Year : 2019

Study of DNA repair and recombination mechanisms in Chinese hamster ovary cells

Bosshard Sandra

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Institut de Biotechnologie

**STUDY OF DNA REPAIR AND RECOMBINATION MECHANISMS
IN CHINESE HAMSTER OVARY CELLS**

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Sandra BOSSHARD

Maîtrise universitaire ès Sciences en sciences moléculaires du vivant
de l'Université de Lausanne, Suisse

Jury

Prof. Jan Roelof van der Meer, Président
Prof. Nicolas Mermod, Directeur de thèse
Prof. Alessandro A. Sartori, expert
Prof. Andrzej Stasiak, expert

Lausanne 2019

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Jan Roelof **Van der Meer**
Directeur·rice de thèse	Monsieur	Prof.	Nicolas **Mermod**
Experts·es	Monsieur	Prof.	Andrzej **Stasiak**
	Monsieur	Prof.	Alessandro A. **Sartori**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Sandra Bosshard

Maîtrise universitaire en sciences moléculaires du vivant, Université de Lausanne

intitulée

Study of DNA repair and recombination mechanisms in Chinese hamster ovary cells

Lausanne, le 3 mai 2019

pour le Doyen
de la Faculté de biologie et de médecine

Prof.   Jan Roelof Van der Meer

Acknowledgments

I would like to thank my thesis advisor Professor Nicolas Mermod for giving me the opportunity to pursue my PhD in his laboratory, for his support and his guidance. I greatly appreciated his curiosity and enthusiasm – they were truly inspiring. Moreover, he always had an open door for discussion.

I would also like to thank Professor Alessandro A. Sartori and Professor Andrzej Stasiak for being members of my thesis committee and for their helpful inputs on DNA repair mechanisms. Finally, I thank Professor Jan Roelof van der Meer for serving as the president of my thesis committee.

I am also grateful to the members of the VitalIT group who performed the bioinformatic analyses. Special thanks go to Emanuel Schmid-Siegert who set off with me on the adventure of detecting mutations upon genome editing.

I also wish to thank the members of the lab for their support and the many laughs. Special thanks go to Kaja Kostyrko for being a great mentor and colleague – notably at the beginning of my PhD studies. I also express my gratitude to Pierre-Olivier Duroy who introduced me to the mysteries of viruses and whom I could always approach for scientific as well as non-scientific matters. Besides my colleagues at the lab, I would also like to thank the members of the BioScience Network Lausanne with whom I had a whale of a time and who backed me during my presidency.

Finally, I would like to express gratitude to my family and friends who supported me during this journey. A big thank also goes to Adrian for all his help and company.

Summary

The CRISPR nuclease systems greatly facilitate targeted genome modifications in mammalian cells. The outcome of genome editing depends on the involved DNA double strand break (DSB) repair pathways. While the classical non-homologous end-joining and the poorly defined alternative end-joining (alt-EJ) DSB repair pathways can cause imprecise repair and thus gene inactivations, the homologous recombination (HR) pathway often introduces precise modifications. Although CRISPR is highly efficient at inactivating single genes, it is inefficient at introducing precise genome modifications. Moreover, its efficiency at inactivating multi-locus DNA sequences such as highly repetitive endogenous viral elements also remains limited.

This thesis addressed these limitations by better characterizing DSB repair pathways in Chinese hamster ovary (CHO) cells – the most widely used production cell host for therapeutic proteins. In this thesis, I first aimed at identifying rate-limiting factors to improve HR-mediated genome editing. Second, I strove for studying approaches to inactivate repetitive endogenous retroviruses (ERV) presumably releasing viral particles into the CHO supernatant.

To identify factors limiting HR, we established two chromosomal CHO assays that measure HR activity based on the correction of a GFP loss-of-function mutation. By using knockdown and overexpression studies, we found that efficient HR-mediated genome editing depended on certain alt-EJ activities. Furthermore, we observed that alt-EJ contribution to HR correlates with the nuclease type and the location of the DSB site relative to the GFP mutation. These observations suggest that alt-EJ and HR repair pathways tightly interact and challenges the common perception of alt-EJ opposing HR. Finally, among the tested repair factors, high Mre11 nuclease and Pari anti-recombinase as well as low Rad51 recombinase levels were the most rate-limiting factors for HR in CHO cells.

