informs on the mechanism of repair leading to the editing outcome. **Figure 6.4** shows a representative indel readout of the 20 most frequent amplicons detected at the EMX1 target site at 30h, revealing the variety of indel sizes, ranging from a single nucleotide insertion to >10 bp deletions. The frequency of the different indels also varied dramatically: the most common indel outcome at the EMX1 on-target is a 1 bp insertion, representing 37.24% of the reads sequenced. This is almost 4-fold higher than the next highest indel outcome (6 bp deletion, 10.66%). Outside of the most frequent editing outcomes, (including unedited reads), the remaining 15 indel products, which included many larger deletions, occurred at frequencies lower than 1%.

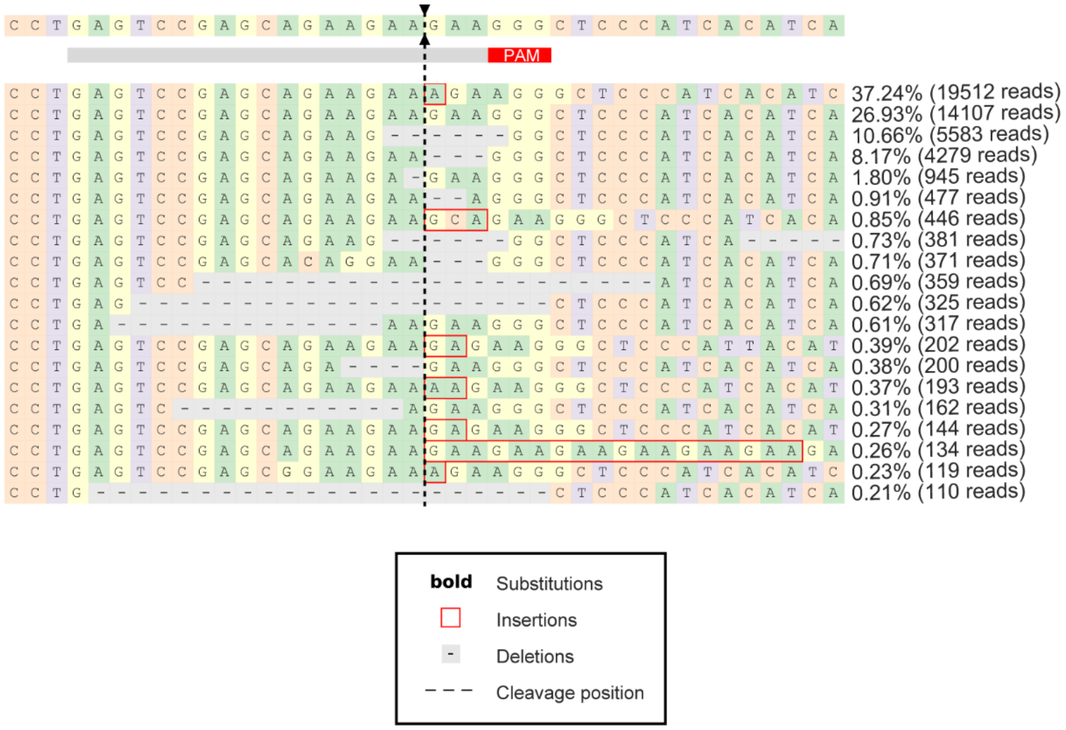


Figure 6.4. Representative indel readout at the EMX1 target site. The position of the guide RNA target protospacer (grey bar) and PAM (red bar) are shown in relation to the cleavage position (vertical dashed line). The number of each sequencing read generated by amplicon sequencing shown and the positions of insertions and deletions are highlighted.

These findings demonstrate the strong preference for formation of a 1 bp insertion at the EMX1 on-target site, which appears to correlate with the cleavage profile shown in the previous chapter. **Figure 6.5** shows the comparison between the on-target cleavage profile, measured by INDUCE-seq at 7h, and the indel profile measured at 30h using amplicon-seq. As shown previously, the INDUCE-seq sequencing read coverage at the on-target site at 7h reveals that the majority of forward (fwd) and reverse (rev) reads overlap by 1 bp at the break site (**Figure 6.5A**). This represents a staggered 1 bp overhanging break end, where the forward reads are positioned 1bp upstream of the expected break position, 3 bp upstream of the PAM (**Figure 6.5B, INDUCE-seq**). Furthermore, the most common indel outcome measured by amplicon-seq is a 1 bp insertion of an adenosine (**Figure 6.5B, amplicon-seq**), which matches the adenosine present at the 1 bp overhanging break site. The same phenomenon is also observed at the two most frequent off-targets, OT1 (**Figure 6.6**) and OT2 (**Figure 6.7**). In both cases, the same 1 bp overlapping break pattern is observed at the break site, shown by the grey coverage tracks and the INDUCE-seq forward and reverse read positioning. Since the EMX1 off-target sequence is present on the minus strand for OT1, but is present on the plus strand at the on-target and OT2, the same pattern is observed but with inverted read orientations. Similar to the on-target, the 1 bp adenosine insertion is the most common indel outcome for both OT1 and OT2, at 18.75% and 1.07% of indels, respectively. These findings suggest that the structure of CRISPR-induced breaks may contribute directly to the type and frequency of the indel editing outcomes.

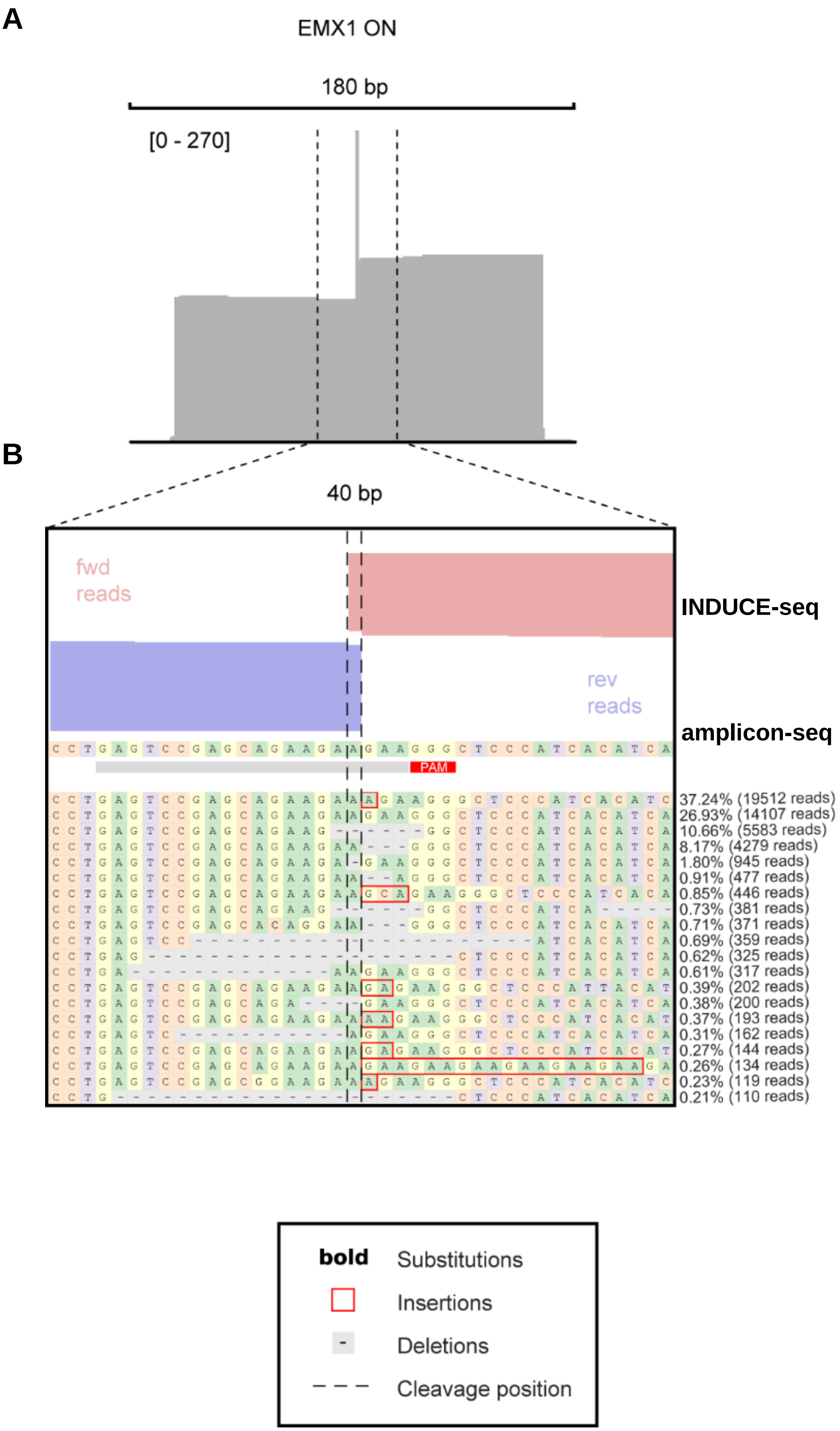

A
EMX1 ON
180 bp
[0 - 270]
B
40 bp
fwd reads
rev reads
INDUCE-seq
amplicon-seq
CCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCATCACATCA
PAM
37.24% (19512 reads)
26.93% (14107 reads)
10.66% (5583 reads)
8.17% (4279 reads)
1.80% (945 reads)
0.91% (477 reads)
0.85% (446 reads)
0.73% (381 reads)
0.71% (371 reads)
0.69% (359 reads)
0.62% (325 reads)
0.61% (317 reads)
0.39% (202 reads)
0.38% (200 reads)
0.37% (193 reads)
0.31% (162 reads)
0.27% (144 reads)
0.26% (134 reads)
0.23% (119 reads)
0.21% (110 reads)
bold Substitutions
Insertions
- Deletions
Cleavage position

Figure 6.5. The INDUCE-seq detected DSB pattern at the EMX1 on-target relates to editing outcome. Coverage track of the EMX1 on-target (**A**), spanning 180bp. (**B**) At a close-up view of the 40bp region surrounding the site, INDUCE-seq shows a distinct 1bp overhanging cleavage pattern rather than the usual Cas9-induced blunt DSB. Corresponding indel spectra, as measured by amplicon sequencing shows the position of the indel mutations in relation to the observed break sites.

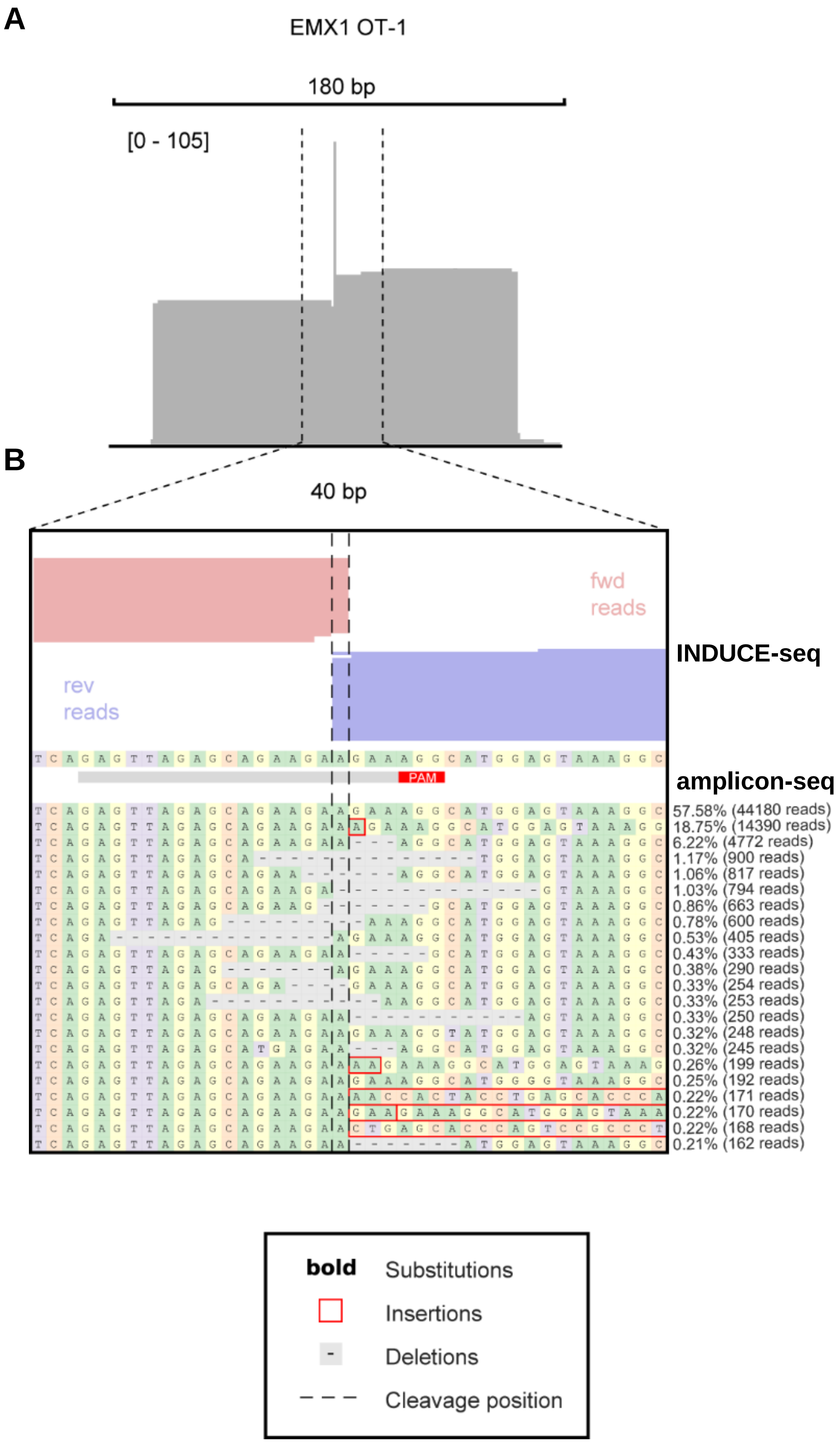

A
EMX1 OT-1
180 bp
[0 - 105]
B
40 bp
fwd reads
rev reads
INDUCE-seq
amplicon-seq
PAM
57.58% (44180 reads)
18.75% (14390 reads)
6.22% (4772 reads)
1.17% (900 reads)
1.06% (817 reads)
1.03% (794 reads)
0.86% (663 reads)
0.78% (600 reads)
0.53% (405 reads)
0.43% (333 reads)
0.38% (290 reads)
0.33% (254 reads)
0.33% (253 reads)
0.33% (250 reads)
0.32% (248 reads)
0.32% (245 reads)
0.26% (199 reads)
0.25% (192 reads)
0.22% (171 reads)
0.22% (170 reads)
0.22% (168 reads)
0.21% (162 reads)
bold Substitutions
Insertions
Deletions
Cleavage position

Figure 6.6. The INDUCE-seq detected DSB pattern at OT1 relates to editing outcome. Coverage track of OT1 (**A**), spanning 180bp. (**B**) At a close-up view of the 40bp region surrounding the site, INDUCE-seq shows a distinct 1bp overhanging cleavage pattern rather than the usual Cas9-induced blunt DSB. Corresponding indel spectra, as measured by amplicon sequencing shows the position of the indel mutations in relation to the observed break sites.

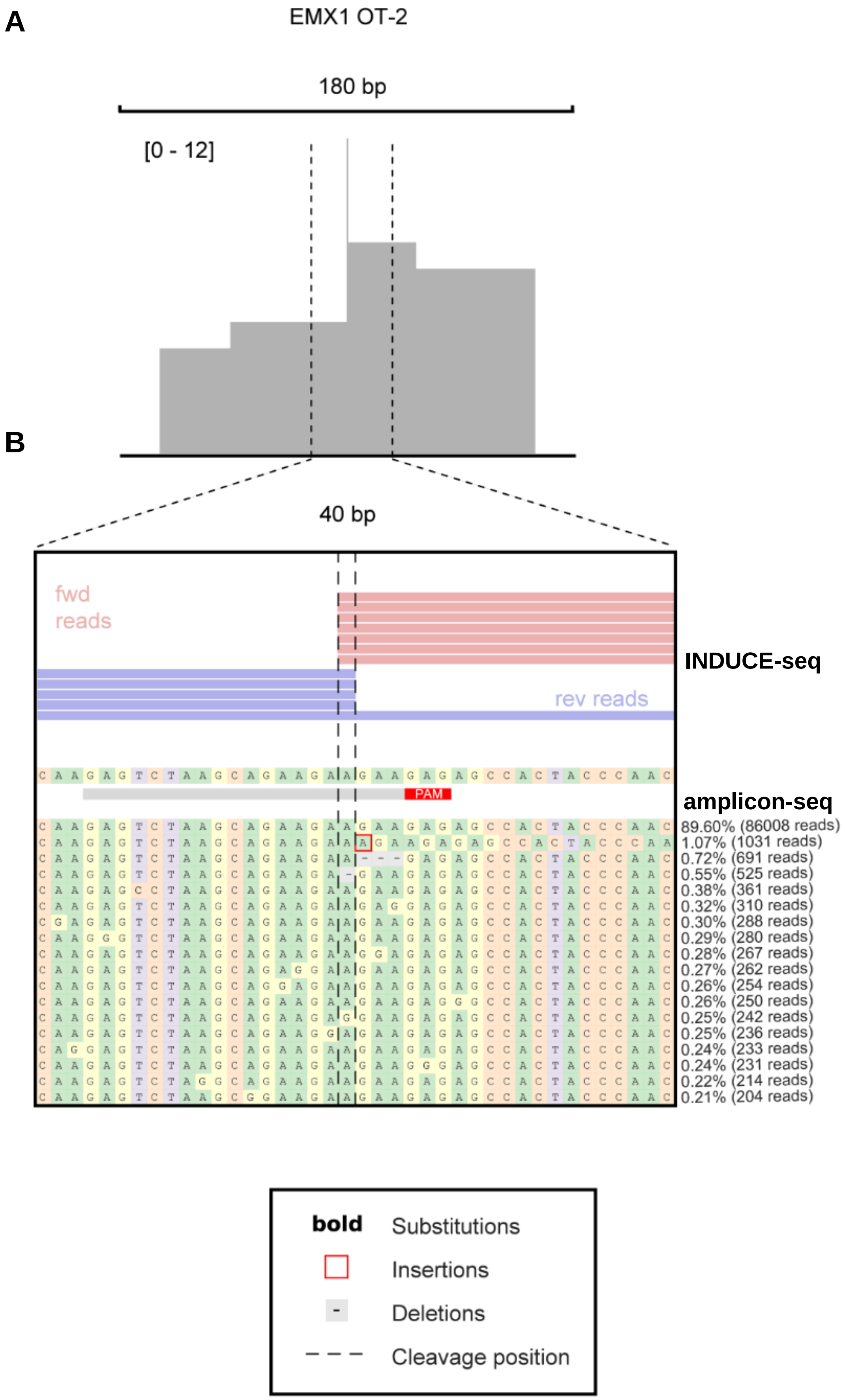

A
EMX1 OT-2
180 bp
[0 - 12]
B
40 bp
fwd reads
rev reads
INDUCE-seq
PAM
amplicon-seq
89.60% (86008 reads)
1.07% (1031 reads)
0.72% (691 reads)
0.55% (525 reads)
0.38% (361 reads)
0.32% (310 reads)
0.30% (288 reads)
0.29% (280 reads)
0.28% (267 reads)
0.27% (262 reads)
0.26% (254 reads)
0.26% (250 reads)
0.25% (242 reads)
0.25% (236 reads)
0.24% (233 reads)
0.24% (231 reads)
0.22% (214 reads)
0.21% (204 reads)
bold Substitutions
Insertions
Deletions
Cleavage position

Figure 6.7. The INDUCE-seq detected DSB pattern at OT2 relates to editing outcome. Coverage track of OT2 (**A**), spanning 180bp. (**B**) At a close-up view of the 40bp region surrounding the site, INDUCE-seq shows a distinct 1bp overhanging cleavage pattern rather than the usual Cas9-induced blunt DSB. Corresponding indel spectra, as measured by amplicon sequencing shows the position of the indel mutations in relation to the observed break sites.