Counteracting these bottlenecks improved HR efficiency by 75%.

To inactivate repetitive ERVs, we transiently expressed a CRISPR-Cas9 nuclease that targets the *gag* gene of a specific transcriptionally active ERV group. Clones bearing a loss-of-function mutation in one particular ERV locus and corresponding mRNA produced considerably fewer particles loaded with viral RNA genomes. These findings indicated that a single ERV locus is responsible for the release of most, if not all, viral particles from CHO cells. Notably, ERV mutagenesis did not compromise cell growth, cell size or therapeutic protein production. In sum, this work provided novel strategies to improve HR-mediated genome editing and to inhibit viral particle release from CHO cells.

Résumé

Le système de nucléases CRISPR permet d'effectuer des modifications ciblées dans le génome des cellules de mammifères. Le résultat de l'édition du génome dépend des voies de réparation de la cassure double brin (CDB) du fragment ADN. Les deux voies de réparation de la CDB par jonction d'extrémités, la voie jonction d'extrémités non homologues classiques et la voie de jonction d'extrémités alternatives (alt-EJ) mal définie, provoquent des réparations imprécises et des inactivations génétiques. Au contraire, la voie de recombinaison homologue (RH) introduit la plupart du temps des modifications précises. Bien que le système CRISPR soit très efficace pour inactiver des gènes individuels, cette approche reste inefficace pour introduire des modifications génomiques précises. En outre, son efficacité à inactiver des séquences d'ADN multi-locus, telles que des éléments viraux endogènes hautement répétés, reste également limitée.

Cette thèse a abordé ces limitations en améliorant la compréhension des voies de réparation de la CDB dans les cellules ovariennes de hamster chinois (CHO), qui est le système cellulaire le plus utilisé pour la production de protéines thérapeutiques. Pour ce faire, j'ai premièrement identifié les facteurs limitant la RH pour améliorer l'édition précise du génome. J'ai ensuite développé des approches visant inactiver les rétrovirus endogènes répétitifs (ERV) qui sont suspectés de libérer des particules dans le surnageant des cellules CHO.

Pour identifier les facteurs limitant la RH, nous avons établi deux essais dans les cellules CHO permettant de mesurer l'activité de la RH par la correction d'une mutation d'une séquence GFP non fonctionnelle. En diminuant ou en augmentant l'expression de facteurs impliqués dans les voies de réparation, nous avons constaté que l'efficacité de la RH dépendait de certains facteurs impliqués dans la voie d'alt-EJ. De plus, nous avons observé que la contribution de l'alt-EJ pour la RH dépendait du type de nucléases utilisés

et de la localisation de la CDB par rapport à la mutation dans le transgène de la GFP. Ces observations suggèrent que les voies de réparation de l'alt-EJ et de la RH interagissent étroitement, et ceci remet en question le consensus selon lequel l'alt-EJ est en compétition avec la RH. Enfin, parmi les facteurs de réparation testés, nous avons constaté que la RH était limitée par des niveaux élevés de nucléase Mre11 et d'anti-recombinase Pari ainsi que par des faibles niveaux de recombinase Rad51. En modifiant le niveau d'expression de ces éléments limitants, nous avons pu améliorer l'efficacité de la HR de 75%.

Pour inactiver les ERV répétitifs, nous avons exprimé de manière transitoire une nucléase CRISPR-Cas9 qui cible le gène *gag* d'un groupe spécifique d'ERV transcriptionnellement actifs. Les clones comprenant une mutation de perte de fonction dans un locus ERV particulier et dans l'ARNm correspondant montraient une production largement réduite de particules chargées d'ARN génomique viral. Ces constatations indiquent qu'un seul locus ERV est responsable de la libération de la plupart, sinon de toutes les particules virales des cellules CHO. Notamment, la mutagenèse d'ERV n'a compromis ni la croissance cellulaire, ni la taille des cellules, ni la production de protéines thérapeutiques. Globalement, ce travail propose de nouvelles stratégies permettant d'améliorer l'édition génomique médiée par la HR, et d'inhiber la libération de particules virales par les cellules CHO.

Abbreviations

53BP1	p53 binding protein 1
Alt-EJ	Alternative end-joining
Amp	Ampicillin resistance gene
AP site	Apurinic or apyrimidinic site
Ape1	AP endonuclease 1
APTX	Aprataxin
ATM	Ataxia telangiectasia mutated
ATR	Ataxia telangiectasia and Rad3-related
B-NHEJ	Backup non-homologous end-joining
B2M	Beta-2-microglobulin
Bard1	Brca1-associated RING Domain 1
BER	Base excision repair
BIR	Break-induced replication
Blm	Bloom syndrome RecQ-like helicase
bp	Base pair
Brca	Breast cancer susceptibility protein
C-NHEJ	Classical non-homologous end-joining
Cas9n	Cas9 nickase
CDC25	Cell division cycle 25
cDNA	Complementary DNA
CHK	Checkpoint kinase
CHO	Chinese hamster ovary
CHO-DG44	Adhererent CHO cells
CHO-K1	Suspension-adapted CHO cells
Cidec	Cell death inducing DFFA like effector c

CO	Crossover
CRISPR-Cas	Clustered regularly interspaced short palindromic repeats - CRISPR-associated
crRNA	CRISPR RNA
CSA and CSB	Cockayne syndrome A and B
CtIP	CtBP-interacting protein
D-loop	Displacement loop
DAPI	4'6-diamidino-2-phenylindole
dCas9	Catalytically inactive Cas9
DDR	DNA damage response
dHJ	Double holliday junction
DNA	Deoxyribonucleic acid
DNA-PKcs	DNA-dependent protein kinase catalytic subunit
Dna2	DNA replication helicase/nuclease 2
DR-GFP	Direct repeat GFP
dRP	Deoxyribose phosphate
DSB	Double strand break
DSBR	Double strand break repair
dsDNA	Double-stranded DNA
dsRed	Discosoma sp. red fluorescent protein
Ercc1	Excision repair cross-complementing group 1
ERV	Endogenous retrovirus
Exo1	Exonuclease I
FA	Fanconi anemia
FACS	Fluorescence-activated cell sorting
FANC	Fanconi anemia complementation group
Fbxo18	F-box protein 18
FeLV	Feline leukemia virus
FISH	Fluorescent in-situ hybridization
FnCas9	Cas9 ortholog from *Francisella novicida*
FokI-dCas9	FokI nuclease fusion to dCas9
FoSTeS	Fork stalling and template switching
FSC	Forward scatter
Gag	Group-specific antigen gene
GFP	Green fluorescent protein

GG-NER	Global genome NER
HDR	Homology-directed repair
HITI	Homology-Independent Targeted Integration
HIV	Human immunodeficiency virus
HR	Homologous recombination
IAP	Intracisternal type-A retrovirus
ICL	DNA interstrand crosslink
Ig	Immunoglobulin
Indel	Mutation consisting of insertions and deletions
IR	Ionization radiation
ITR	Inverted terminal repeat
Jagn1	Jagunal homolog 1
kb	Kilobase
Keap1	Kelch-like ECH-associated protein 1
KI	Knock-in
Lig1	Ligase I
Lig3	Ligase III
Lig4	Ligase IV
LOH	Loss of heterozygosity
LTR	Long terminal repeat
Mdc1	Mediator of DNA-damage checkpoint 1
MH	Microhomology
Mlh1	MutL homolog 1
MLV	Murine leukemia virus
MM-BIR	Microhomology-mediated BIR
MM-SDSA	Microhomology-mediated SDSA
MMC	Mitomycin C
MMEJ	Microhomology-mediated end-joining
MMR	DNA mismatch repair
Mre11	Meiotic recombination 11
MRN	Complex composed of Mre11-Rad50-Nbs1
mRNA	Messenger RNA
Msh2	MutS homolog 2
Myr	Myristoylation