6.2.1 Characterising the break structure of CRISPR-induced break sites

Having established the correlation between the structure of the induced breaks measured by INDUCE-seq and the indel editing outcome, the next step was to quantify the position of break ends around each induced break site. As discussed in the previous chapter, break ends positioned away from the cleaved site are likely to arise from DNA end-resection during DSB repair. Given that the degree of DNA end-resection is an important determinant of the DSB repair pathway utilised (Chapman et al. 2012; Ceccaldi et al. 2016), the positions of these resected break ends could provide insights into repair outcomes measured by amplicon-seq.

To calculate and quantify the resection profile around the on- and off-target sites, the position of each minus strand (reverse) read was subtracted from the position of each plus strand (forward) read within a +/- 50 bp window around the expected break site (**Figure 6.8A**). This was then expressed as a frequency. Because the first stage of break labelling blunts break ends via 5’>3’ polymerase activity and 3’>5’ exonuclease activity, a range of break structures can be envisaged using this approach. First, when the start positions of the right-side labelled (forward) and left-side labelled (reverse) reads are the same, this could represent a blunt-ended DSB (**Figure 6.8B, top**). Furthermore, when breaks labelled on the left are positioned downstream of breaks labelled on the right, a positive score is generated, which could represent an overhanging end (**Figure 6.8B, middle**). Finally, a negative score is generated when breaks labelled on the left side are positioned upstream from breaks labelled on the right, which could represent the resection distance in bp (**Figure 6.8B, bottom**).

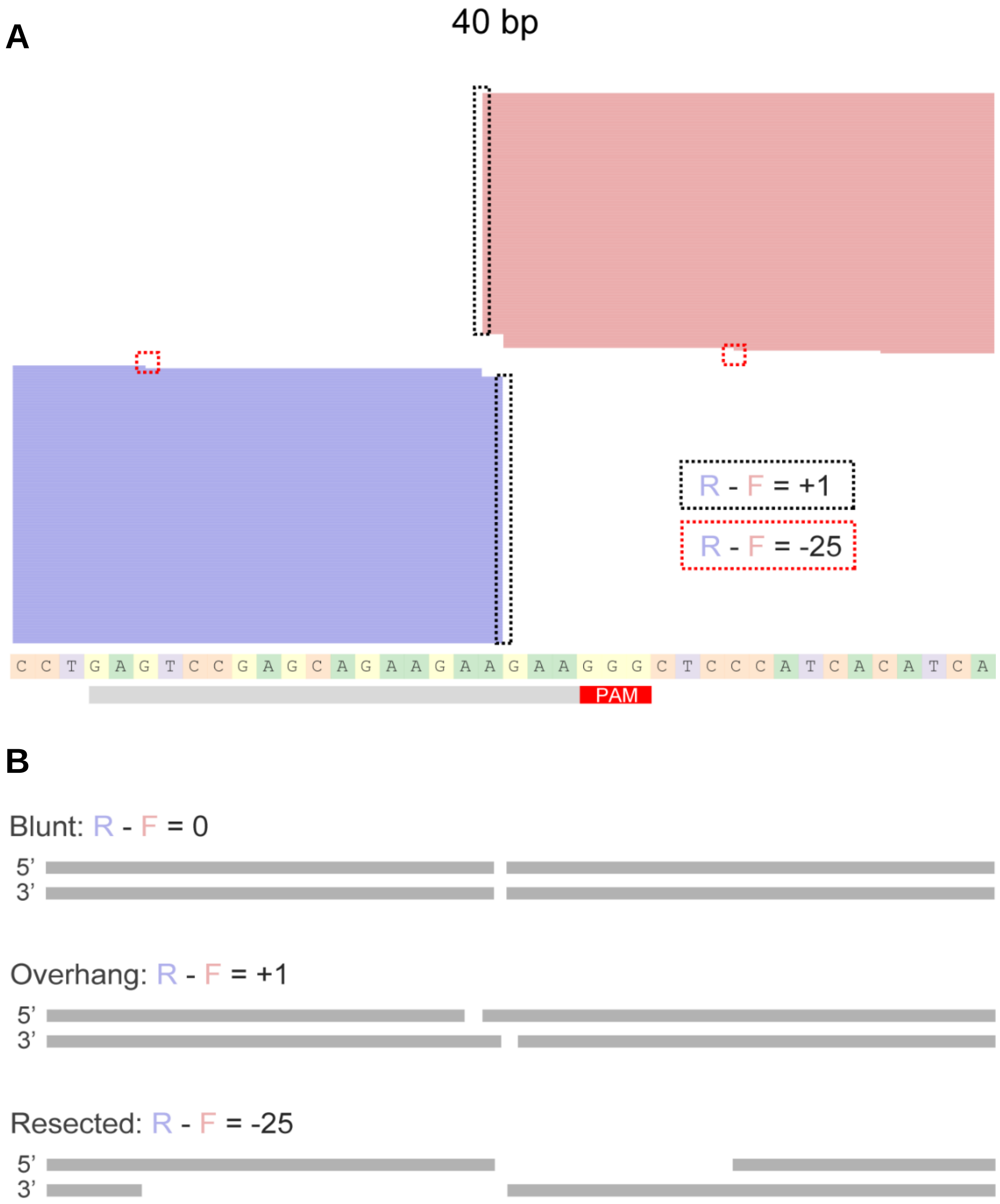


Figure 6.8. Schematic detailing the distance calculation to quantify the resection profile around CRISPR induced break sites. (**A**) Representative break structure within a 40bp window around the EMX1 on-target site the start position of all reverse reads are subtracted from the start position of forward reads to calculate the break distance. (**B**) Examples of the break structures can be envisaged using the distance calculation.

6.2.2 Comparing DSB end-resection and indel formation

Using all ten INDUCE-seq datasets (0h-30h, replicates r1 and r2), a single resection profile was generated for the EMX1 on-target and OT1 to represent the overall pattern of break resection for both sites. In a similar way, the 30h r1 and r2 replicate amplicon-seq datasets were also combined to generate a single indel frequency distribution for the on-target site and OT1. Insertions, deletions, and unedited DNA fragments are represented by positive, negative, and 0 indel size values, when viewed as a frequency distribution. **Figure 6.9**, **Figure 6.10,** and **Figure 6.11** show the comparison between the indel frequency measured at 30h and the aggregate resection profile at the EMX1 on-target, and corresponding profiles at OT1. As was demonstrated previously, the most common editing outcome at both sites is a 1 bp insertion (35.12% on target, 32.49% OT1), followed by the unedited fraction (31.36% on-target, 52.56% OT1) (**Figure 6.9A**, **Figure 6.10A**, **Figure 6.11**). Across both sites, a range of small deletions (<10bp) occur at frequencies above 1%, peaking at 8.98% for a 6bp deletion at the on-target site. Interestingly, the deletion profile differs considerably between the two sites, with smaller deletions appearing enriched at the on-target and larger deletions occurring at higher frequencies at OT1 (**Figure 6.9B**, **Figure 6.10B**, **Figure 6.11)**. Both sites also show significant levels of small insertion events that are larger than the most common 1bp insertion measured. Notably, the frequencies of these are higher at the on-target. Looking at the positioning of break ends (represented here as resection profiles), confirms that a +1bp overhang makes up the vast majority of break structures measured (~84.4% on-target, 77.3% OT1) (**Figure 6.9C**, **Figure 6.10C**, **Figure 6.11).** Correspondingly, 'blunt-ended' break represent only 8.75% (on-target) and 6.71% (OT1) of the breaks measured. This demonstrates the high prevalence of induced overhanging breaks and resected break ends at the time points measured. A greater diversity of low frequency resected ends was measured at both sites compared with the deletions in the indel profile, ranging from 0.01% to 3.08% (**Figure 6.11**). Despite the fact that many small insertions (<5bp) and deletions (<10bp) are found at the on-target, few corresponding resection products or overhangs are found of the same size. This suggests that the precise resection distance does not correlate with the editing outcome (**Figure 6.9D**, **Figure 6.11).** Interestingly, OT1 shows a greater range and variety of resection products and overhanging sequences despite the lower overall editing frequency. This indicates differences in the way the two sites are repaired following break induction (**Figure 6.10D)**.

Overall, these results suggest that break structure, while contributing significantly to certain repair outcomes, does not directly correlate with the indel distribution at each site.

Instead, end-resection and break structure are likely to be one of many factors that determine the mutagenic editing outcome following genome editing using CRISPR-Cas9.

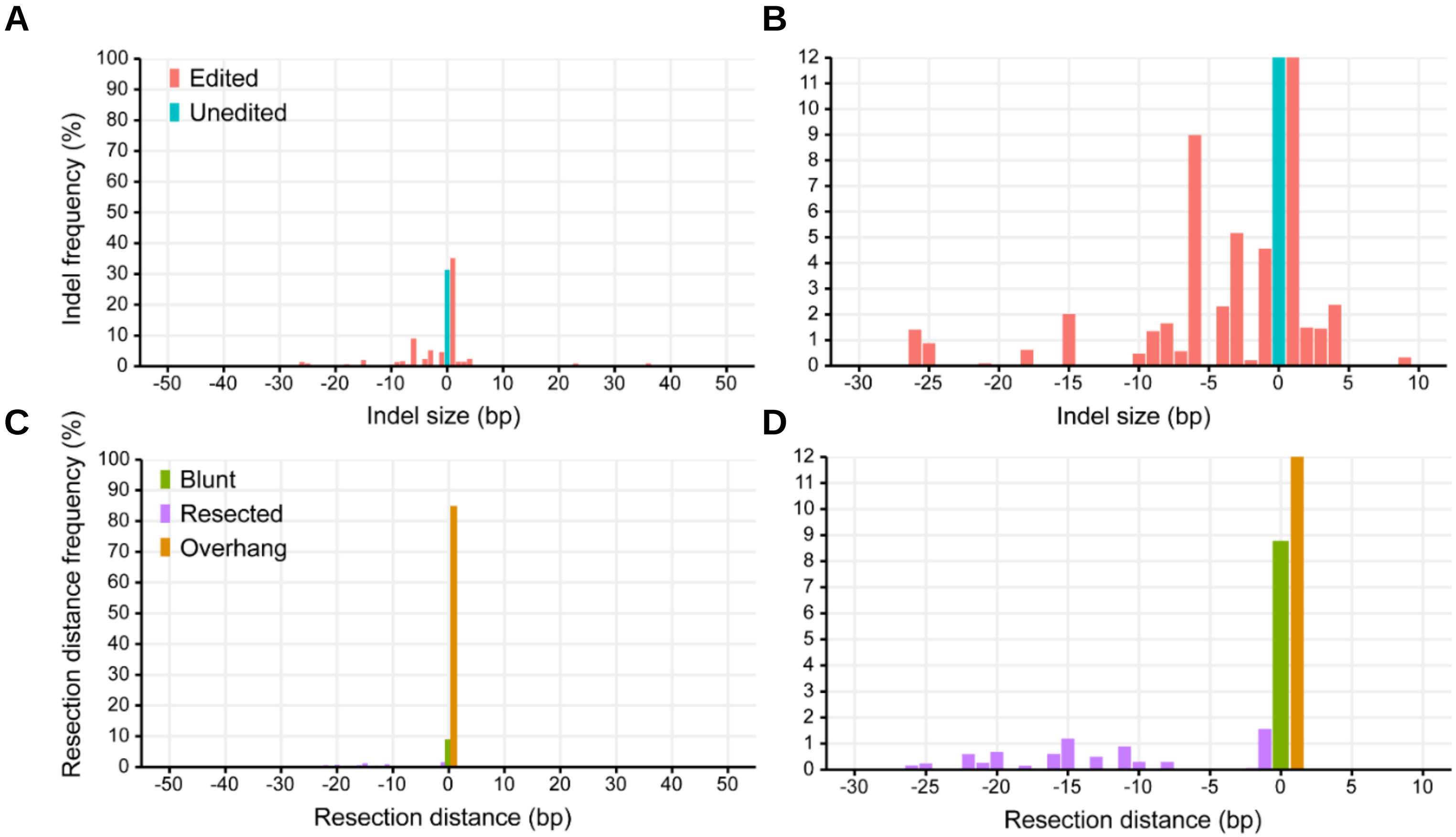


Figure 6.9. Comparing the indel profile and the range of break structures at the EMX1 on-target site. (**A**) The indel profile measured at the EMX1 on-target. The frequency of edited amplicons is shown by pink bars and unedited by blue bars. Deletions are represented as negative values and insertions as positive. (**B**) Zoomed view of (**A**). (**C**) The frequency of different end structures at the EMX1 on-target. Blunt break structures are shown by green bars, resected ends by purple bars and negative resection distance values, and overhanging ends by orange bars and positive resection distance values. (**D**) Zoomed view of (**C**).

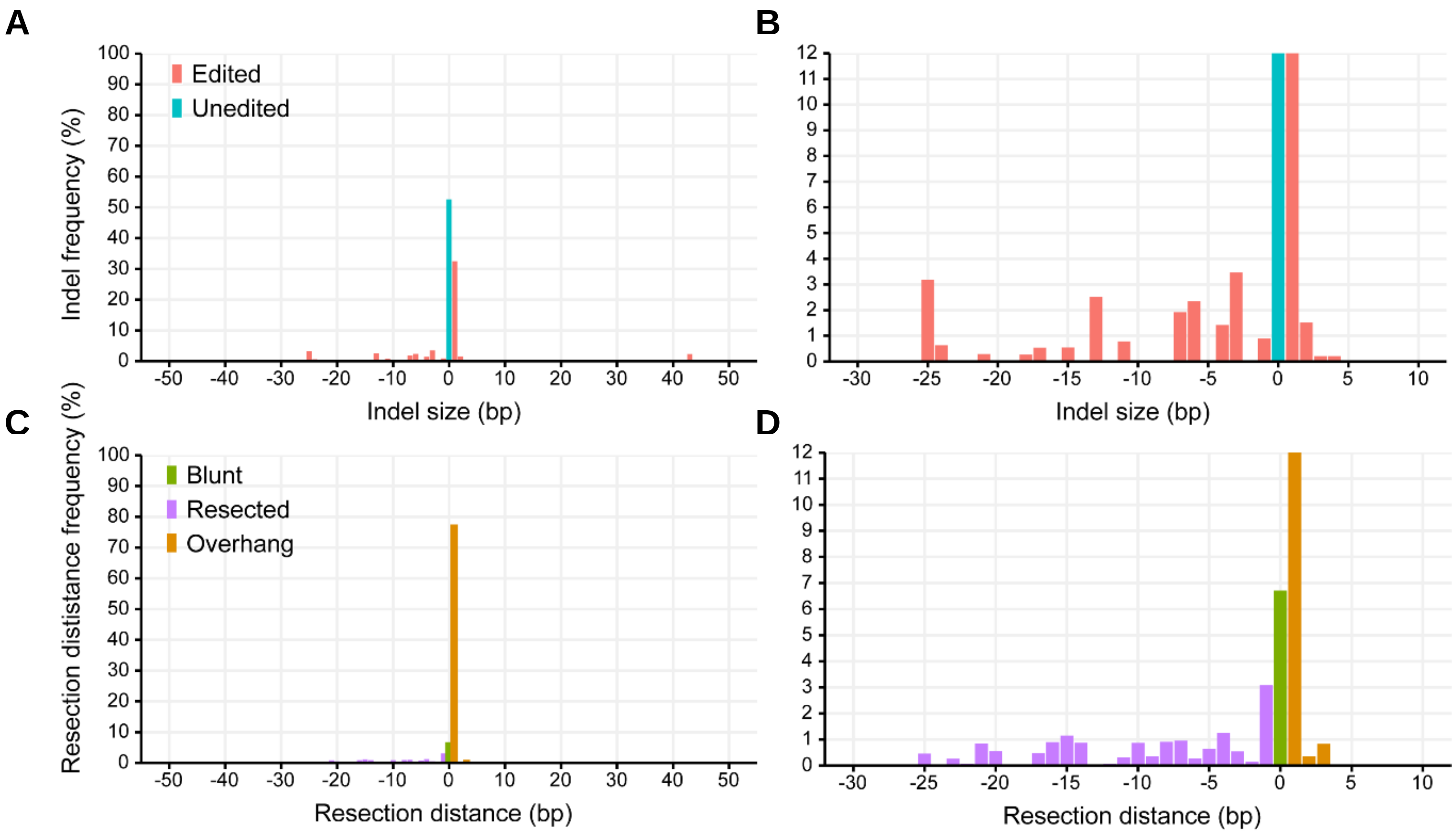


Figure 6.10. Comparing the indel profile and the range of break structures at the EMX1 OT1. (**A**) The indel profile measured at the EMX1 OT1. The frequency of edited amplicons is shown by pink bars and unedited by blue bars. Deletions are represented as negative values and insertions as positive. (**B**) Zoomed view of (**A**). (**C**) The frequency of different end structures at the EMX1 OT1. Blunt break structures are shown by green bars, resected ends by purple bars and negative resection distance values, and overhanging ends by orange bars and positive resection distance values. (**D**) Zoomed view of (**C**).

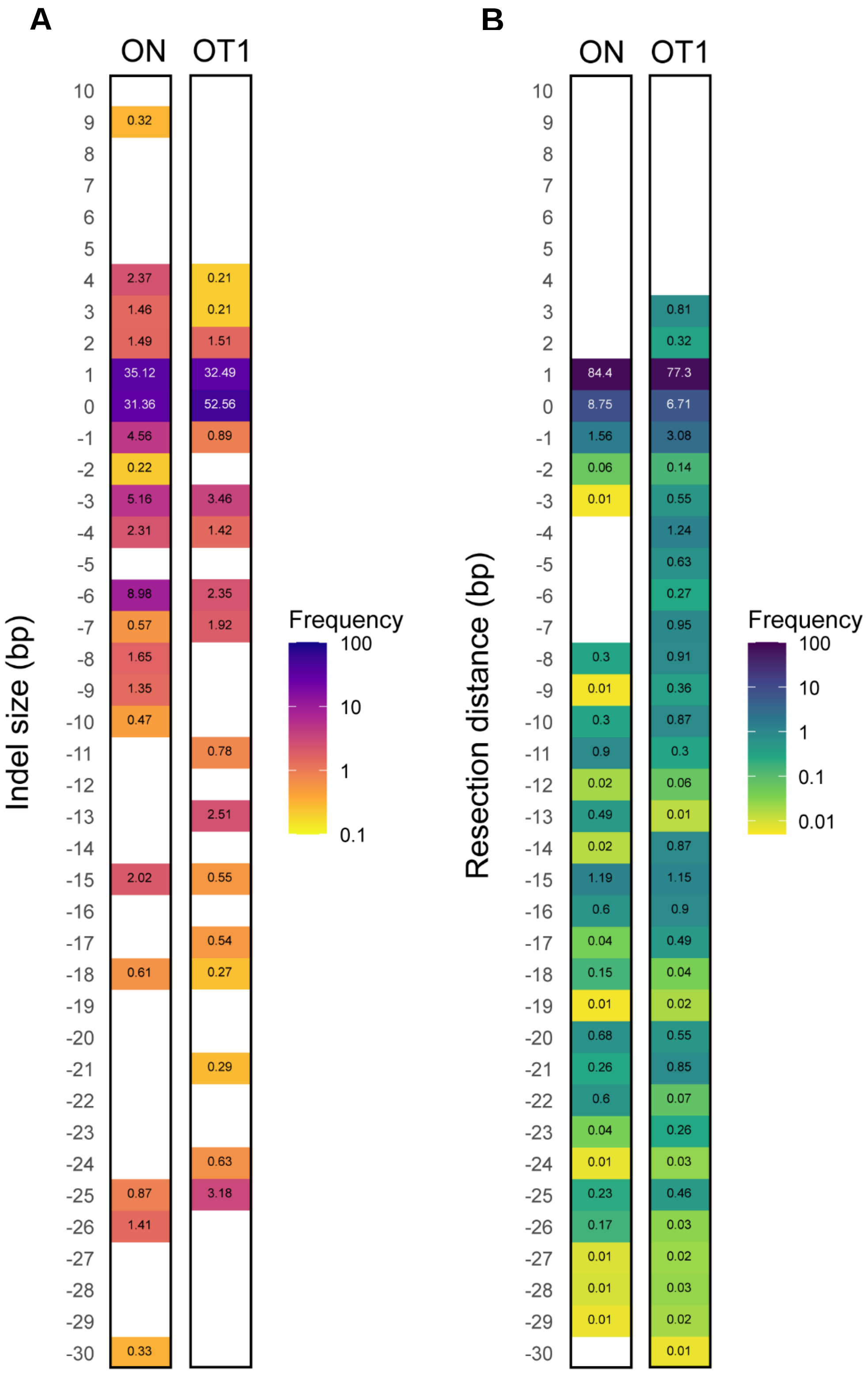


Figure 6.11. Heatmaps comparing the indel profile and resection distance at the EMX1 on-target and OT1. (**A**) The frequency of indels around each site. Deletions are represented by negative indel size values, insertions by positive values and unedited amplicons by 0. (**B**) The frequency of each resected end structure at both sites. Resected break ends are represented by negative distance values, overhang structures by positive values and blunt DSBs by 0.