Nbs1	Nijmegen breakage syndome 1
NCO	Non-crossover
NER	Nucleotide excision repair
nt	Nucleotide
ObLiGaRe	Obligate Ligation-Gated Recombination
ORF	Open reading frame
Palb2	Partner and localizer of BRCA2
PAM	Protospacer adjacent motif
Pari	Proliferating Cell Nuclear Antigen (PCNA)-interacting factor
Parp1	Poly (ADP-ribose) polymerase 1
Paxx	Paralog of Xrcc4 and Xlf
PB	PiggyBac
PCNA	Proliferating cell nuclear antigen
PCR	Polymerase chain reaction
PITCh	Precise Integration into Target Chromosomes
PNKP	Polynucleotide kinase 3-phosphatase
Pol theta	DNA polymerases θ
Pold3	DNA polymerase δ subunit 3
PPYP	CHO-specific PPxY-related budding motif
pre-crRNA	Precursor CRISPR RNA
PTM	Post-translational modification
Puro	Puromycin resistance gene
qPCR	Quantitative real-time polymerase chain reaction
Recql5	RecQ-like helicase 5
Rmi	RecQ-mediated genome instability protein
RNA	Ribonucleic acid
ROS	Reactive oxygen species
RPA	Replication protein A
RPKM	Reads per kilobase per million
RT-qPCR	Quantitative reverse transcription polymerase chain reaction
Rtel1	Regulator of telomere elongation helicase 1
s.e.m	Standard error of the mean
SD-MMEJ	Synthesis-dependent microhomology-mediated end-joining
SDSA	Synthesis-dependent strand annealing

sgRNA	Single guide RNA
siNeg	Non-targeting negative control siRNA
siRNA	Small interfering RNA
SpCas9	Cas9 ortholog from *Streptococcus pyogenes*
SSA	Single strand annealing
SSB	Single strand break
SSC	Side scatter
ssDNA	Single-stranded DNA
ssODN	Single-stranded oligodeoxyribonucleotide
SSTR	Single-stranded template repair
TALEN	Transcription activator-like effector nuclease
TC-NER	Transcription-coupled NER
TLS	Translesion synthesis
TMEJ	Polymerase θ-mediated end-joining
Topo3α	DNA topoisomerase III α
TPM	Transcripts per kilobase million
tracrRNA	Trans-activating CRISPR RNA
UV	Ultra-violet light
UV-DDB	UV-damaged DNA-binding protein
VLP	Viral-like particle
Wrn	Werner syndrome RecQ-like helicase
WT	Wild-type
Xlf	Xrcc4-like factor
XP	Xeroderma pigmentosum complementation group
Xpf	Xeroderma pigmentosum complementation group F
Xrcc4	X-ray repair cross-complementing protein 4
ZFN	Zinc finger nuclease

List of publications and patents

Publications and manuscripts

Duroy P.O.*, **Bosshard S.***, Schmid-Siegert, E., Neuenschwander, S., Arib G., Lemercier P., Masternak J., Roesch L., Buron F., Girod P.A., Xenarios, I., Mermod, N. Characterization and inactivation of endogenous retroviruses in Chinese hamster ovary cells. *Manuscript in preparation.* * equal contribution

Bosshard S., Mermod N. A role for alternative end-joining factors in homologous recombination and genome editing. *Submitted for publication.*

Kostyrko K., Neuenschwander S., Junier T., Regamey A., Iseli C., Schmid-Siegert E., **Bosshard S.**, Majocchi S., Girod P.A., Xenarios I., Mermod N. (2017) MAR-mediated transgene integration into permissive chromatin and increased expression by recombination pathway engineering. *Biotechnology and Bioengineering*, **114**(2), 384–396.

Kostyrko K., **Bosshard S.**, Urban Z., Mermod N. (2015) A role for homologous recombination proteins in cell cycle regulation. *Cell Cycle*, **14**(17), 2853-2861.

Patent applications

Duroy P.O., **Bosshard S.**, Le Mercier P., Schmid-Siegert, E., Mermod N., (2018) Characterization and inactivation of endogenous retroviruses in Chinese hamster ovary cells. *Provisional patent application*, 62784566.

Mermod N., Duroy P.O., **Bosshard S.**, Le Mercier P. (2015) Improved eukaryotic cells for protein manufacturing and methods of making them. *Patent application*, WO2017109177 A1.

Table of contents

Chapter 1

Introduction

Genome editing techniques, such as the CRISPR-Cas9 nuclease system, allow to efficiently introduce targeted genome modifications in many types of cells and organisms. Genome editing is widely used to study fundamental biological processes and holds great promise for translational applications, for instance, to cure genetic diseases. The outcome of genome editing depends on the cellular DNA double strand break (DSB) repair machinery. In the present thesis, I focused on the interconnection between DNA DSB repair pathways and CRISPR-Cas9-mediated genome editing in Chinese hamster ovary (CHO) cells, the most widely used cell system to produce therapeutic proteins. In particular, I aimed at developing new genome editing strategies to improve precise homologous recombination-mediated genome modifications as well as to prevent the release of viral-like particles from endogenous retroviruses (ERVs) embedded in the CHO genome. In the first section of this chapter, I reviewed the current knowledge of DSB repair mechanisms in mammalian cells. In the second section, I described genome editing systems and their current limitations. Finally, in the third section, I summarized the CHO expression system and the threat of ERVs as source for viral contaminations.

1.1   DNA lesions and the DNA damage response

Cells encounter hundreds to thousands of spontaneous DNA lesions per day [1]. Typical DNA lesions include modified DNA bases, mispaired or entirely missing nucleotides, intra- and interstrand crosslinks, as well as single and double strand breaks (SSBs and

DSBs, respectively). All these lesions threaten genome stability and integrity as well as cell viability. To cope with and recover from complex DNA lesions, mammalian cells possess an intricate network of signaling and repair pathways, termed the DNA damage response (DDR) [2–5]. The DDR essentially detects and signals the presence of DNA lesions and activates downstream effectors to induce cell cycle checkpoints, DNA repair pathways or apoptosis (Fig. 1.1).
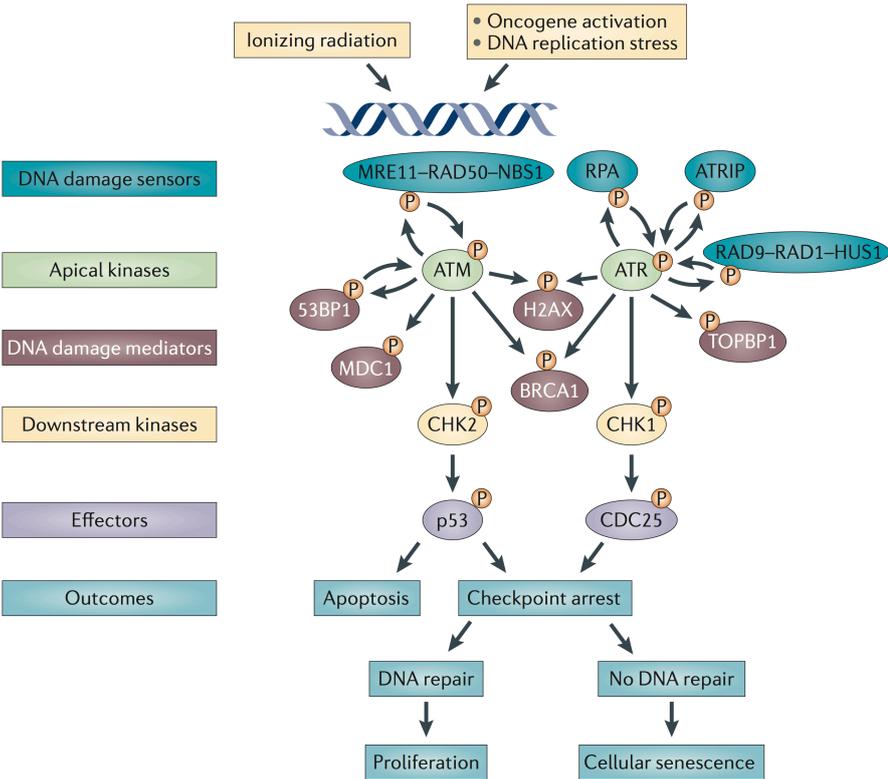
Figure 1.1: The DNA damage response is a signaling cascade able to sense and propagate the damage signals to effector proteins that activate cell cycle checkpoints, DNA repair or apoptosis (reproduced from [4]).

The key mediators of the DDR are the ataxia-telangiectasia mutated (ATM) and ataxia telangiectasia and Rad3-related (ATR) kinases [6, 7]. The ATM kinase is activated by DSBs, which may arise from exposure to ionizing radiation (IR) and are detected by the primary DSB damage sensor Mre11-Rad50-Nbs1 (MRN) complex. The MRN complex consists of a Mre11-Rad50 heterotetramer, which translocates to the nucleus through Nbs1, and tethers broken DSB ends together [8]. In contrast, the ATR kinase activity depends on replication protein A (RPA)-covered single-stranded DNA (ssDNA), which may emerge from stalling DNA replication forks as well as resected DSB ends.