6.1 Discussion

Understanding the mutagenic outcome of genome editing is a critical step that enables the safety assessment of any genome edited cell and gene therapy. In this regard, measuring indel formation at on- and off-target sites using targeted amplicon sequencing is undoubtedly a powerful way to understand unwanted off-target mutagenesis, in addition to characterising the variety of editing outcomes that occur at the on-target site. The purpose of the experiments described in this chapter was to determine the editing outcome at off-target break sites identified previously by INDUCE-seq. This reveals the relationship between the frequency of break induction and the frequency of mutational outcome. Limitations of amplicon sequencing became apparent, however, when many fewer off-target indels were detected than had been determined as off-target breaks by INDUCE-seq. Given the well-established detection limit of amplicon sequencing of ~0.1%, and the high sensitivity of INDUCE-seq, this is not surprising. This was further corroborated by data shown previously using the same guide RNA and experimental design (Vakulskas et al. 2018). Bounded by the mutation detection limit of amplicon-seq, it is possible that failure to detect edits at many of the off-targets detected by INDUCE-seq is a true representation of faithful repair outcome. However, direct measurements of error-free repair would be required to confirm this. Unfortunately, there are currently no assays with the same level of sensitivity as INDUCE-seq to be able to do this. The current paradigm for confirming true positive off-target sites is to validate any potential site using amplicon sequencing, only deeming it 'real' if indels can be detected in a cellular or *in vivo* system (Akcakaya et al. 2018; Chaudhari et al. 2020). While this paradigm is appropriate when measuring CRISPR off-targets using an *in vitro* discovery method, which require cellular and/or *in vivo* confirmation, it is not appropriate when using highly sensitive cell-based off-target detection methods such as INDUCE-seq. In future, confirmation of indel formation at off-targets should be performed using an error-corrected sequencing approach such as duplex sequencing, which is not limited by the 0.1% detection limit of amplicon sequencing (Kennedy et al. 2014). Taking this approach could help to determine with more accuracy whether the off-targets identified by INDUCE-seq result in a low-frequency indel outcome or are faithfully repaired by the cell in an error-free manner, which is critical to understand when considering the safety of CRISPR genome editing. Furthermore, expanding the editing outcome analysis to regions beyond the 150-250bp captured for amplicon sequencing, using long-read sequencing would also enable the discovery of larger indels and structural alterations such as translocations, which have been identified at the on-target site by several groups (Kosicki et al. 2018; Bi et al. 2020).

In addition to confirming the presence of editing at the off-targets identified by INDUCE-seq, the results in this chapter explored the relationship between the kinetics of break induction and the editing outcome. At the on-target and most frequent off-target site (OT1), the number of accumulated breaks correlated strongly with the frequency of editing outcome. This suggests that the number of breaks measured by INDUCE-seq over time is proportional to the frequency of indel formation, which presents the intriguing possibility that early break measurements could predict the frequency of the eventual editing outcome. Strong correlation was not observed at OT2, however, which is likely due to the fact that many of the indel measurements were zero values. The causes of this could have been technical, or simply because the indel frequency fell below the detection limit of amplicon sequencing for those particular samples. Another possibility is that the kinetics of break induction and repair efficiency are unlinked at this position in the genome due to structural features of the DNA and/or chromatin that affect either process. Future kinetics experiments using a range of guide RNAs would confirm whether these findings are applicable more broadly across targets, or are specific for the EMX1 guide RNA used here.

Despite the sensitivity limitations of amplicon sequencing, the approach reveals the full spectrum of the indels present at sites edited with a higher frequency. Analysis of the indel spectrum was performed for the on-target and OT1, which revealed a range of insertion and deletion events ranging from single nucleotide insertions to larger >20bp deletions. Given the importance of understanding the range of editing outcomes at on- and off-target sites, the distribution of indels generated following CRISPR editing has been well characterised in recent years. Several groups have determined that cell-line dependent bias and target sequence context drives the DNA repair outcome in a non-random manner (van Overbeek et al. 2016; Taheri-Ghahfarokhi et al. 2018; Allen et al. 2019). At the EMX1 on-target site and off-target sites OT1 and OT2, a 1bp overhanging break structure was identified by INDUCE-seq. This correlates with the most common indel editing outcome (1bp insertion) measured at these sites. These findings suggest that the structure of break ends measured by INDUCE-seq, which represents repair intermediates during the editing process, may reveal the mechanisms of end repair that explain the non-random editing outcome.

To further investigate this, the distribution of break ends around the EMX1 on-target site and OT1 was determined. This revealed a range of end structures, in addition to the 1bp overhang described previously. Comparison of the indel distribution and the break structure distribution showed that in addition to site-specific indel profiles, site specific break structures were present at the on-target and OT1. Furthermore, although the high frequency 1bp overhang correlated with the high frequency 1bp insertion at both sites, the general profiles of end-resection and indel length were not directly comparable. This suggests that multiple factors contribute to the resection profiles observed. Indeed, it is formally possible that an overhanging break-end directly drives the majority of editing outcomes toward the 1bp insertion, while the remaining indels are formed as a result of DNA end-resection, exposure of regions of microhomology, and repair by the MMEJ DSB repair pathway. Although a direct 1:1 relationship was not observed for DSB end position and indel size, the characterisation of DNA-end resection using INDUCE-seq could become critical to understanding the mechanism of DSB repair, at least in the context of genome editing where the break position is known. For the measurement of end-resection at sites where the break position is unknown, would require a modified version of INDUCE-seq which measures resected-ends directly. Future analysis should integrate the surrounding sequence and chromatin context with the resection profile to understand if the two factors are related, and whether cell-specific repair activity coupled with sequence context can be used to reliably predict the outcomes of genome editing on- and off-target.

7 Chapter VII – General Discussion

The aim of this thesis was to establish a novel tool capable of measuring off-target genomic DSBs caused by CRISPR-Cas9 genome editing. The emergence of CRISPR-Cas9 in recent years has catalysed a genome editing revolution, resulting in a new wave of cell and gene therapies that are made possible by directly and permanently altering our genetic code. Despite this potential to provide transformative therapies for genetic diseases, safety concerns remain around the issue of off-target genome editing that can be induced by CRISPR-Cas9. Specifically, these off-target DSBs have the potential to result in often lethal mutagenic outcomes that can directly drive oncogenesis, calling in to question the safety of current genome editing strategies. Because of this issue it has become evident that empirical measurement of off-target activity in the genome is required to assess the risk of any CRISPR guide RNA. However, current approaches to do this remain flawed and do not provide the sensitivity or accuracy required to enable the safe and sustainable development of CRISPR-Cas9 based cell and gene therapies.

To address this problem, a novel method, INDUCE-seq, was developed that could provide an accurate, unbiased, and sensitive measurement of DSBs in the genome. INDUCE-seq exploits a novel PCR-free library preparation that eliminates amplification bias and enables the simultaneous enrichment and sequencing of genomic DSBs. Using this approach, each sequencing read generated represents a single break-end labelled in the cell, and thus produces a uniquely unbiased and non-distorted representation of breaks. The characteristics of INDUCE-seq were first determined using cells that had been digested *in situ* using the restriction endonuclease HindIII. These experiments showed the unrivalled dynamic range of INDUCE-seq, demonstrating that the method could simultaneously detect ~150 million on-target induced breaks in addition individual low frequency off-target sites that were represented as few as five times throughout the cell population. Furthermore, these experiments demonstrated the exquisite sensitivity possible using INDUCE-seq. In the same experiment, ~1.5x10^{8} HindIII induced-break DNA fragments were enriched and sequenced from a sample comprising ~8x10^{10} genomic DNA fragments that were unrelated to HindIII-induced breaks. This represents a ~500-fold enrichment. The untreated samples from the same experiment set demonstrate INDUCE-seq's enhanced sensitivity even further, which revealed ~200,000 endogenous breaks from a genomic DNA sample comprising ~2.5x10^{11} DNA fragments. This represents a staggering ~1.2M-fold enrichment. INDUCE-seq's enhanced dynamic range and sensitivity is directly enabled by the novel flow cell enrichment strategy that it employs. This novel use of the Illumina flow cell, coupled with the PCR-free library preparation, results in a digital 1:1 readout for break-labelled DNA fragments. Enrichment of non-amplified DNA fragments via hybridisation on the Illumina flow cell dramatically enhances the dynamic range of detection as the measurement is only limited by the sequencing capacity of the flow cell that is used. It is therefore conceivable that both the sensitivity and dynamic range of this approach will be further advanced in the future by improvements in flow cell technology, which is likely to scale with the increasing capacity of modern sequencing platforms. As was discussed in Chapter III, a range of other genomic factors including other types of DNA damage and DNA end resection can be conceived of that could be measured using a similar approach. Any genomic feature that can be converted into a DNA end can potentially be captured, enriched for, and sequenced using this transformative approach. This was not an exhaustive list, and other applications may emerge in the future that take advantage of this technology in currently unimagined ways.

In Chapter IV, the characteristics of INDUCE-seq were determined by measuring induced and endogenous breaks in different cells. First, INDUCE-seq's ability to reproducibly measure enzymatically-induced breaks in live cells was examined using the AsiSI-inducible DIvA cell system (Iacovoni et al. 2010). Because of INDUCE-seq's eventual application for CRISPR-Cas9 off-target detection, it was important to quantify its reproducibility using a well-characterised system that induced breaks at many predetermined locations of the genome. INDUCE-seq was highly reproducible across experimental replicates, and revealed consistent site-specific differences in the levels of break induction across the AsiSI sites. Furthermore, these experiments also revealed that only ~200 of the ~1200 AsiSI recognition sequences in the human genome are cleaved. This is likely because of the methylation sensitivity of the enzyme, which has been shown previously in several other studies genome (Iacovoni et al. 2010; Massip et al. 2010; Aymard et al. 2017). Interestingly, at the AsiSI sites where break induction was measured by INDUCE-seq, a distribution emerged showing that the level of break induction varied considerably between different sites. The number of breaks found per AsiSI site appeared to follow a pareto-like distribution, where the majority of breaks measured originated from a small number of sites, and many sites were found with very few breaks. These experiments demonstrated how INDUCE-seq's unbiased and quantitative readout can be used to better understand the precise characteristics of a complex biological system, which would otherwise be masked by amplification bias when using alternative PCR-based break capture methodologies. In future, information concerning the differential levels of induced break induction in the genome may be valuable for the design of CRISPR guide RNAs. In this manner, guides with off-target sites at locations of the genome that are more susceptible to break formation, or are repaired less efficiently, may be avoided. Furthermore, the same information may enable the design of highly efficacious guide RNAs that induce high levels of on-target editing at specific locations in the genome, even when administered at very low concentrations. It has become evident that understanding this differential rate of break induction and repair throughout the genome will become key for the development of safe genome editing therapies in the future.

The second part of Chapter IV investigated the endogenous break landscape in a variety of different cell types. These experiments revealed cell-type specific differences in the number of DSBs found in different cells. Remarkably, break numbers ranged three orders of magnitude across the cell types from ~10^5 to ~10^7 per sample. The samples with the most breaks were the *in vitro* differentiated neural progenitor cells and mature neurons, which showed substantially higher break counts than the other cell types examined. Elevated numbers of breaks have been observed previously in *in vitro* differentiated neural progenitor cells, and it has been postulated that these breaks are involved in transcriptional regulation during neurodevelopment (Alt and Schwer 2018; Tena et al. 2020; Tully 2020). However, these finding also highlight the intriguing possibility that the *in vitro* differentiation process itself drives genome instability in a manner independent of cell type. Indeed, studies have shown that human IPS cell (hiPSCs) reprogramming increases DSB number in the presence of functioning DSB repair (Simara et al. 2017). Furthermore, after prolonged *in vitro* culturing of hiPSCs, repair capacity decreases and non-random chromosomal abnormalities have been observed (Na et al. 2014). Better understanding of the number and distribution of endogenous breaks in long-term cultured and developing hiPSCs would therefore help to elucidate the mechanisms driving the acquisition of genomic damage in these systems, which in turn would lower the barriers for the delivery of regenerative medicines. Furthermore, this would have significant and serious implications when considering different tissues or cell types for targeting during genome editing in cell or gene therapy.

When examining the relationship between the total number of endogenous breaks and the number of recurrent breaks across the various cell types, a very strong linear correlation was observed. These findings initially suggested that the number of recurrent breaks sites was simply a reflection of an increasing number of randomly positioned breaks. However, given the non-random sequence composition of the genome and the biological processes that act upon it, this is very unlikely. To determine whether DSBs were positioned stochastically throughout the genome, the experimental break datasets were compared with a comparable set of randomised, *in silico* generated datasets. These experiments revealed that experimentally defined break positions cluster closer together than an equivalent number of random positions. Furthermore, when recurrent break sites were defined based on exact overlapping 1bp endogenous break positions, break recurrence appeared independent of total break number. Collectively, these findings demonstrate how different classes of recurrent break sites can be defined from a non-random distribution of single endogenous break positions. This kind of analysis has not been possible before using PCR-based DSB capture approaches, which typically can only measure positions with substantial recurrent break formation (Canela et al. 2016). Using INDUCE-seq, it is now possible to quantify and characterise low-level sporadic break positions that occur as a result of genotoxic agents, ionising radiation, DDR drugs, or even spontaneously from endogenous processes such as DNA replication and transcription. These findings have significant implications for the safe development of novel therapeutic chemicals and innovative DDR drugs in the future, that previously, generated undetectably low levels of DNA breaks. Indeed, a new paradigm of data-driven risk assessment is emerging in the toxicology field, which is being driven by the adoption of next generation risk assessment (NGRA) approaches over legacy cell/fluorescence-based methodologies (Liu et al. 2019b). In the future, INDUCE-seq could potentially replace animal-based assays for measuring genome breaks caused by a variety of substances. Furthermore, because of its enhanced sensitivity, a range of novel chemicals, drugs, and material may be taken to market that are deemed safe at the concentrations relevant for human use.

When INDUCE-seq was applied to the study of CRISPR-induced off-targets in the genome, it was revealed that the approach could detect numerous novel off-targets in live cells that had not been previously observed by a range of other methods, including *in vitro* screening methods CIRCLE-seq and Digenome-seq (Kim et al. 2015; Tsai et al. 2017). These findings are of great importance as *in vitro* methods are considered a 'catch-all' for revealing all CRISPR-Cas9 off-target activity. Based on these findings, it seems inappropriate to use an *in vitro* methodology to define off-targets because of the tendency to significantly overestimate the number of off-target sites while simultaneously missing sites that can be induced in the cellular environment. Further exacerbating this was the discrepancy observed between off-target breaks, detected by INDUCE-seq, and indel editing outcome measured by targeted amplicon sequencing. Very few of the INDUCE-seq-identified off-target breaks were reflected by indels in the amplicon sequencing readout, calling in to question the sensitivity of this approach for validating editing outcome. These results are not surprising given the known detection limit of amplicon sequencing of ~1 in 1000. However, this is particular concerning as a proposed strategy for the risk assessment of CRISPR-Cas9 off-targets it to use an *in vitro* discovery tool such as CIRCLE-seq, followed by amplicon sequencing in the relevant edited cellular or *in vivo* system (Akcakaya et al. 2018). It is now evident that a more suitable approach for off-target risk assessment would be to couple an ultra-sensitive DSB mapping approach such as INDUCE-seq with an equally sensitive, error-corrected sequencing approach such as duplex sequencing (Kennedy et al. 2014), measuring entirely using material from the relevant cell and/or animal model. Conceivably, by combining the INDUCE-seq library preparation with the retention of duplex information could enable a novel approach for, PCR-free error-corrected sequencing. This would significantly reduce the cost of ultra-sensitive whole-genome mutation analysis while retaining the quantitative characteristics of the INDUCE-seq library preparation. In addition to revealing the safety of novel genome editing tools, the full understanding of both genomic break induction, and the accumulation of mutations in the genome would provide significant insight into the forces driving genome instability.

Beyond simple site identification, INDUCE-seq also enabled the accurate profiling of mismatches present at the CRISPR-Cas9 off-target sites. This analysis was made possible by the quantitative nature of the INDUCE-seq readout, which reveals the non-distorted and proportionate distribution of off-target sites. Analysis of the mismatches occurring the off-target sites revealed consistent mismatch types at various positions of the guide RNA sequence. Furthermore, classification of the mismatches based on the location within the sequence revealed an off-target mismatch ‘signature’ for the EMX1 guide RNA. This quantitative information is ideally suited for the application of machine learning or other AI approaches in the future. It can therefore be conceived that the unbiased measurement of many thousands of guide RNAs using INDUCE-seq, would enable the data-driven design of novel guide RNAs with safer off-target profiles. This methodology could be applied to a range of CRISPR systems, including alternative Cas nucleases such as Cas12A, and even base editing approaches which have been shown to induce low-level DSBs at the on-target (Komor et al. 2016). Regardless of the system, the development of a technology platform dedicated to assessing the safety of any therapeutic CRISPR enzyme and/or guide RNA would certainly enable the safer development and introduction of novel cell and gene therapies to patients.

Finally, the break structure identified at CRISPR-cleaved sites was calculated to determine if break-end position was predictive of the indel repair outcomes. At the on-target and most frequent off-target sites, a 1bp overlapping break structure predominated, which correlated with the most common editing outcome, a 1bp insertion. Despite this, when the full spectrum of break-end positions around the induced break site were calculated, no direct correlation was observed between the resection distance and deletion size. These findings suggest that several factors, including resection distance, are at play, which lead to the observed indel distribution at the CRISPR-induced sites. Limitations of this work include the need for predefined, induced-break positions in order to define the position of the resected ends. Furthermore, because of exonuclease and polymerase activity during the blunting stage of the break labelling procedure, this was not a direct measurement of DNA end-resection. As described previously, it is also evident that the sensitivity of amplicon sequencing becomes limiting when interrogating low-frequency indels. This approach does not generate a genome-wide measurement, as amplicon positions must be predetermined, and cannot reveal indels larger than the amplicon size. The application of error-corrected whole-genome sequencing, and whole-genome long-read sequencing, would enable the unbiased and ultra-sensitive interrogation of a range of indels up to several kilobases in size. Future work should focus on combining the direct measurement of resected DNA ends, with such strategies for mutation detection, which in turn would reveal the mechanisms that drive DNA end-resection during DSB repair. This information would be extremely valuable in a range of basic research and applied contexts. Better understanding of how genetic diversity is facilitated through the transfer of genetic material between different regions of the genome will be pivotal in determining the precise mechanisms of how DSBs drive diversity in the immune system, or lead to the development of lethal cancers (Chapman et al. 2012). Furthermore, understanding how different sequence strings expand and contact will be essential when considering the rational design of synthetic genomes and chromosomes. Indeed, the stability of ‘minimal’ or ‘safe’ human genomes that lack many repetitive elements could become a major barrier for the application of such technologies (Schindler et al. 2018). Finally, this knowledge could help to overcome the key limitation of heterogeneous repair outcomes following CRISPR-Cas9 genome editing, which would enable precise and directed genome editing via the formation of a DSB.

8 Appendix

Appendix A – INDUCE-seq oligonucleotides and rhAmpSeq primers

Table A1. INDUCE-seq adapter sequences

P5 adapter top strand	5'	A*ATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T	3'
P5 adapter complement C3	5'	/5Phos/GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT/3SpC3/	3'
P7 adapter top short	5'	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCAC/3SpC3/	3'
P7 AD001-B	5'	C*AAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD002-A	5'	C*AAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD003-B	5'	C*AAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD004-A	5'	C*AAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD005-A	5'	C*AAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD006-A	5'	C*AAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD007-A	5'	C*AAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD008-B	5'	C*AAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD009-B	5'	C*AAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD010-B	5'	C*AAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD011-B	5'	C*AAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD012-A	5'	C*AAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD013-A	5'	C*AAGCAGAAGACGGCATACGAGATTGTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD014-A	5'	C*AAGCAGAAGACGGCATACGAGATACGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD015-A	5'	C*AAGCAGAAGACGGCATACGAGATTCTGACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD016-A	5'	C*AAGCAGAAGACGGCATACGAGATCGGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD018-A	5'	C*AAGCAGAAGACGGCATACGAGATGTGCGGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD019-A	5'	C*AAGCAGAAGACGGCATACGAGATCGTTTCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD020-B	5'	C*AAGCAGAAGACGGCATACGAGATAAGGCCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD021-B	5'	C*AAGCAGAAGACGGCATACGAGATTCCGAAACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD022-B	5'	C*AAGCAGAAGACGGCATACGAGATTACGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD023-B	5'	C*AAGCAGAAGACGGCATACGAGATATCCACTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD025-B	5'	C*AAGCAGAAGACGGCATACGAGATATATCAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'
P7 AD027-B	5'	C*AAGCAGAAGACGGCATACGAGATAAAGGAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T	3'

A2. rhAmpSeq primers for Amplicon sequencing.

rhAmpSeq Fwd	rhAmpSeq Rev
AGGACAAAGTACAAACGGCAGAAGC	TTGCCCACCCTAGTCATTGGAGGT
GAGTCTGACACCTTTTAAGATCTGACAG	GAAAACTCAAAGAAATGCCCAATCATTG
AATGTGCTTCAACCCATCACGGCCT	CATCAGTGTTGGCTTTCACAAGGATG
CACACCAGCAATGCTCTCGTCTT	CTTGGCCCTTCCTCTGTACTCTA
AAAATCCTCCCCGGCTTAGGTAC	TCTCCCTATCCCTCACAAACCGTC
GAAGACCAGACTCAGTAAAGCCTGGA	AAATCCTTCTTTAGGGCCCAGCTGT
GATGTAGTTCTGACATTCCTCCTGAG	TTTTGGTCAATATCTGAAAGGTTTATTTG
AGATGGGTGTCGGGATCCCGCT	CCGCCCTTACTAGGACATGTCCA
GGGATGTTCTTCTTGCCTTGGTAG	ACCATTCACTCCACCTGATCTCGG
TTTCCTAGCTATCTTAACACCTCCCTTTG	GAAACAAAGGCTAACCTGGTTAACAGTA
CATCAGCTGTTCGTAGGACATGT	GGTTAGCCATTTACGCTCACATCT
GAGTCAGGTGAACAAAATTCATTGGTTCC	TTCAGTGGTTTTGAAGACTCTTTTCCCTTC
GAATAAAGGCGAGGAAGCGGGAGC	CTGTCTGCCTCTGACGACGAGCA
TGACTTTTCTTTCTGGCTTGTCCTCTC	ATATTTGTAGGGTCCAGGCAAGAGAA
GGGAATAAACTTGTGCTTATTTGTTGGAAG	TGAATATGTTTTAAATTCTCCACAGTAAAAGAG
CAAACTCTGGCTGTTTAGAAGGCTAGT	GGCCTGAAACCTCTGAGAATATGATTA
AGAGGCAAACTAGGACAACCACTTAC	CCGGTCTGAGTGATGGAAGGAACC
CTCGTAGGACATGTCCAGCAGCT	TCCAGTATGCTGCTCTTCTGAGGA
CTCTGAACTCTGTTGCCAGCAAAG	TGGATTCTGGTGTCTCCTCCTAC
AAATTGAGTGTCTAGGAGTTGTGTGAC	TTCACCTAATTTGGGTAGTTTTTAGGCA
AGATAGCCTCTCATTCCATAATCCAGGG	GAAATTGTGGGTAACTGTTTAGATGAGCC
GCTCCTTTTAGCCACATCTGGAGTG	AGGGAATGCTAGAGAATGTCACAGCA
TGGCATTTGAATAAAGTGTTAGGATTACAGGC	TGTCTGTCACATACATATGTACACACGCAA
CCCTCATTGCATTTATCATGGTTGTAAC	GCAAATTCAGGAACTCTTCATTCACTTAC
TTTCTCACTGCCCTGGTCCAAC	CAAAGCTGGCATCTGTCTCAGGA
TCTAACATTATCTCACTCAGAAGCTCCATGT	GATTTCTTGCTAAATCCTAGAACAGTGGTTT
CATGGTTCTCTTGCTTACCTGTGAT	CAGTTAGTTGGCTGAATAGGAAATGT
CCCTTTAACACAGGGCAAATGCATA	TAAAAGCAGGGTACTTGTGCCTTGG
CAAGCTGCAAGTCTTGGTGTCTT	TGCACTGCCCTAGTCAAGGTTCT
GAACTCACCTCTCTAATTCCGGCAAA	GTGACTTGAATAAACCCTCACCCTAG
CATTTTCCTGGGTTCACAGCCTCC	CTGAGTGGCCCAATATAATCACAAGG
CTGAATATTTGGCAGTTTGTAAGGTGAATG	ACATTTACTACACCACATCTGTTTGGCAT
CCCATCCTTTGTCCCAATTCACTT	ATCCTGATTCACATGGGAAGGGGTA
AGATAAACTTTGCTTCTCTTTTCGGCTTTG	AATGGACTTTTACTGAATCCGGAATAGTAAC
GGTACTTCATGTCCGTGCTCTTGA	CACTTGGAGAGTCAGAGGTCACAA
AGGCCAACTCATGACAGCATC	CAGCTTCCTGCAGTGAGAAAG
GTAATGATTCTGCCTTAGAGTCCCAGGT	AGTACTATATTATTTATGGCAGGGCATGGG
GGCTTAATATAATGGCAAACGGAGATCA	TCCAGTTATCAACATGTTGGCTGTTCT
AAACATGATGCTGGCATCTGCTGCA	TGGTGCTGTCCTCGATATAGTGAGTG
CTAGTTGGAGTTCATTGAGCTTCTTGG	CAGAGACCTGTGGGACATTATCAAGT
GTCAAGCTGTGCATAAAAGGACTTGT	GTAACTGGAGCTGGATTTGACACAA
CTCCATCTTCTGCTGGTTCCGATT	GGGAAAATGGGAACAGTTGGGAAG
TAGCAGGTTGTGGGAGACAGGA	GACCCTCCTATCTGCTTCCCTTC
GGCCATTCTTTATAGTTTGACTCTGGAGA	TGAAAATGTCACATACTCATTCCTCAACC
ACCAGAAGTCCTCAGGGCCCTG	AGGAAAGCTGTCTGGAGAGGCCA
GGTTATTGTAGCTTCCTAAGTGGTCCT	AAGGAAGTCGTTAGAAGACTCTACAGAC
TTTCAGTTTTGTTTCTCTTTCCTTGCCTC	TCACCATAGAGGACTTTAGAAGAGGTGA
CCACCAAATTCCTGCCAAATAGTTCC	GGTAGCATCTGTGACTTCACTTGCAGT

CACAGGACACCTCTCGTTCCATTT	AAAAGGATGCCACGTGACAGCCATT
GAGGTTGATGAAGTAAGGTCAGCATGT	GGCTTGAGATTTTGCATTTCTAACACAC
CAAAGGCGAGAGCCGACTCAA	AGAAGGTTCTCGCGGTTGCGTT
CCTAGAAGTTGGGACTACTCAGCAGA	ATGCTAGTGCTAGCCTTCTGGCTTC
AGGAGATGATTGAATGGCATTTGGAAG	TCCTGAGAAAATCCAAGACCTCACAT
CACAACCAAGGTTTGAATTTCAAGACGG	AATAAGAAGTCTCTCGCTAACAGCCGGC
AGCCAGGCACAGAGAGGAGAG	TCTCTGCCACACTGACTCACTC
AAGGAAATCATTGACAGAGGGGTCTG	AGCATTTCCCAGACCATCGTCCT
AGATTAGGCCGCGAGCCGAAG	ATCCGATCCNGTGGTGCCCCA
CAGCATTAGTTTTCTGGTGTCTCATGA	GCCTAGATACTTCATCCATAAGCAGCTT
AAAACAGCAGTAGGCCAAGACAATG	GACAGTTGTGCCAAACATGTAAAGA
CATTGGTCCGCCGACTCATCT	GGAAACCCAGGCTGGTTCTTGT

Appendix B – Scripts used for bioinformatics analysis of INDUCE-seq data

INDUCE-seq processing script

```
#!/usr/bin/env bash
clear
# Setting variables
method=INDUCE_SEQ
platform=NEXTSEQ
BAM=BAM
SAM=SAM
BED=BED
FASTQ=FASTQ
SUM=summary
tmpdir=tmpdir
blacklist=accessory_files/hg19.blacklist.bed
chromsizes=accessory_files/hg19.chrom.sizes.bed
chromends=accessory_files/hg19.chrom.ends.bed
refseq=/home/postdoc/bwa_refseq/Homo_sapiens_assembly19.fa
quality=30
threads=15

#Detect FASTQ files
for file in "$FASTQ"/*.fq*; do (
filename=$(echo "$file" | awk -F'[/]' '{print $2}')
echo "$filename" >> filelist.txt
experiment=$(echo "$filename" | awk -F'[.]' '{print $1}')
echo "Found" "$filename" "from experiment:" "$experiment"
# rm filelist_"$experiment" #comment this line if you want to test
whether the above functions actually finds your files

# Generate folders
mkdir -p "$experiment"
mkdir -p "$experiment"/preprocessing/"$BAM"
BAM="$experiment"/preprocessing/"$BAM"
mkdir -p "$experiment"/preprocessing/"$SAM"
SAM="$experiment"/preprocessing/"$SAM"
mkdir -p "$experiment"/preprocessing/"$BED"
BED="$experiment"/preprocessing/"$BED"
mkdir -p "$experiment"/preprocessing/"$SUM"
SUM="$experiment"/preprocessing/"$SUM"
mkdir -p "$experiment"/preprocessing/"$tmpdir"
tmpdir="$experiment"/preprocessing/"$tmpdir"

# Alignment
if [ ! -e "$BAM"/"$experiment".q30.srt.bam ]; then
  bwa mem -t "$threads" -M -R
"@RG\tID:"$method"\tPL:"$platform"\tPU:0\tLB:"$method"\tSM:"$experime
nt"" "$refseq" "$FASTQ"/"$experiment"* > "$tmpdir"/"$experiment".sam
  awk '$6 !~ /[0-9]S/{print}' "$tmpdir"/"$experiment".sam | samtools
view -Shu -q "$quality" -F 256 -@ "$threads" - | \
  samtools sort -@ "$threads" -m 4G - -o
"$BAM"/"$experiment".q30.srt.bam
  samtools index "$BAM"/"$experiment".q30.srt.bam
  mv "$tmpdir"/"$experiment".sam "$SAM"/"$experiment".sam
else
  echo "BAM file found in:" "$BAM" "skipping alignment"
fi

################################################################
# Filter out blacklist regions
```

```
echo "Preparing and filtering bed files"
bedtools bamtobed -i "$BAM"/"$experiment".q30.srt.bam |
bedtools intersect -v -bed -a - -b "$blacklist" |
bedtools intersect -v -bed -a - -b "$chromends" |
bedtools intersect -wa -a - -b "$chromsizes" >
"$BED"/"$experiment".q30.srt.bed

###############################################################
# Split by strand
echo "Defining break ends and counting break events and locations"

awk '{ if ($6 == "+")
print $1,$2,$3=$2+1,$4,$5,$6;
else
print $1,$2=$3-1,$3,$4,$5,$6 }' OFS="\t"
"$BED"/"$experiment".q30.srt.bed | \
LC_ALL=C sort --parallel="$threads" -k1,1 -k2,2n | \
sed 's/:/ /g' | awk '{print
$1"_"$2"_"$3"_"$4":"$5":"$6":"$7":"$8":"$9":"$10"_"$11"_"$12, $8, $10
}' |\
awk 'NR==1{p=$2;q=$3;next}     #p=tile and q=y coordinate
{print $1, $2-p, $3-q; p=$2 ; q=$3}' | \
sed 's/_/ /g' |
awk -v BED="$BED" '{ if (($7 == 0 || $7 == 1) && $8 <= 40 && $8 >=-
40)
{print $1, $2, $3, $4, $5, $6 > BED"/optical_duplicates.txt"}
else {print $1, $2, $3, $4, $5, $6}}' OFS="\t" >
"$BED"/"$experiment".breakends.bed
bedtools merge -s -c 2 -o count -i "$BED"/"$experiment".breakends.bed
| awk -v var="$experiment" '{print $1,$2,$3,var"_"(FNR FS),$5,$4}'
OFS="\t" > "$BED"/"$experiment".breakcount.bed

echo "Breakcounts files ready for secondary analysis"

################################################################
# Get Summary for the data (number of reads after each filter, etc.)
echo "Writing summary file"
summaryfile="$SUM"/"$experiment"_summary

echo "Alignment statistics:" > "$summaryfile"
samtools flagstat "$BAM"/"$experiment".q30.srt.bam >> "$summaryfile"

echo "Number of tagged break-ends:" >> "$summaryfile"
wc -l "$BED"/"$experiment".breakends.bed | cut -d' ' -f1 >>
"$summaryfile"

echo "Number of tagged break-ends on the forward strand:" >>
"$summaryfile"
awk '{ if ($6 == "+") print}' "$BED"/"$experiment".breakends.bed | wc
-l >> "$summaryfile"

echo "Number of tagged break-ends on the reverse strand:" >>
"$summaryfile"
awk '{ if ($6 == "-") print}' "$BED"/"$experiment".breakends.bed | wc
-l >> "$summaryfile"

echo "Number of DSB locations:" >> "$summaryfile"
wc -l "$BED"/"$experiment".breakcount.bed | cut -d' ' -f1 >>
"$summaryfile"

echo "Average break frequency at unique sites:" >> "$summaryfile"
```

```
awk '{x+=$5} END{print x/NR}' "$BED"/"$experiment".breakcount.bed >>
"$summaryfile"

echo "Histogram frequency at unique sites:" >> "$summaryfile"
awk '{printf "%0.0f\n", $5 }' "$BED"/"$experiment".breakcount.bed |
sort -k1,1n | uniq -c | awk '{print $2,$1}' >> "$summaryfile"

)
done
```

Sequence around break sites

```
#!/usr/bin/env bash
#TO DO: SET width as variable (50-100bp)
clear
BED=BED
FASTA=FASTA
SEQUENCES=sequences
refseq=/home/postdoc/bwa_refseq/Homo_sapiens_assembly19.fa
BACKGROUND=background.txt

# both strands breakcounts 50bp
for file in */preprocessing/"$BED"/*.breakcount.bed; do (
filename=$(echo "$file" | awk -F'[/]' '{print $4}')
experiment=$(echo "$filename" | awk -F'[.]' '{print $1}');
mkdir -p "$experiment"/seq_at_break/"$BED"
postBED="$experiment"/seq_at_break/"$BED"
preBED="$experiment"/preprocessing/"$BED"
mkdir -p "$experiment"/seq_at_break/"$FASTA"
FASTA="$experiment"/seq_at_break/"$FASTA"
mkdir -p "$experiment"/seq_at_break/"$SEQUENCES"
SEQUENCES="$experiment"/seq_at_break/"$SEQUENCES"
echo "Found" "$filename" "Extracting Both Strand Sequences"
awk '{
if ($6 == "+")
print $1,$2-=25,$3+=24,$4,$5,$6;
else
print $1,$2-=24,$3+=25,$4,$5,$6;
}' OFS="\t" "$preBED"/"$experiment".breakcount.bed |
bedtools getfasta -bedOut -fi "$refseq" -s -bed - |
awk '{print $7}' - > "$SEQUENCES"/"$experiment".breakcount.sequences
pr -mts' ' "$preBED"/"$experiment".breakcount.bed
"$SEQUENCES"/"$experiment".breakcount.sequences >
"$postBED"/"$experiment".breakcount.50bp.bed
sed 's/^/>/' "$postBED"/"$experiment".breakcount.50bp.bed | awk '{print
$1,$2,$3,$4,$6,"\n"$7}' > "$FASTA"/"$experiment".breakcount.50bp.fasta
) done

# both strands breakends 50bp
for file in */preprocessing/"$BED"/*.breakends.bed; do (
filename=$(echo "$file" | awk -F'[/]' '{print $4}')
experiment=$(echo "$filename" | awk -F'[.]' '{print $1}');
mkdir -p "$experiment"/seq_at_break/"$BED"
postBED="$experiment"/seq_at_break/"$BED"
preBED="$experiment"/preprocessing/"$BED"
mkdir -p "$experiment"/seq_at_break/"$FASTA"
FASTA="$experiment"/seq_at_break/"$FASTA"
mkdir -p "$experiment"/seq_at_break/"$SEQUENCES"
SEQUENCES="$experiment"/seq_at_break/"$SEQUENCES"
echo "Found" "$filename" "Extracting Both Strand Sequences"
awk '{
if ($6 == "+")
print $1,$2-=25,$3+=24,$4,$5,$6;
else
print $1,$2-=24,$3+=25,$4,$5,$6;
}' OFS="\t" "$preBED"/"$experiment".breakends.bed |
bedtools getfasta -bedOut -fi "$refseq" -s -bed - |
awk '{print $7}' - > "$SEQUENCES"/"$experiment".breakends.sequences
pr -mts' ' "$preBED"/"$experiment".breakends.bed
"$SEQUENCES"/"$experiment".breakends.sequences >
"$postBED"/"$experiment".breakends.50bp.bed
sed 's/^/>/' "$postBED"/"$experiment".breakends.50bp.bed | awk '{print
$1,$2,$3,$4,$6,"\n"$7}' > "$FASTA"/"$experiment".breakends.50bp.fasta
) done
```

Logo generation

```
#!/usr/bin/env bash

mkdir -p Logo_plots
for file in */seq_at_break/FASTA/*.breakends.50bp.fasta; do (
filename=$(echo "$file" | awk -F'[/]' '{print $4}')
experiment=$(echo "$filename" | awk -F'[.]' '{print $1}');
fasta="$experiment"/seq_at_break/FASTA
logo="$fasta"/logo
mkdir -p "$fasta"/logo
weblogo -A DNA -F pdf --composition 'H. sapiens' -n 50 -U 'probability' -c
classic -t "$experiment" -f "$fasta"/"$experiment".breakends.50bp.fasta -o
"$fasta"/logo/"$experiment".breakends.50bp.logo
cp "$fasta"/logo/"$experiment".breakends.50bp.logo
Logo_plots/"$experiment".breakends.50bp.logo
echo "Generated logo for "$experiment""
) done
```

Amplicon sequencing analysis

```
#!/usr/bin/env bash
mkdir -p CRISPResso
amplicons=accessory_files/EMX1_11-20_OT_stranded_amplicon_sequences.txt

nrow=$(wc -l < "$amplicons")
for ((i=1; i<=nrow; i++)); do
  amplicon_nm=$(awk -v a="$i" 'NR==a {print $4}' "$amplicons")
  for fastq in $(ls FASTQ/*.fq.gz | awk '-F[//./_]' '{ print $2,$3,$4 }'
OFS="_" | uniq); do
    fastq_R1=$(echo FASTQ/"$fastq"_R1_val_1.fq.gz)
    fastq_R2=$(echo FASTQ/"$fastq"_R2_val_2.fq.gz)
    if [[ $(find CRISPResso/ -maxdepth 1 -name "Amplicon_$amplicon_nm") &&
$(find CRISPResso/Amplicon_"$amplicon_nm"/ -maxdepth 1 -type d -name
"CRISPResso_on_$fastq*") ]]; then
      echo "Amplicon" "$amplicon_nm" "already analysed for sample" "$fastq"
    else
      echo "Amplicon" "$amplicon_nm" "is being analysed for sample" "$fastq"
      amplicon_seq=$(awk -v a="$i" 'NR==a {print $7}' "$amplicons")
      echo "$amplicon_seq"
      guide_seq=$(awk -v a="$i" 'NR==a {print substr($4,length($4)-24,20)}'
"$amplicons")
      echo "$guide_seq"
      CRISPResso -r1 "$fastq_R1" -r2 "$fastq_R2" -a "$amplicon_seq" -an
"$amplicon_nm" -g "$guide_seq" -gn EMX1_sgRNA -q 30 --
max_paired_end_reads_overlap 151 -o CRISPResso/Amplicon_"$amplicon_nm"
    fi
  done
done
```

Appendix C – Raw data for Chapter 3

Table A3. Raw data for Figure 3.10.