ATM and ATR phosphorylate a large variety of targets, including checkpoint kinase 1

and 2 (CHK1 and CHK2), which are mainly phosphorylated by ATR and ATM, respectively [9]. These diffusible kinases spread the DDR signal away from the DNA damage site and further phosphorylate downstream effectors, such as the p53 transcription factor or the cell division cycle 25 (CDC25) phosphatases. Phosphorylation of p53 induces the G1/S checkpoint, while phosphorylation of CDC25 triggers the G1/S, intra-S and G2/M checkpoints, delaying cell cycle progression to allow time for repair [10].

ATM and ATR also phosphorylate Ser139 of histone variant H2AX, known as γH2AX [11, 12]. γH2AX induces global changes in the chromatin structure flanking the DSB site and serves as a "landing runway" for other DDR factors that help to sustain and amplify the DDR signal [5]. For instance, γH2AX attracts the DNA damage mediator MDC1 [13], which recruits additional MRN and ATM molecules to the DSB site, and it promotes recruitment of p53 binding protein 1 (53BP1) and Breast cancer susceptibility protein 1 (Brca1) to the DNA damage site, which are both also targets of ATM and/or ATR [7]. 53BP1 and Brca1 further amplify DDR signaling but also contribute to the classical non-homologous end-joining (C-NHEJ) and homologous recombination (HR) DNA DSB repair pathways, respectively, as described below in Sections 1.2.1 and 1.2.2. The accumulation of numerous DDR and repair factors at the DSB creates nuclear foci visible by fluorescence microscopy.

Severe DNA damage or failure to remove it leads to persistent DDR signaling. This may trigger apoptosis, mediated by p53-induced transcription of the Bax and Puma proteins [14], or permanent cell cycle arrest, known as cellular senescence [4].

1.2 Double strand break (DSB) repair

DSBs are one of the most cytotoxic DNA lesions since both strands of the DNA double helix are broken simultaneously. A mammalian cell encounters ~50 DSBs per cell cycle [15]. DSBs can occur spontaneously from endogenous and exogenous DNA damaging agents. The major sources of endogenous DSB are stalled or collapsed DNA replication forks, reactive oxygen species (ROS) and other cellular metabolites, while exogenous DSBs generally result from IR and anticancer chemotherapeutic drugs [16, 17]. In addition, DSBs can arise naturally during meiosis, increasing genetic diversity and ensuring proper chromosome segregation, as well as during immune system maturation by V(D)J recombination, class switch recombination, and somatic hypermutations [18–20]. Incor-

rect or missing repair of DSBs may result in mutations and chromosome rearrangements, predisposition to cancer, genetic disorders or cell death.

To ensure proper DSBs repair, eukaryotes possess several DSB repair pathways. The two main DSB repair pathways are classical non-homologous end-joining (C-NHEJ) and homologous recombination (HR). Moreover, DSBs can be repaired by single strand annealing (SSA) and by a family as of yet poorly described alternative end-joining (alt-EJ) pathways.

1.2.1 Classical non-homologous end-joining (C-NHEJ)

Classical non-homologous end-joining (C-NHEJ) is the predominant DSB repair pathway in most mammalian cells. It is the principal DSB repair pathway for IR-induced DNA damage and for diversification of the antigen receptor during V(D)J recombination [19, 21]. Therefore, mutations in C-NHEJ genes are often associated with radiation sensitivity and severe immunodeficiency [22]. C-NHEJ is a fast DSB repair pathway, with most C-NHEJ repair processes being completed within 30 min [23]. It is active in non-dividing and diving cells throughout the cell cycle but is particularly important in G1/G0 and early S phase, when HR is not active [21]. Unlike other DSB repair pathways, C-NHEJ relies on no or only on minimal end processing of broken DNA ends and can ligate the DNA ends regardless of sequence homology, although short microhomologies (MHs) may be used to facilitate annealing of broken DNA ends [24, 25].