	All library fragments			Sequenceable library fragments			
Sample	**Conc. [pM]**	**Conc. [pg/ul]**	**Total DNA [pg]**	**Conc. [pM]**	**Conc. [pg/ul]**	**Total DNA [pg]**	**% Sequenceable**
WT 1 AD001	8,490	2,170	108,500	0.286	0.071	3.5	0.0033%
WT 2 AD002	8,190	2,200	110,000	0.440	0.109	5.4	0.0049%
WT 3 AD004	3,910	9,98	49,900	0.284	0.070	3.5	0.0070%
WT 4 AD005	4,770	1,420	71,000	0.230	0.065	3.3	0.0046%
WT 5 AD006	5,770	1,510	75,500	0.278	0.069	3.4	0.0046%
HindIII 1 AD012	10,900	2,810	140,500	191.999	47.454	2,372.0	1.6888%
HindIII 2 AD019	10,600	2,780	139,000	211.708	52.326	2,616.0	1.8822%

Table A4. Calculation of Illumina flow cell hybridization efficiency.

	Pass filter clusters		Loading (0.4kb library)			Hybridization efficiency	
Illumina platform	Million (lower)	Million (upper)	Vol (µL)	Conc pM	Molecules (Million)	Lower hybridisation efficiency (%)	Upper hybridisation efficiency (%)
MiSeq (v3)	10	20	600	8	2899.7	0.34	0.69
Miseq nano (v2)	0.5	1	600	8	2899.7	0.02	0.03
NextSeq (High Output)	380	440	1300	1.8	1409.1	26.97	31.22
HiSeq 2500 (v4) 1x lane	240	280	75	18	811.9	29.56	34.49

Appendix D – Additional text and figures for Chapter 5

Additional figures for section 5.2.1 - Validation of EMX1 on-target editing

The ICE analysis tool quantifies the proportion and distribution of indels within the Sanger sequence trace generated from the treated samples (**Figure A1**). This analysis revealed that the edited fraction for both 30h samples was greater than the unedited fraction, with r1 showing a 68% indel frequency (**Figure A2, A**) and r2 59% (**Figure A2, B**). Furthermore, ICE analysis determined the profile of indels within the edited fraction, revealing a 1bp insertion to be the most common indel (47% for r1 and 35% for r2). Deletions for both samples were also similar; 1bp, 3bp and 6bp deletions were present in both r1 and r2 and were detected at similar frequencies. Only a single larger deletion of 26bp, which was detected in r2 was not detected in r1. The generation of similar indel frequencies and profiles at the endpoints of both time courses confirmed that successful and reproducible editing had been achieved in both sets of samples.

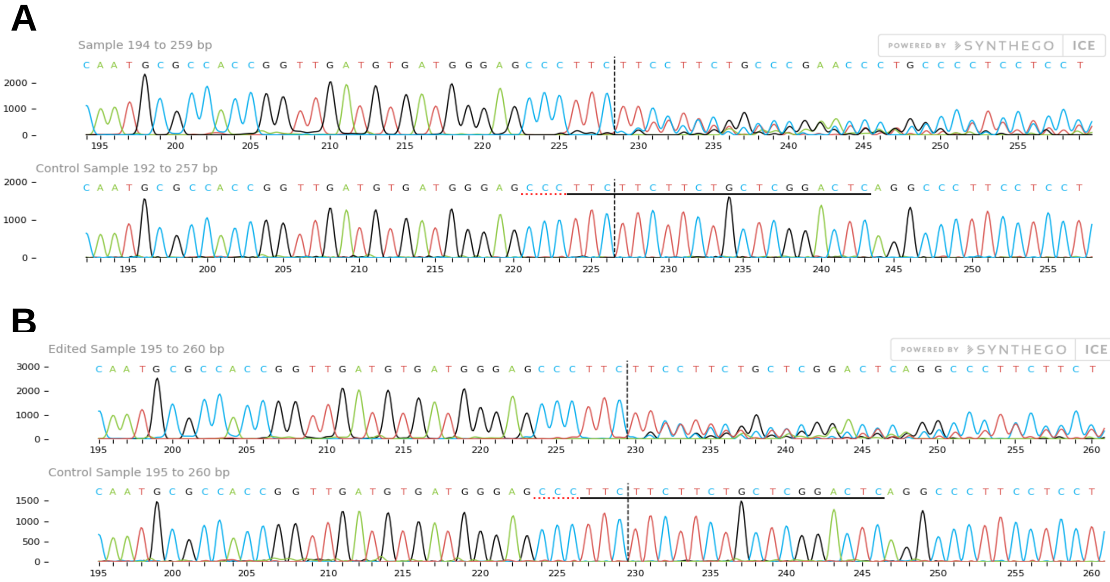


Figure A1. Sanger sequencing traces spanning the EMX1 target site (**A**) Sanger sequencing traces generated from PCR-products amplified from the 30h r1 treated (top panel) and r1 control sample (bottom panel). (**B**) Sanger sequencing traces generated from PCR-products amplified from the 30h r2 treated (top panel) and r2 control sample (bottom panel). The EMX1 target sequences is aligned to the Sanger sequencing traces (horizontal black line and red dotted line) and the predicted EMX1 cleavage site is marked (dashed line).

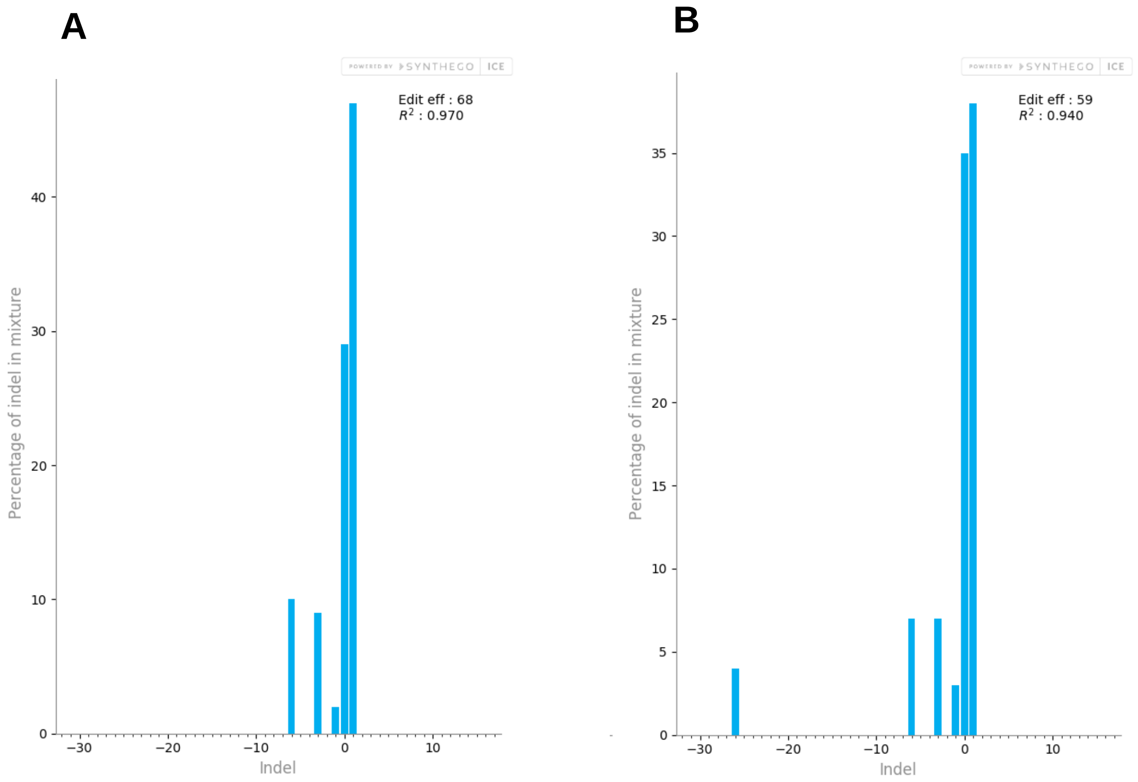


Figure A2. Quantification of indel frequency at the EMX1 on-target site. The quantities of insertions (>0) and deletions (<0) are shown for both r1 30h (a) and r2 30h (b) samples.

Table A5. Raw data for Figure 5.11 and Figure 6.1.

INDUCE-seq identified on- and off-targets from r1 and r2 experiments					Break number at each time point (hours)					Breaks per million reads at each time point (hours)					Also found using (1 = yes, 0 = no)					Editing outcome at each time point (hours)					
																									(Vakulskas et al. 2018)
Coord (GRCh37/hg19)	Sequence	Mismatch	Strand	Total breaks	0	7	12	24	30	0	7	12	24	30	CIRCLE-seq	Digenome-seq	GUIDE-seq	BLISS	HTGTS	0	7	12	24	30	48
2:73160981-73161004	GAGTCCGAGCAGAAGAAGAANGG	0	+	971	74	474	277	89	57	23.34	43.50	30.22	26.80	12.19	1	1	1	1	1	0	13.05	27.81	45.37	66.50	94.89
5:45359060-45359083	GAGTTAGAGCAGAAGAAGAANGG	2	-	486	59	202	139	47	39	18.61	18.54	15.16	14.15	8.34	1	1	1	1	1	0	9.43	14.83	26.04	46.67	27.20
15:44109746-44109769	GAGTCTAAGCAGAAGAAGAANAG	3	+	48	3	28	9	5	3	0.95	2.57	0.98	1.51	0.64	1	1	1	1	1	0	0	0.77	0.52	2.04	6.62
6:9118792-9118815	ACGTCTGAGCAGAAGAAGAANGG	3	-	35	7	16	10	2	0	2.21	1.47	1.09	0.60	0	1	1	1	1	1	0	0	0	0	0	0
2:219845055-219845078	GAGGCCGAGCAGAAGAAAGANGG	3	+	25	2	5	10	3	5	0.63	0.46	1.09	0.90	1.07	1	1	1	1	1	0	0	0	0.19	0.16	0.37
8:128801241-128801264	GAGTCCTAGCAGGAGAAGAANAG	3	+	15	1	7	4	2	1	0.32	0.64	0.44	0.60	0.21	1	1	1	1	1	0	0	0	0	0	0.12
5:9227145-9227168	AAGTCTGAGCACAAGAAGAANGG	3	+	13	2	7	1	1	2	0.63	0.64	0.11	0.30	0.43	1	1	1	1	1	0	0	0	0	0.18	0.34
19:1438808-1438831	GAAGTAGAGCAGAAGAAGAANCG	5	+	13	6	3	4	0	0	1.89	0.28	0.44	0	0	1	1	0	0	0	0	0	0	0	0	0
5:146833183-146833206	GAGCCGGAGCAGAAGAAGGANGG	3	-	12	1	4	6	1	0	0.32	0.37	0.65	0.30	0	1	1	1	0	1	0	0	0	0	0	0
3:95690179-95690202	TCATCCAAGCAGAAGAAGAANAG	5	-	12	4	6	0	0	2	1.26	0.55	0	0	0.43	1	1	1	0	0	0	0	0	0	0	0
1:234492858-234492881	GAAGTAGAGCAGAAGAAGAANCG	5	-	11	6	5	0	0	0	1.89	0.46	0	0	0	1	1	0	0	0	0	0	0	0	0	0
1:151027591-151027614	TTCTCCAAGCAGAAGAAGAANAG	5	-	7	2	5	0	0	0	0.63	0.46	0	0	0	1	1	0	0	0	0	0	0	0	0	0
X:53467704-53467727	GAGTCCGGGAAGGAGAAGAANGG	3	-	6	2	1	0	0	3	0.63	0.09	0	0	0.64	1	0	1	1	1	0	0	0	0	0	0
19:24250496-24250519	GAGTCCAAGCAGTAGAGGAANGG	3	-	5	0	2	3	0	0	0	0.18	0.33	0	0	1	1	0	0	0	0	0	0	0	0	0
3:5031597-5031620	GAATCCAAGCAGGAGAAGAANGA	4	+	4	0	1	2	1	0	0	0.09	0.22	0.30	0	1	1	1	0	0	0	0	0	0	0	0
1:231750724-231750747	GAGTCAGAGCAAAAGAAGTANTG	4	+	3	1	2	0	0	0	0.32	0.18	0	0	0	1	1	0	0	0	0	0	0	0	0	0
6:110491396-110491419	AAGTCAGAGCAGAAAAAGAGNGG	4	+	3	0	3	0	0	0	0	0.28	0	0	0	0	1	0	0	1	0	0	0	0	0	0
2:172374197-172374220	GAAGTAGAGCAGAAGAAGAANCG	5	-	3	0	3	0	0	0	0	0.28	0	0	0	1	1	0	0	0	0	0	0	0	0	0
1:33606473-33606496	GAGCCTGAGCAGAAGGAGAANGG	3	-	2	1	0	1	0	0	0.32	0	0.11	0	0	1	1	0	1	0	0	0	0	0	0	0
14:48932102-48932125	GAGTCCCAGCAAAAGAAGAANAG	3	+	2	0	1	1	0	0	0	0.09	0.11	0	0	1	1	0	0	0	0	0	0	0	0	0
11:43747931-43747954	AAGCCCGAGCAAAGGAAGAANGG	4	+	2	0	2	0	0	0	0	0.18	0	0	0	1	1	0	0	1	0	0	0	0	0	0
14:43156800-43156823	GAGGCCAAGCAGAAAAAAAANGG	4	-	2	0	2	0	0	0	0	0.18	0	0	0	1	1	0	0	0	0	0	0	0	0	0
15:100292461-100292484	AAGTCCCGGCAGAGGAAGAANGG	4	+	2	0	2	0	0	0	0	0.18	0	0	0	1	1	1	1	0	0	0	0	0	0	0

INDUCE-seq identified on- and off-targets from r1 and r2 experiments					Break number at each time point (hours)					Breaks per million reads at each time point (hours)					Also found using (1 = yes, 0 = no)					Editing outcome at each time point (hours)					
Coord (GRCh37/hg19)	Sequence	Mismatch	Strand	Total breaks	0	7	12	24	30	0	7	12	24	30	CIRCLE-seq	Digenome-seq	GUIDE-seq	BLISS	HTGTS	0	7	12	24	30	(Vakulskas et al. 2018) 48
1:59902174-59902197	GAGCCAGGGCAGAAGAAGAANGA	4	-	2	0	2	0	0	0	0	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17:78609059-78609082	GAGCCCGTGCAGAGGAAGAANGA	4	-	2	0	0	2	0	0	0	0	0.22	0	0	1	0	0	0	0	0	0	0	0	0	0
3:63468110-63468133	AAGTTGGAGCAGGAGAAGAANGG	4	+	2	1	0	0	1	0	0.32	0	0	0.30	0	1	0	0	0	0	0	0	0	0	0	0
7:141972555-141972578	AAGTCCGGGCAAAAGAGGAANGG	4	-	2	0	0	1	1	0	0	0	0.11	0.30	0	1	1	0	0	0	0	0	0	0	0	0
8:82733133-82733156	GAGTCAGAGAAGAGGAGGAANGG	4	-	2	0	1	0	1	0	0	0.09	0	0.30	0	0	0	0	0	0	0	0	0	0	0	0
10:91416144-91416167	ATGTCCAAGCAGAAGAAGTCNGG	5	+	2	0	2	0	0	0	0	0.18	0	0	0	1	1	0	0	0	0	0	0	0	0	0
16:73177721-73177744	TCTTCCGAGCTGAAGAAGAANAG	5	+	2	0	2	0	0	0	0	0.18	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2:216731830-216731853	GGGTCAGAGAAGGAGAAGATNGG	5	-	2	0	2	0	0	0	0	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2:54284994-54285017	AAGGCAGAGCAGAGGAAGAGNGG	5	+	2	2	0	0	0	0	0.63	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
12:119985924-119985947	GACTCCTAGCAAAAGAAGAANGG	3	-	1	0	0	0	0	1	0	0	0	0	0.21	1	0	0	0	0	0	0	0	0	0	0
1:225534266-225534289	GATTCCTACCAGAAGAAGAANGG	3	+	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1:23720611-23720634	AAGTCCGAGGAGAGGAAGAANGG	3	-	1	0	1	0	0	0	0	0.09	0	0	0	1	1	1	1	1	0	0	0	0	0	0
15:61646860-61646883	AAGTCAGAGGAGAAGAAGAANGG	3	+	1	1	0	0	0	0	0.32	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4:87256685-87256708	GAGTAAGAGAAGAAGAAGAANGG	3	-	1	1	0	0	0	0	0.32	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
10:5401770-5401793	TAATCCAATCAGAAGAAGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	1	1	0	0	0	0	0	0	0	0	0
10:58848711-58848734	GAGCACGAGCAAGAGAAGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	1	1	1	0	0	0	0	0	0	0	0
10:94423051-94423074	GAGTCCTAGTAGAAGAGAAANGG	4	-	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11:26852512-26852535	GAGTATGAGCAGAAGTAGACNGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11:62365266-62365289	GAATCCAAGCAGAAGAAGAGNAG	4	-	1	1	0	0	0	0	0.32	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0
12:104276937-104276960	GAGTCAGAGGAAAAGAAGAANGA	4	+	1	0	0	0	1	0	0	0	0	0.30	0	0	0	0	0	0	0	0	0	0	0	0
12:106646073-106646096	AAGTCCATGCAGAAGAGGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	1	1	0	0	0	0	0	0	0	0	0
13:31765208-31765231	AAGTCCCAGCCGAGGAAGAANGG	4	-	1	0	0	0	1	0	0	0	0	0.30	0	0	0	0	0	0	0	0	0	0	0	0
1:35818885-35818908	TATACGGAGCAGAAGAAGAANGG	4	-	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14:50538462-50538485	TAGTCCTAGCAAAAGCAGAANGG	4	-	1	0	1	0	0	0	0	0.09	0	0	0	1	0	0	0	0	0	0	0	0	0	0

INDUCE-seq identified on- and off-targets from r1 and r2 experiments					Break number at each time point (hours)					Breaks per million reads at each time point (hours)					Also found using (1 = yes, 0 = no)					Editing outcome at each time point (hours)					
Coord (GRCh37/hg19)	Sequence	Mismatch	Strand	Total breaks	0	7	12	24	30	0	7	12	24	30	CIRCLE-seq	Digenome-seq	GUIDE-seq	BLISS	HTGTS	0	7	12	24	30	(Vakulskas et al. 2018) 48
14:70375672-70375695	GAGGCAGAGAAGAAGAAGAGNGG	4	-	1	1	0	0	0	0	0.32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15:91761953-91761976	GAGTCAGGGCAGAAGAAGAANAT	4	-	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16:85121027-85121050	CAGTCAGGGCAGAGGAAGAANGG	4	+	1	0	0	0	0	1	0	0	0	0	0.21	0	0	0	0	0	0	0	0	0	0	0
18:7515889-7515912	GAAACCAAGCAGAAGAGGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2:219824755-219824778	GCGTCCGCCAAGAAGAAGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2:230180651-230180674	TATTCAGAGCTGAAGAAGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3:123771739-123771762	CATTCCTAGCAGAGGAAGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5:36616618-36616641	AAGTCTGAGGACAAGAAGAANGG	4	-	1	0	0	1	0	0	0	0	0.11	0	0	0	0	0	0	0	0	0	0	0	0	0
6:158029597-158029620	GAGCCCGGGCAGGAGAAGATNGG	4	+	1	0	0	0	1	0	0	0	0	0.30	0	0	0	0	0	0	0	0	0	0	0	0
6:167069776-167069799	GAGGCAGGGAAGAAGAAGAANGG	4	+	1	0	1	0	0	0	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7:149462283-149462306	CAGTCCGGGCAGAAGAAGGANCG	4	+	1	0	0	0	0	1	0	0	0	0	0.21	0	0	0	0	0	0	0	0	0	0	0
8:10048559-10048582	TAGTCTAAGCAGCAGAAGAANGG	4	-	1	1	0	0	0	0	0.32	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
8:105164108-105164131	GAGCCCAAGAAGAAGAAGAANGA	4	+	1	0	0	0	0	1	0	0	0	0	0.21	1	1	0	0	0	0	0	0	0	0	0
8:109199391-109199414	GAGTCAGAGCAGAAGAAAGANGA	4	-	1	0	1	0	0	0	0	0.09	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Additional figures for section 5.2.8. - Characterising the mismatches at EMX1 off-targets.

Quantification of different mismatch frequency across the EMX1 off-target sites at each time point.

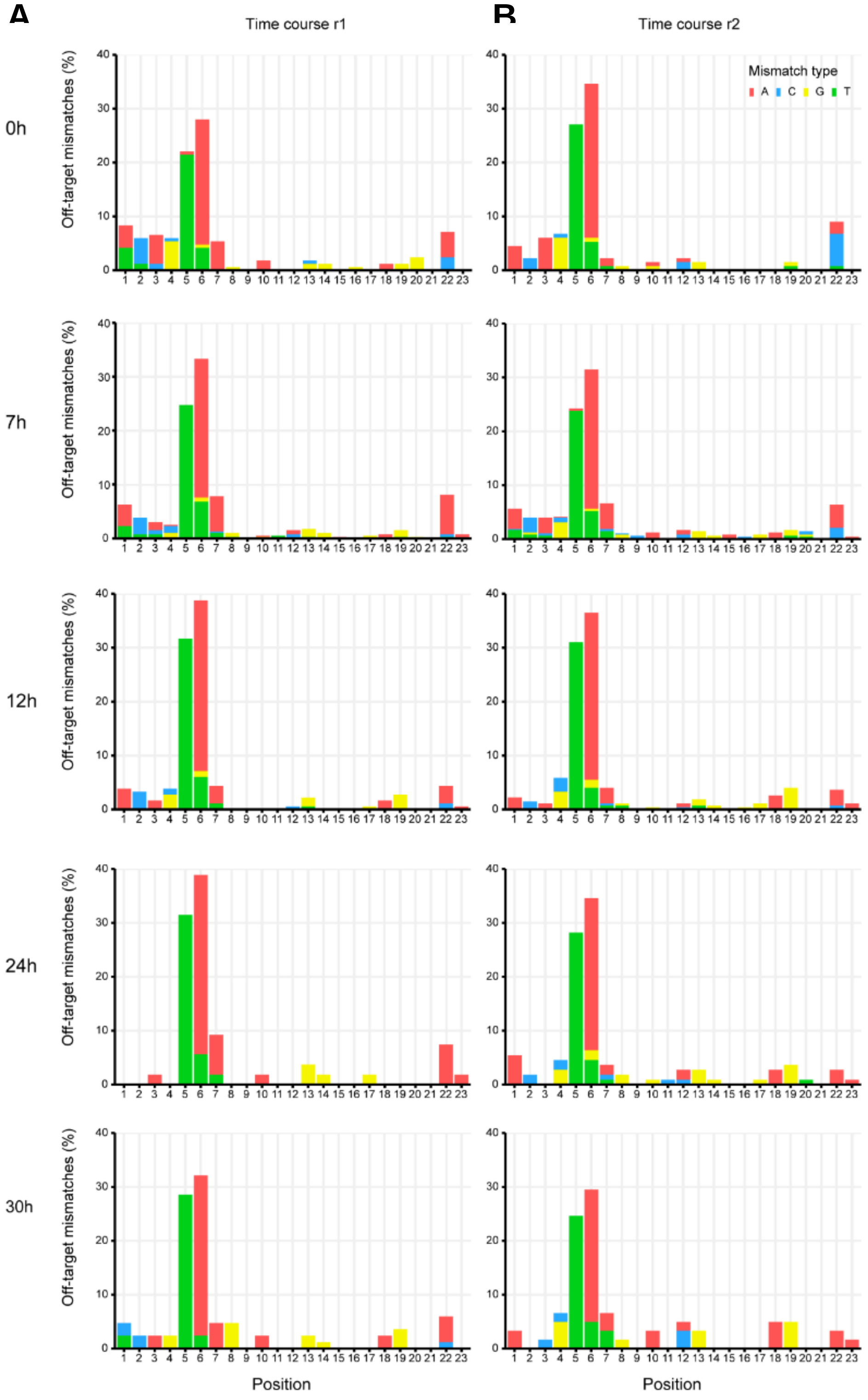


Figure A3. The mismatch profile at EMX1 off-targets at each of the time point samples. (A) and **(B)** Frequency of mismatches found at off-target sites across the EMX1 target sequence for each of the different time points for r1 (**A**) and r2 (**B**).

Full description of Figure 5.14.

When considering the full target sequence (**Figure 5.14A, pink bars**), fewer breaks are identified at off-targets with decreasing mismatches number. Outside of OT1 at 2 mismatches, sites with 3 mismatches account for ~20% of the breaks identified at off-target sites, whereas sites with 4 and 5 mismatches each account for ~8%. Interestingly, an inverse relationship becomes obvious when considering off-target site frequency: generally, more off-target sites are identified with more mismatches, peaking at ~50% of sites containing 4 mismatches in total. Interestingly, significantly fewer sites are identified with 5 mismatches despite the fact that an equivalent number of breaks are found at sites with 4 and 5 mismatches.

To better understand which of the mismatch groups were contributing most to off-target site selection, the spacer distal (**Figure 5.14, green bars**), spacer seed (**Figure 5.14, blue bars**), and PAM region (**Figure 5.14, purple bars**), were also plotted in an equivalent manner. As with the full sequence, when looking at break frequency, most breaks originate from a single off-target, OT1, which possesses 2 mismatches in the spacer distal region and 0 mismatches in the spacer seed and PAM. Interestingly, not a single off-target was identified without mismatches in the 8 bp spacer distal region, and only ~20% of the off targets, which accounted for ~10% of off-target breaks, had just 1 mismatch in the spacer distal region (**Figure 5.14, green bars**). Correspondingly, most sites contained ≥2 mismatches in the spacer distal region, with ~70% of sites containing 2 or 3 mismatches. In contrast to the distal region, the majority of breaks were identified at sites with 0 mismatches in the seed region and the PAM, both at ~80%. The contribution of OT1, with only two mismatches in the spacer distal, is significant and can easily be accounted for when interpreting break activity across the remaining off-targets. Removing the ~60% break contribution from OT1 means that for all other off-target breaks, approximately 50% have 0 mismatches in the seed region and 50% ≥1 (20% each of the total) (**Figure 5.14, blue bars**). Given this, and the fact that ~80% of sites seemed to contain some form of spacer seed mismatch, it suggests that even outside of OT1 high breakage occurs preferentially at sites with no seed mismatches, agreeing with that observed in Figure 5.13 and 5.15. A similar pattern can be observed with PAM mismatches (**Figure 5.14 purple bars**), although fewer mismatch variations are possible because of the 3 bp size. Outside of OT1, approximately 50% of all breaks are at off-targets with a single PAM mismatch (~18% of total). This break frequency occurs at relatively fewer sites than shown for 0 PAM mismatches, (~30% and ~60% of sites, respectively), suggesting that, on average once OT1 has been accounted for, off-targets with a single PAM mismatch are broken approximately 2-fold more often than those without. Taken together, these results show the characteristics of the unique profile of off-targets identified for EMX1 using INDUCE-seq.

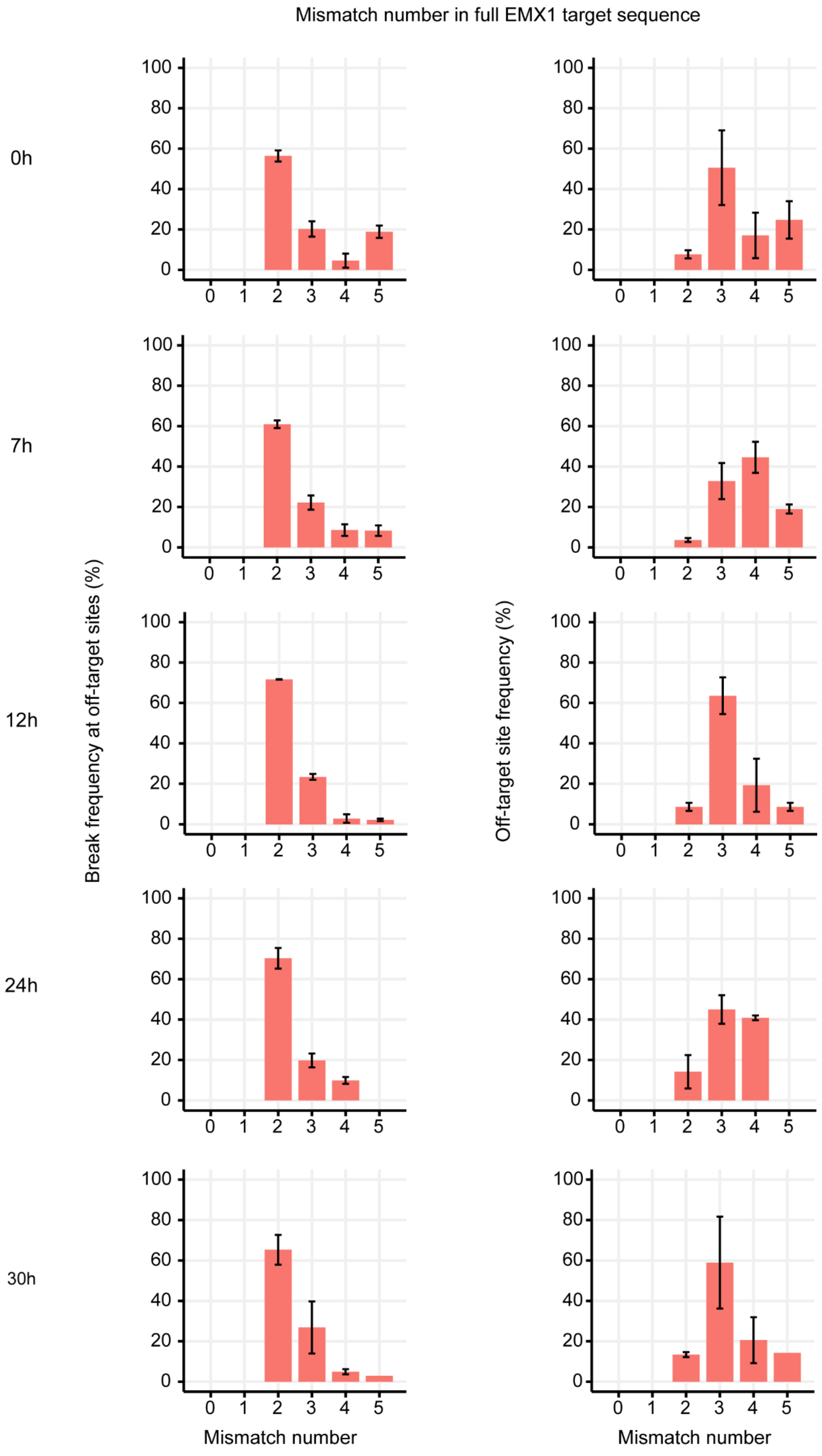

Mismatch number in full EMX1 target sequence
0h
7h
12h
24h
30h
Break frequency at off-target sites (%)
Off-target site frequency (%)
100
80
60
40
20
0
0
1
2
3
4
5
Mismatch number
Mismatch number

Figure A4. The quantification of EMX1 off-target mismatches across the full target sequence at each time point. Error bars as SD, n=2.

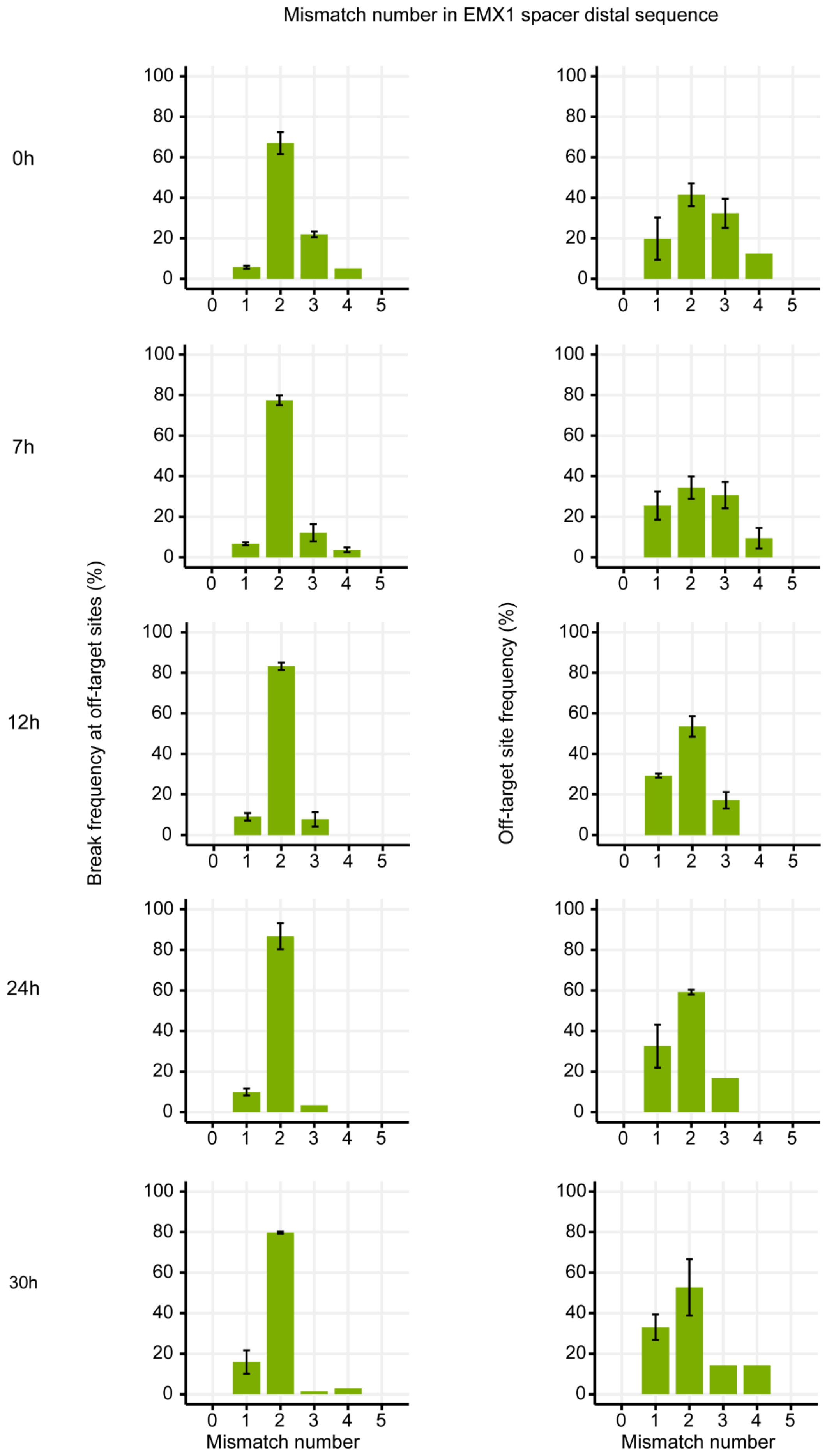

Mismatch number in EMX1 spacer distal sequence
0h
7h
12h
24h
30h
Break frequency at off-target sites (%)
Off-target site frequency (%)
100
80
60
40
20
0
0 1 2 3 4 5
Mismatch number
Mismatch number

Figure A5. The quantification of EMX1 off-target mismatches across the spacer distal sequence at each time point. Error bars as SD, n=2.

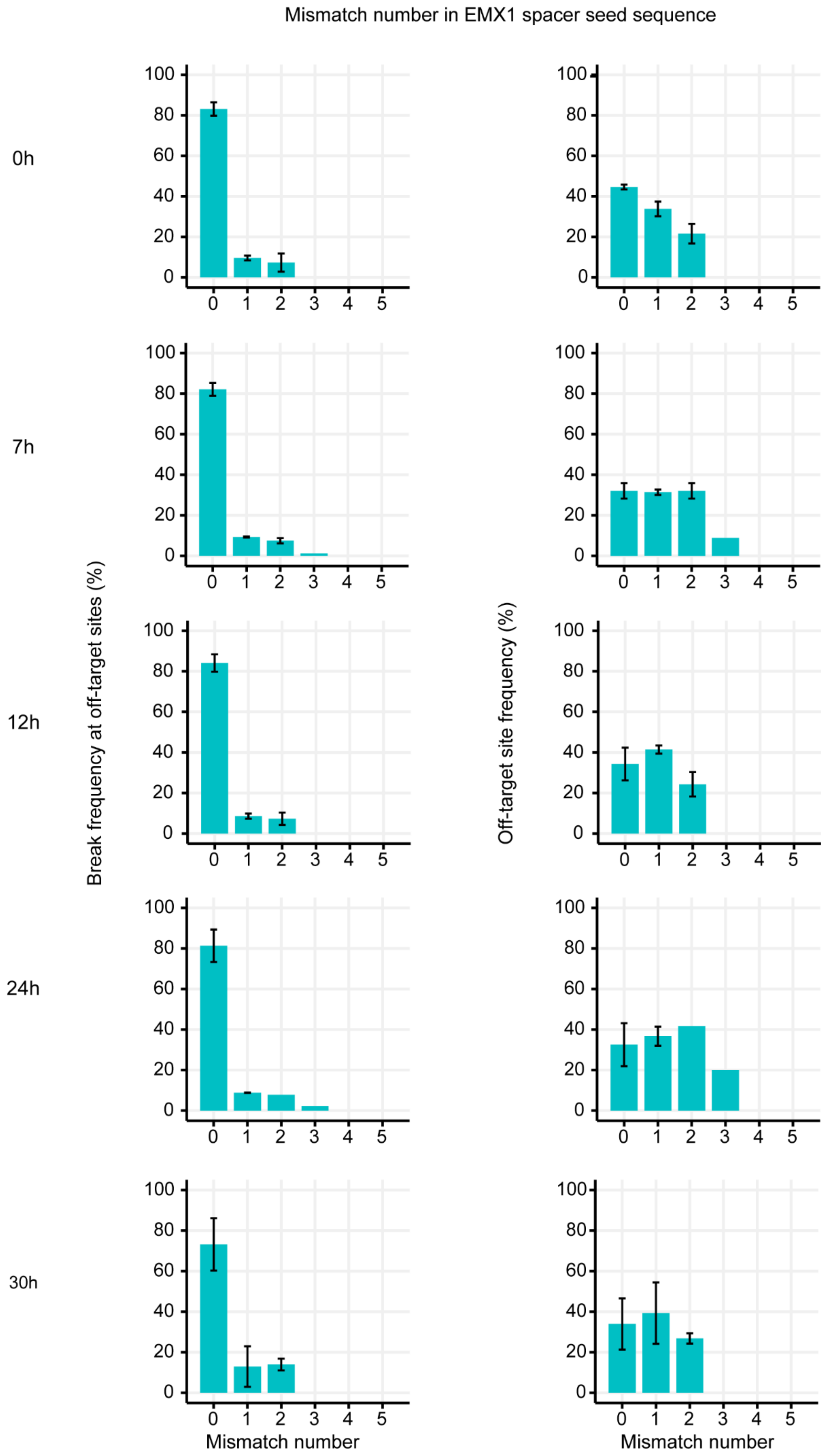

Mismatch number in EMX1 spacer seed sequence
0h
7h
12h
24h
30h
Break frequency at off-target sites (%)
Off-target site frequency (%)
100
80
60
40
20
0
0
1
2
3
4
5
Mismatch number
Mismatch number

Figure A6. The quantification of EMX1 off-target mismatches across the spacer seed sequence at each time point. Error bars as SD, n=2.

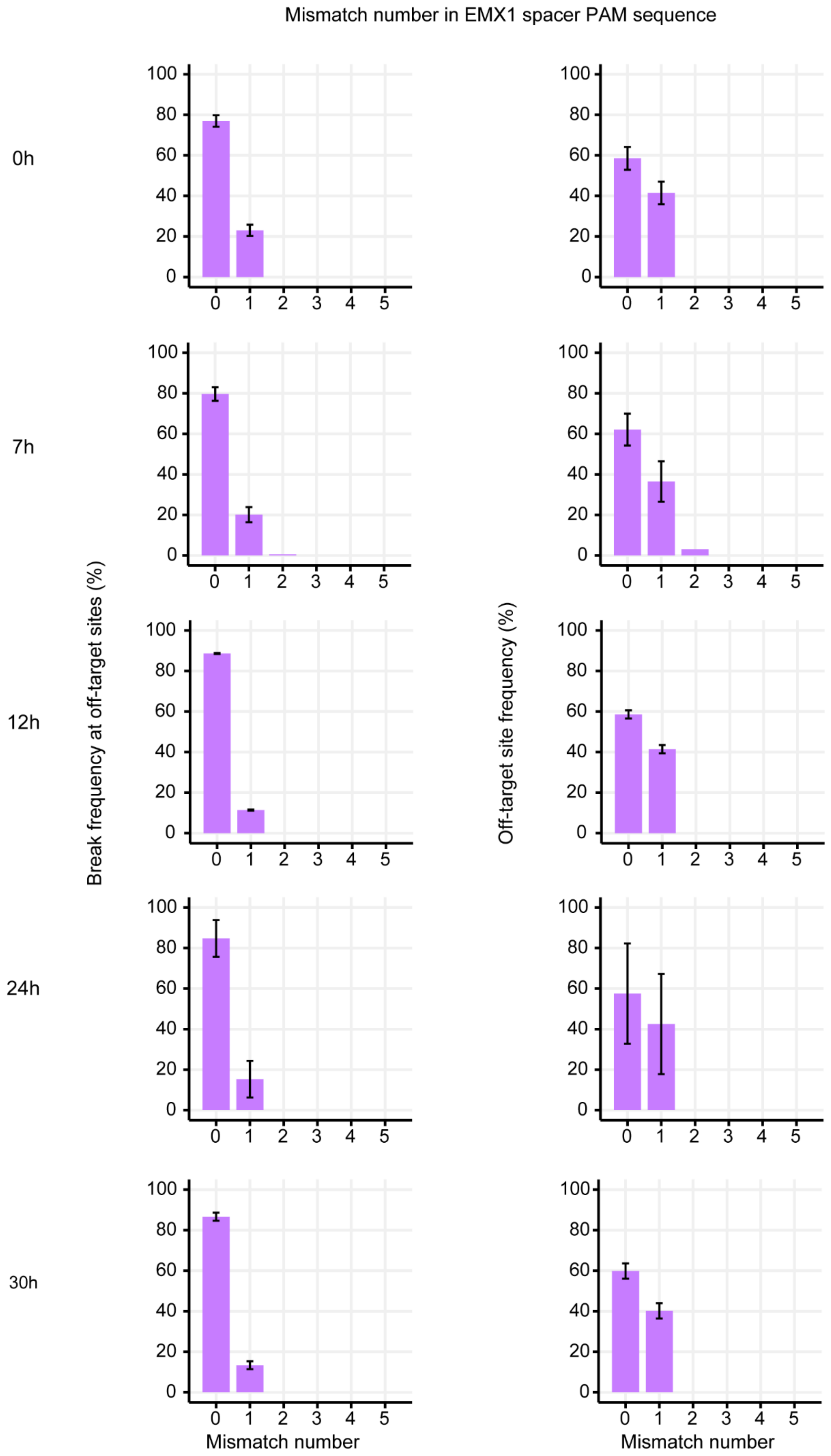

Mismatch number in EMX1 spacer PAM sequence
0h
7h
12h
24h
30h
Break frequency at off-target sites (%)
Off-target site frequency (%)
0
20
40
60
80
100
0 1 2 3 4 5
Mismatch number
Mismatch number

Figure A7. The quantification of EMX1 off-target mismatches across the PAM sequence at each time point. Error bars as SD, n=2.

Appendix E – Additional figures for Chapter 6

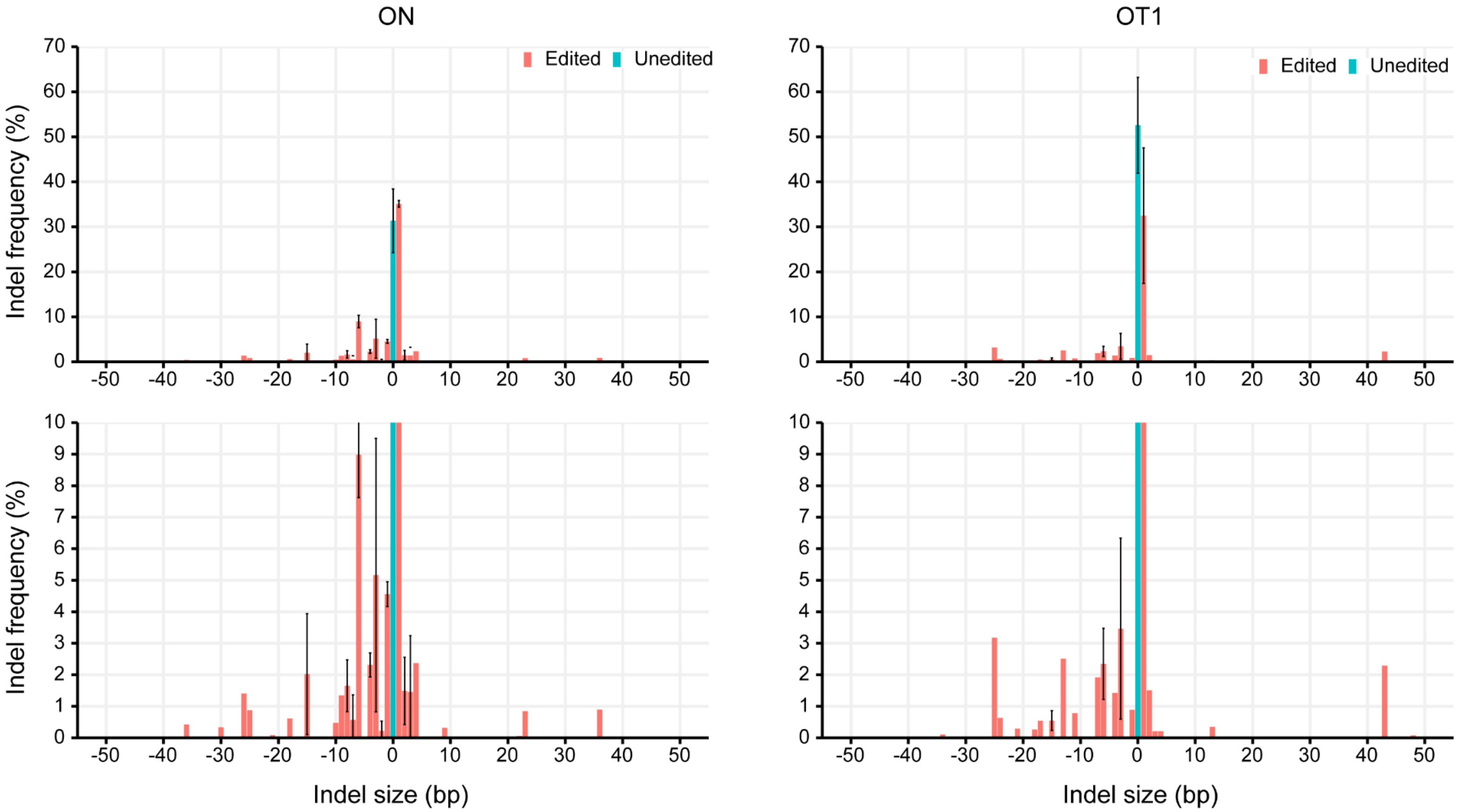


Figure A8. Detailed comparison of the indel frequencies measured at the EMX1 on-target and OT1 at 30h. Error bars as SD, n=2.

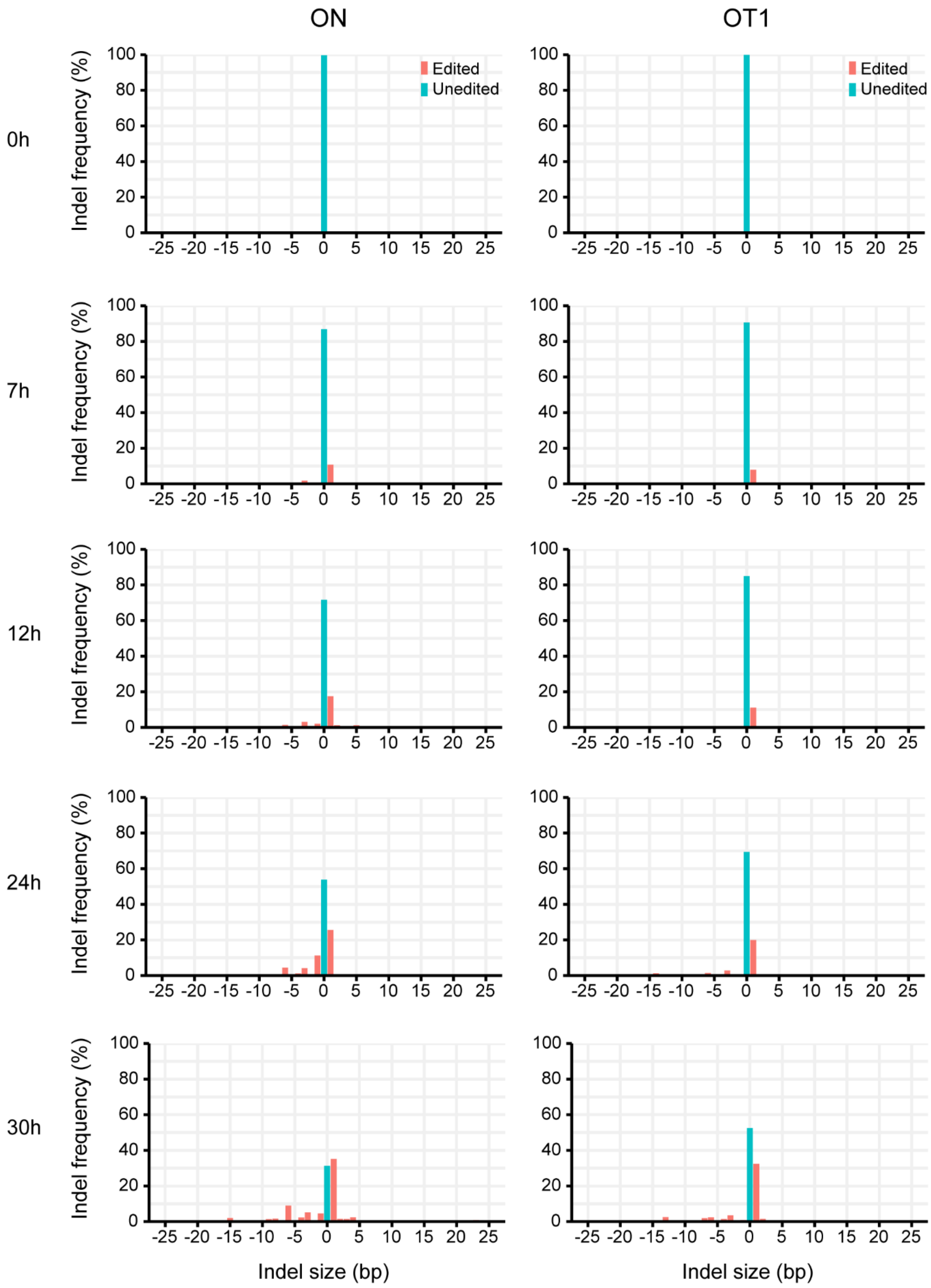


Figure A9. Comparison of the indel frequencies measured at the EMX1 on-target and OT1 across each of the time points.

9 Bibliography

Abadi, S. et al. 2017. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *Plos Computational Biology* 13(10), doi: 10.1371/journal.pcbi.1005807

Aird, D. et al. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* 12(2), doi: 10.1186/gb-2011-12-2-r18

Akcakaya, P. et al. 2018. In vivo CRISPR editing with no detectable genome-wide off-target mutations. *Nature* 561(7723), pp. 416-+. doi: 10.1038/s41586-018-0500-9

Allen, F. et al. 2019. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nature Biotechnology* 37(1), pp. 64-+. doi: 10.1038/nbt.4317

Alt, F. W. and Schwer, B. 2018. DNA double-strand breaks as drivers of neural genomic change, function, and disease. *DNA Repair* 71, pp. 158-163. doi: 10.1016/j.dnarep.2018.08.019

Anders, C. et al. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* 513(7519), pp. 569-+. doi: 10.1038/nature13579

Anzalone, A. V. et al. 2020. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology* 38(7), pp. 824-844. doi: 10.1038/s41587-020-0561-9

Anzalone, A. V. et al. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576(7785), pp. 149-+. doi: 10.1038/s41586-019-1711-4

Aymard, F. et al. 2017. Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes. *Nature Structural & Molecular Biology* 24(4), pp. 353-+. doi: 10.1038/nsmb.3387

Bae, S. et al. 2014. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* 30(10), pp. 1473-1475. doi: 10.1093/bioinformatics/btu048

Ballinger, T. J. et al. 2019. Modeling double strand break susceptibility to interrogate structural variation in cancer. *Genome Biology* 20, doi: 10.1186/s13059-019-1635-1

Baranello, L. et al. 2014. DNA Break Mapping Reveals Topoisomerase II Activity Genome-Wide. *International Journal of Molecular Sciences* 15(7), pp. 13111-13122. doi: 10.3390/ijms150713111

Barrangou, R. et al. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315(5819), pp. 1709-1712. doi: 10.1126/science.1138140

Bennett, E. P. et al. 2020. INDEL detection, the 'Achilles heel' of precise genome editing: a survey of methods for accurate profiling of gene editing induced indels. *Nucleic Acids Research* 48(21), pp. 11958-11981. doi: 10.1093/nar/gkaa975

Bentley, D. R. et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218), pp. 53-59. doi: 10.1038/nature07517

Bi, C. W. et al. 2020. Long-read individual-molecule sequencing reveals CRISPR-induced genetic heterogeneity in human ESCs. *Genome Biology* 21(1), doi: 10.1186/s13059-020-02143-8

Biernacka, A. et al. 2018. i-BLESS is an ultra-sensitive method for detection of DNA double-strand breaks. *Communications Biology* 1, doi: 10.1038/s42003-018-0165-9

Bodnar, A. G. et al. 1998. Extension of life-span by introduction of telomerase into normal human cells. *Science* 279(5349), pp. 349-352. doi: 10.1126/science.279.5349.349

Bolotin, A. et al. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology-Sgm* 151, pp. 2551-2561. doi: 10.1099/mic.0.28048-0

Bouwman, B. A. M. and Crosetto, N. 2018. Endogenous DNA Double-Strand Breaks during DNA Transactions: Emerging Insights and Methods for Genome-Wide Profiling. *Genes* 9(12), doi: 10.3390/genes9120632

Cameron, P. et al. 2017. Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nature Methods* 14(6), pp. 600-+. doi: 10.1038/nmeth.4284

Canela, A. et al. 2016. DNA Breaks and End Resection Measured Genome-wide by End Sequencing. *Molecular Cell* 63(5), pp. 898-911. doi: 10.1016/j.molcel.2016.06.034

Ceccaldi, R. et al. 2016. Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends in Cell Biology* 26(1), pp. 52-64. doi: 10.1016/j.tcb.2015.07.009

Chapman, J. R. et al. 2012. Playing the End Game: DNA Double-Strand Break Repair Pathway Choice. *Molecular Cell* 47(4), pp. 497-510. doi: 10.1016/j.molcel.2012.07.029

Chaudhari, H. G. et al. 2020. Evaluation of Homology-Independent CRISPR-Cas9 Off-Target Assessment Methods. *Crispr Journal* 3(6), pp. 440-453. doi: 10.1089/crispr.2020.0053

Cheng, Y. and Tsai, S. Q. 2018. Illuminating the genome-wide activity of genome editors for safe and effective therapeutics. *Genome Biology* 19, doi: 10.1186/s13059-018-1610-2

Chiarle, R. et al. 2011. Genome-wide Translocation Sequencing Reveals Mechanisms of Chromosome Breaks and Rearrangements in B Cells. *Cell* 147(1), pp. 107-119. doi: 10.1016/j.cell.2011.07.049

Cho, S. W. et al. 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology* 31(3), pp. 230-232. doi: 10.1038/nbt.2507

Cho, S. W. et al. 2014. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Research* 24(1), pp. 132-141. doi: 10.1101/gr.162339.113

Choulika, A. et al. 1995. INDUCTION OF HOMOLOGOUS RECOMBINATION IN MAMMALIAN CHROMOSOMES BY USING THE I-SCEI SYSTEM OF SACCHAROMYCES-CEREVISIAE. *Molecular and Cellular Biology* 15(4), pp. 1968-1973.

Christian, M. et al. 2010. Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. *Genetics* 186(2), pp. 757-U476. doi: 10.1534/genetics.110.120717

Chuai, G. H. et al. 2018. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biology* 19, doi: 10.1186/s13059-018-1459-4

Cohen-Tannoudji, M. et al. 1998. I-SceI-induced gene replacement at a natural locus in embryonic stem cells. *Molecular and Cellular Biology* 18(3), pp. 1444-1448. doi: 10.1128/mcb.18.3.1444

Cong, L. et al. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* 339(6121), pp. 819-823. doi: 10.1126/science.1231143

Cromwell, C. R. et al. 2018. Incorporation of bridged nucleic acids into CRISPR RNAs improves Cas9 endonuclease specificity. *Nature Communications* 9, doi: 10.1038/s41467-018-03927-0

Crooks, G. E. et al. 2004. WebLogo: A sequence logo generator. *Genome Research* 14(6), pp. 1188-1190. doi: 10.1101/gr.849004

Crosetto, N. et al. 2013. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature Methods* 10(4), pp. 361-+.

Dabney, J. and Meyer, M. 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* 52(2), pp. 87-+. doi: 10.2144/000113809

Deltcheva, E. et al. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471(7340), pp. 602-+. doi: 10.1038/nature09886

Densham, R. M. and Morris, J. R. 2019. Moving Mountains-The BRCA1 Promotion of DNA Resection. *Frontiers in Molecular Biosciences* 6, doi: 10.3389/fmolb.2019.00079

Doudna, J. A. and Charpentier, E. 2014. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346(6213), pp. 1077-+. doi: 10.1126/science.1258096

Ernst, M. P. T. et al. 2020. Ready for Repair? Gene Editing Enters the Clinic for the Treatment of Human Disease. *Molecular Therapy-Methods & Clinical Development* 18, pp. 532-557. doi: 10.1016/j.omtm.2020.06.022

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 8(3), pp. 186-194. doi: 10.1101/gr.8.3.186

Fellows, M. D. 2016. Targeting safety in the clinic for precise genome editing using CRISPR: a genotoxicologist's perspective. *Personalized Medicine* 13(4), pp. 279-282. doi: 10.2217/pme-2016-0026

Fernandez, A. et al. 2017. A history of genome editing in mammals. *Mammalian Genome* 28(7-8), pp. 237-246. doi: 10.1007/s00335-017-9699-2

Fisher, S. et al. 2011. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology* 12(1), doi: 10.1186/gb-2011-12-1-r1

Frock, R. L. et al. 2015. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nature Biotechnology* 33(2), pp. 179-186. doi: 10.1038/nbt.3101

Fu, Y. F. et al. 2013. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology* 31(9), pp. 822-+. doi: 10.1038/nbt.2623

Fu, Y. F. et al. 2014. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature Biotechnology* 32(3), pp. 279-284. doi: 10.1038/nbt.2808

Fujita, T. et al. 2016. Allele-specific locus binding and genome editing by CRISPR at the p16INK4a locus. *Scientific Reports* 6, doi: 10.1038/srep30485

Gabriel, R. et al. 2011. An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nature Biotechnology* 29(9), pp. 816-U872. doi: 10.1038/nbt.1948

Geurts, A. M. et al. 2009. Knockout Rats via Embryo Microinjection of Zinc-Finger Nucleases. *Science* 325(5939), pp. 433-433. doi: 10.1126/science.1172447

Ghezraoui, H. et al. 2014. Chromosomal Translocations in Human Cells Are Generated by Canonical Nonhomologous End-Joining. *Molecular Cell* 55(6), pp. 829-842. doi: 10.1016/j.molcel.2014.08.002

Goodwin, J. F. and Knudsen, K. E. 2014. Beyond DNA Repair: DNA-PK Function in Cancer. *Cancer Discovery* 4(10), pp. 1126-1139. doi: 10.1158/2159-8290.cd-14-0358

Haft, D. H. et al. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *Plos Computational Biology* 1(6), pp. 474-483. doi: 10.1371/journal.pcbi.0010060

Hess, J. F. et al. 2020. Library preparation for next generation sequencing: A review of automation strategies. *Biotechnology Advances* 41, doi: 10.1016/j.biotechadv.2020.107537

Hockemeyer, D. et al. 2011. Genetic engineering of human pluripotent cells using TALE nucleases. *Nature Biotechnology* 29(8), pp. 731-734. doi: 10.1038/nbt.1927

Hoeijmakers, W. A. M. et al. 2011. Linear amplification for deep sequencing. *Nature Protocols* 6(7), pp. 1026-1036. doi: 10.1038/nprot.2011.345

Hoffman, E. A. et al. 2015. Break-seq reveals hydroxyurea-induced chromosome fragility as a result of unscheduled conflict between DNA replication and transcription. *Genome Research* 25(3), pp. 402-412. doi: 10.1101/gr.180497.114

Hsiau, T. et al. 2019. Inference of CRISPR Edits from Sanger Trace Data. *bioRxiv*, p. 251082. doi: 10.1101/251082

Hsu, P. D. et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* 31(9), pp. 827-+. doi: 10.1038/nbt.2647

Hu, J. Z. et al. 2016. Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nature Protocols* 11(5), pp. 853-871. doi: 10.1038/nprot.2016.043

Hwang, W. Y. et al. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology* 31(3), pp. 227-229. doi: 10.1038/nbt.2501

Iacovoni, J. S. et al. 2010. High-resolution profiling of gamma H2AX around DNA double strand breaks in the mammalian genome. *Embo Journal* 29(8), pp. 1446-1457. doi: 10.1038/emboj.2010.38

Ishino, Y. et al. 1987. NUCLEOTIDE-SEQUENCE OF THE IAP GENE, RESPONSIBLE FOR ALKALINE-PHOSPHATASE ISOZYME CONVERSION IN ESCHERICHIA-COLI, AND IDENTIFICATION OF THE GENE-PRODUCT. *Journal of Bacteriology* 169(12), pp. 5429-5433. doi: 10.1128/jb.169.12.5429-5433.1987

Jackson, S. P. and Bartek, J. 2009. The DNA-damage response in human biology and disease. *Nature* 461(7267), pp. 1071-1078. doi: 10.1038/nature08467

Jansen, R. et al. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* 43(6), pp. 1565-1575. doi: 10.1046/j.1365-2958.2002.02839.x

Jiang, F. G. and Doudna, J. A. 2017. CRISPR-Cas9 Structures and Mechanisms. *Annual Review of Biophysics, Vol 46* 46, pp. 505-529. doi: 10.1146/annurev-biophys-062215-010822

Jiang, F. G. et al. 2016. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* 351(6275), pp. 867-871. doi: 10.1126/science.aad8282

Jiang, F. G. et al. 2015. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* 348(6242), pp. 1477-1481. doi: 10.1126/science.aab1452

Jinek, M. et al. 2012. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337(6096), pp. 816-821. doi: 10.1126/science.1225829

Jinek, M. et al. 2013. RNA-programmed genome editing in human cells. *Elife* 2, doi: 10.7554/eLife.00471

Jinek, M. et al. 2014. Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. *Science* 343(6176), pp. 1215-+. doi: 10.1126/science.1247997

Jones, M. B. et al. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences of the United States of America* 112(45), pp. 14024-14029. doi: 10.1073/pnas.1519288112

Joung, J. K. and Sander, J. D. 2013. INNOVATION TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology* 14(1), pp. 49-55. doi: 10.1038/nrm3486

Kebschull, J. M. and Zador, A. M. 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research* 43(21), doi: 10.1093/nar/gkv717

Kennedy, S. R. et al. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols* 9(11), pp. 2586-2606. doi: 10.1038/nprot.2014.170

Khanna, K. K. and Jackson, S. P. 2001. DNA double-strand breaks: signaling, repair and the cancer connection. *Nature Genetics* 27(3), pp. 247-254. doi: 10.1038/85798

Kim, D. et al. 2015. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature Methods* 12(3), pp. 237-+. doi: 10.1038/nmeth.3284

Kim, D. and Kim, J. S. 2018. DIG-seq: a genome-wide CRISPR off-target profiling method using chromatin DNA. *Genome Research* 28(12), pp. 1894-1900. doi: 10.1101/gr.236620.118

Kim, S. et al. 2014. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Research* 24(6), pp. 1012-1019. doi: 10.1101/gr.171322.113

Kircher, M. et al. 2011. Addressing challenges in the production and analysis of illumina sequencing data. *Bmc Genomics* 12, doi: 10.1186/1471-2164-12-382

Klein, I. A. et al. 2011. Translocation-Capture Sequencing Reveals the Extent and Nature of Chromosomal Rearrangements in B Lymphocytes. *Cell* 147(1), pp. 95-106. doi: 10.1016/j.cell.2011.07.048

Knott, G. J. and Doudna, J. A. 2018. CRISPR-Cas guides the future of genetic engineering. *Science* 361(6405), pp. 866-869. doi: 10.1126/science.aat5011

Kocak, D. D. et al. 2019. Increasing the specificity of CRISPR systems with engineered RNA secondary structures. *Nature Biotechnology* 37(6), pp. 657-+. doi: 10.1038/s41587-019-0095-1

Komor, A. C. et al. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533(7603), pp. 420-+. doi: 10.1038/nature17946

Koonin, E. V. et al. 2017. Diversity, classification and evolution of CRISPR-Cas systems. *Current Opinion in Microbiology* 37, pp. 67-78. doi: 10.1016/j.mib.2017.05.008

Kosicki, M. et al. 2018. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* 36(8), pp. 765-+. doi: 10.1038/nbt.4192

Kozarewa, I. et al. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G plus C)-biased genomes. *Nature Methods* 6(4), pp. 291-295. doi: 10.1038/nmeth.1311

Krenning, L. et al. 2019. Life or Death after a Break: What Determines the Choice? *Molecular Cell* 76(2), pp. 346-358. doi: 10.1016/j.molcel.2019.08.023

Kunne, T. et al. 2014. Planting the seed: target recognition of short guide RNAs. *Trends in Microbiology* 22(2), pp. 74-83. doi: 10.1016/j.tim.2013.12.003

Lazzarotto, C. R. et al. 2020. CHANGE-seq reveals genetic and epigenetic effects on CRISPR-Cas9 genome-wide activity. *Nature Biotechnology* 38(11), pp. 1317-+. doi: 10.1038/s41587-020-0555-7

Leduc, F. et al. 2011. Genome-Wide Mapping of DNA Strand Breaks. *Plos One* 6(2), doi: 10.1371/journal.pone.0017353

Lensing, S. V. et al. 2016. DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods* 13(10), pp. 855-+. doi: 10.1038/nmeth.3960

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14), pp. 1754-1760. doi: 10.1093/bioinformatics/btp324

Li, H. et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16), pp. 2078-2079. doi: 10.1093/bioinformatics/btp352

Li, T. et al. 2011. TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Research* 39(1), pp. 359-372. doi: 10.1093/nar/gkq704

Listgarten, J. et al. 2018. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering* 2(1), pp. 38-47. doi: 10.1038/s41551-017-0178-6

Liu, J. J. et al. 2019a. CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* 566(7743), pp. 218-+. doi: 10.1038/s41586-019-0908-x

Liu, Z. C. et al. 2019b. Toxicogenomics: A 2020 Vision. *Trends in Pharmacological Sciences* 40(2), pp. 92-103. doi: 10.1016/j.tips.2018.12.001

Lord, C. J. and Ashworth, A. 2012. The DNA damage response and cancer therapy. *Nature* 481(7381), pp. 287-294. doi: 10.1038/nature10760

Maeder, M. L. et al. 2008. Rapid "Open-Source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular Cell* 31(2), pp. 294-301. doi: 10.1016/j.molcel.2008.06.016

Maier, D. A. et al. 2013. Efficient Clinical Scale Gene Modification via Zinc Finger Nuclease-Targeted Disruption of the HIV Co-receptor CCR5. *Human Gene Therapy* 24(3), pp. 245-258. doi: 10.1089/hum.2012.172

Mali, P. et al. 2013. RNA-Guided Human Genome Engineering via Cas9. *Science* 339(6121), pp. 823-826. doi: 10.1126/science.1232033

Mao, P. et al. 2016. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* 113(32), pp. 9057-9062. doi: 10.1073/pnas.1606667113

Mao, Z. Y. et al. 2008. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle* 7(18), pp. 2902-2906. doi: 10.4161/cc.7.18.6679

Massip, L. et al. 2010. Deciphering the chromatin landscape induced around DNA double strand breaks. *Cell Cycle* 9(15), pp. 2963-2972. doi: 10.4161/cc.9.15.12412

Mirzazadeh, R. et al. 2018. Genome-Wide Profiling of DNA Double-Strand Breaks by the BLESS and BLISS Methods. *Methods in molecular biology (Clifton, N.J.)* 1672, pp. 167-194. doi: 10.1007/978-1-4939-7306-4_14

Modrzejewski, D. et al. 2020. Which Factors Affect the Occurrence of Off-Target Effects Caused by the Use of CRISPR/Cas: A Systematic Review in Plants. *Frontiers in Plant Science* 11, doi: 10.3389/fpls.2020.574959

Mojica, F. J. M. et al. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of Molecular Evolution* 60(2), pp. 174-182. doi: 10.1007/s00239-004-0046-3

Mojica, F. J. M. et al. 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, bacteria and mitochondria. *Molecular Microbiology* 36(1), pp. 244-246. doi: 10.1046/j.1365-2958.2000.01838.x

Na, J. et al. 2014. Aneuploidy in pluripotent stem cells and implications for cancerous transformation. *Protein & Cell* 5(8), pp. 569-579. doi: 10.1007/s13238-014-0073-9

Newton, M. D. et al. 2019. DNA stretching induces Cas9 off-target activity. *Nature Structural & Molecular Biology* 26(3), pp. 185-+. doi: 10.1038/s41594-019-0188-z

Nishimasu, H. et al. 2014. Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* 156(5), pp. 935-949. doi: 10.1016/j.cell.2014.02.001

Nunez, J. K. et al. 2015. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 519(7542), pp. 193-+. doi: 10.1038/nature14237

Palermo, G. et al. 2016. Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *Acs Central Science* 2(10), pp. 756-763. doi: 10.1021/acscentsci.6b00218

Pattanayak, V. et al. 2013. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature Biotechnology* 31(9), pp. 839-+. doi: 10.1038/nbt.2673

Pavletich, N. P. and Pabo, C. O. 1991. ZINC FINGER DNA RECOGNITION - CRYSTAL-STRUCTURE OF A ZIF268-DNA COMPLEX AT 2.1-A. *Science* 252(5007), pp. 809-817. doi: 10.1126/science.2028256

Pinello, L. et al. 2016. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nature Biotechnology* 34(7), pp. 695-697. doi: 10.1038/nbt.3583

Pourcel, C. et al. 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology-Sgm* 151, pp. 653-663. doi: 10.1099/mic.0.27437-0

Qasim, W. et al. 2017. Molecular remission of infant B-ALL after infusion of universal TALEN gene-edited CAR T cells. *Science Translational Medicine* 9(374), doi: 10.1126/scitranslmed.aaj2013

Quinlan, A. R. and Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6), pp. 841-842. doi: 10.1093/bioinformatics/btq033

Radecke, S. et al. 2010. Zinc-finger Nuclease-induced Gene Repair With Oligodeoxynucleotides: Wanted and Unwanted Target Locus Modifications. *Molecular Therapy* 18(4), pp. 743-753. doi: 10.1038/mt.2009.304

Ramalingam, S. et al. 2014. TALEN-Mediated Generation and Genetic Correction of Disease-Specific Human Induced Pluripotent Stem Cells. *Current Gene Therapy* 14(6), pp. 461-472. doi: 10.2174/1566523214666140918101725

Robertson, M. 1980. BIOLOGY IN THE 1980S, PLUS OR MINUS A DECADE. *Nature* 285(5764), pp. 358-359. doi: 10.1038/285358a0

Romay, G. and Bragard, C. 2017. Antiviral Defenses in Plants through Genome Editing. *Frontiers in microbiology* 8, pp. 47-47. doi: 10.3389/fmicb.2017.00047

Rouet, P. et al. 1994. INTRODUCTION OF DOUBLE-STRAND BREAKS INTO THE GENOME OF MOUSE CELLS BY EXPRESSION OF A RARE-CUTTING ENDONUCLEASE. *Molecular and Cellular Biology* 14(12), pp. 8096-8106. doi: 10.1128/mcb.14.12.8096

Sansbury, B. M. et al. 2019. Understanding the diversity of genetic outcomes from CRISPR-Cas generated homology-directed repair. *Communications Biology* 2, doi: 10.1038/s42003-019-0705-y

Santiago, Y. et al. 2008. Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases. *Proceedings of the National Academy of Sciences of the United States of America* 105(15), pp. 5809-5814. doi: 10.1073/pnas.0800940105

Schindler, D. et al. 2018. Synthetic genomics: a new venture to dissect genome fundamentals and engineer new functions. *Current Opinion in Chemical Biology* 46, pp. 56-62. doi: 10.1016/j.cbpa.2018.04.002

Schipler, A. and Iliakis, G. 2013. DNA double-strand-break complexity levels and their possible contributions to the probability for error-prone processing and repair pathway choice. *Nucleic Acids Research* 41(16), pp. 7589-7605. doi: 10.1093/nar/gkt556

Schmid-Burgk, J. L. et al. 2020. Highly Parallel Profiling of Cas9 Variant Specificity. *Molecular Cell* 78(4), pp. 794-+. doi: 10.1016/j.molcel.2020.02.023

Schwer, B. et al. 2016. Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proceedings of the National Academy of Sciences of the United States of America* 113(8), pp. 2258-2263. doi: 10.1073/pnas.1525564113

Scott, S. J. et al. 2020. Synchronization of human retinal pigment epithelial-1 cells in mitosis. *Journal of Cell Science* 133(18), doi: 10.1242/jcs.247940

Shen, M. W. et al. 2018. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* 563(7733), pp. 646-+. doi: 10.1038/s41586-018-0686-x

Shen, W. et al. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *Plos One* 11(10), doi: 10.1371/journal.pone.0163962

Sherba, J. J. et al. 2020. The effects of electroporation buffer composition on cell viability and electro-transfection efficiency. *Scientific Reports* 10(1), doi: 10.1038/s41598-020-59790-x

Shi, X. et al. 2019. Cas9 has no exonuclease activity resulting in staggered cleavage with overhangs and predictable di- and tri-nucleotide CRISPR insertions without template donor. *Cell Discovery* 5, doi: 10.1038/s41421-019-0120-z

Shou, J. et al. 2018. Precise and Predictable CRISPR Chromosomal Rearrangements Reveal Principles of Cas9-Mediated Nucleotide Insertion. *Molecular Cell* 71(4), pp. 498-+. doi: 10.1016/j.molcel.2018.06.021

Simara, P. et al. 2017. DNA double-strand breaks in human induced pluripotent stem cell reprogramming and long-term in vitro culturing. *Stem Cell Research & Therapy* 8, doi: 10.1186/s13287-017-0522-5

Slaymaker, I. M. et al. 2016. Rationally engineered Cas9 nucleases with improved specificity. *Science* 351(6268), pp. 84-88. doi: 10.1126/science.aad5227

Sternberg, S. H. et al. 2015. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* 527(7576), pp. 110-113. doi: 10.1038/nature15544

Sternberg, S. H. et al. 2014. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* 507(7490), pp. 62-+. doi: 10.1038/nature13011

Taheri-Ghahfarokhi, A. et al. 2018. Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Research* 46(16), pp. 8417-8434. doi: 10.1093/nar/gky653

Tang, P. Z. et al. 2018. TEG-seq: an ion torrent-adapted NGS workflow for in cellulo mapping of CRISPR specificity. *Biotechniques* 65(5), pp. 259-266. doi: 10.2144/btn-2018-0105

Tena, A. et al. 2020. Induction of recurrent break cluster genes in neural progenitor cells differentiated from embryonic stem cells in culture. *Proceedings of the National Academy of Sciences of the United States of America* 117(19), pp. 10541-10546. doi: 10.1073/pnas.1922299117

Tsai, S. Q. and Joung, J. K. 2016. Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. *Nature Reviews Genetics* 17(5), pp. 300-312. doi: 10.1038/nrg.2016.28

Tsai, S. Q. et al. 2017. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR Cas9 nuclease off-targets. *Nature Methods* 14(6), pp. 607-+. doi: 10.1038/nmeth.4278

Tsai, S. Q. et al. 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology* 33(2), pp. 187-197. doi: 10.1038/nbt.3117

Tully, I. J. 2020. *Exploring the link between CHD2 mutations and double strand break repair in developing neurons.* PhD, Cardiff University.

Urnov, F. D. 2018. Genome Editing B.C. (Before CRISPR): Lasting Lessons from the "Old Testament". *Crispr Journal* 1(1), pp. 34-46. doi: 10.1089/crispr.2018.29007.fyu

Vakulskas, C. A. et al. 2018. A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nature Medicine* 24(8), pp. 1216-+. doi: 10.1038/s41591-018-0137-0

van der Oost, J. et al. 2014. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Reviews Microbiology* 12(7), pp. 479-492. doi: 10.1038/nrmicro3279

van Overbeek, M. et al. 2016. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell* 63(4), pp. 633-646. doi: 10.1016/j.molcel.2016.06.037

Verkuijl, S. A. N. and Rots, M. G. 2019. The influence of eukaryotic chromatin state on CRISPR-Cas9 editing efficiencies. *Current Opinion in Biotechnology* 55, pp. 68-73. doi: 10.1016/j.copbio.2018.07.005

Vilenchik, M. M. and Knudson, A. G. 2003. Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. *Proceedings of the National Academy of Sciences of the United States of America* 100(22), pp. 12871-12876. doi: 10.1073/pnas.2135498100

Vitelli, V. et al. 2017. Recent Advancements in DNA Damage-Transcription Crosstalk and High-Resolution Mapping of DNA Breaks. *Annual Review of Genomics and Human Genetics, Vol 18* 18, pp. 87-113. doi: 10.1146/annurev-genom-091416-035314

Wang, X. L. et al. 2015. Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nature Biotechnology* 33(2), pp. 175-178. doi: 10.1038/nbt.3127

Wienert, B. et al. 2019. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science* 364(6437), pp. 286-+. doi: 10.1126/science.aav9023

Wu, J. Z. et al. 2018. Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *Journal of the American Chemical Society* 140(31), pp. 9783-9787. doi: 10.1021/jacs.8b03715

Yan, W. X. et al. 2017. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications* 8, doi: 10.1038/ncomms15058

Yang, H. et al. 2020. Methods Favoring Homology-Directed Repair Choice in Response to CRISPR/Cas9 Induced-Double Strand Breaks. *International Journal of Molecular Sciences* 21(18), doi: 10.3390/ijms21186461

Yin, J. H. et al. 2019. Optimizing genome editing strategy by primer-extension-mediated sequencing. *Cell Discovery* 5, doi: 10.1038/s41421-019-0088-8

Zetsche, B. et al. 2015. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* 163(3), pp. 759-771. doi: 10.1016/j.cell.2015.09.038

Zhang, X. H. et al. 2015. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular Therapy-Nucleic Acids* 4, doi: 10.1038/mtna.2015.37

Zheng, Z. L. et al. 2014. Anchored multiplex FOR for targeted next-generation sequencing. *Nature Medicine* 20(12), pp. 1479-1484. doi: 10.1038/nm.3729