C-NHEJ is generally described as an error-prone repair pathway [26]. Indeed, most naturally occurring DSBs produce non-complementary ends (e.g. mismatching overhangs, hairpins, damaged or chemically modified bases) that prevent direct end re-ligation. Hence, to create ligation-compatible DNA ends, C-NHEJ requires minimal end processing by nucleases and/or polymerases to trim the ends and/or to add nucleotides [24, 27, 28]. This processing often results in small heterogeneous insertions and deletions (indels) at the DSB site. However, C-NHEJ can also give rise to accurate repair outcomes [26]. This is the case for fully compatible DSB ends, such as blunt and cohesive ends introduced by restriction enzymes, that can be directly ligated, leading to perfectly reconstituted DNA sequences [27].

Upon ATM-mediated DSB recognition, accumulation of the DDR protein 53BP1 at broken DNA ends is a key step towards C-NHEJ repair [29] (Fig. 1.1). Besides DSB

4

checkpoint activation and recruitment of other repair factors, 53BP1 acts together with the downstream factors Rif1 and Rev7 (also known as Mad2l2) to prevent DSB end resection in the G1 cell cycle phase [29–31]. Hence, 53BP1-mediated end protection favors C-NHEJ over the resection-dependent HR, SSA and alt-EJ pathways. Moreover, 53BP1 contributes to C-NHEJ repair by bridging the two DSB ends [32].

Subsequently, 53BP1-primed DSB ends are recognized by the Ku heterodimer consisting of the two subunits Ku70 (Xrcc6) and Ku80 (Xrcc5). This further counteracts DSB end resection (Fig. 1.2). Most cellular Ku70 and Ku80 proteins are bound as stable Ku70-Ku80 heterodimer complex and deficiency in either Ku70 or Ku80 protein results in low levels of the other Ku subunit [33, 34]. Ku proteins are highly abundant in eukaryotes and possess a tight affinity for DSB ends, yet a poor affinity for ssDNA [35–37].
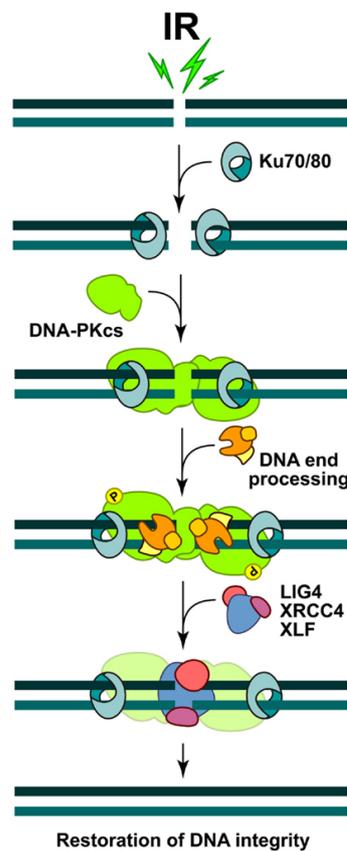
Figure 1.2: Schematic representation of DSB repair by C-NHEJ (reproduced from [38]).

Binding of Ku70-Ku80 facilitates recruitment of other C-NHEJ components, such as the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) to DNA ends, which forms together with the Ku heterodimer the DNA-PK holoenzyme [39]. Upon DSB binding, DNA-PKcs autophosphorylates, helps end bridging and activates Artemis, the main

endonuclease for C-NHEJ [40]. It is currently believed that Artemis or other endonucleases are dispensable for simple end-joining but important for processing damaged or incompatible DSB ends by resecting 5' and 3' overhangs as wells as opening hairpins [24, 41].

DNA-bound Ku70-Ku80 also serves as scaffold for various DNA polymerases, notably the Pol μ and Pol λ. These polymerases can add nucleotides in a template-dependent or template-independent (mainly Pol μ) manner and, thus, help to fill the gaps of Artemis-resected DNA ends [42, 43]. Finally, DNA ligase IV (Lig4) and X-ray repair cross-complementing protein 4 (Xrcc4) mediate DSB end ligation [44]. They are stimulated by the two accessory proteins Xrcc4-like factor (Xlf; also known as Cernunnos) as well as paralog of Xrcc4 and Xlf (Paxx) [45, 46].

1.2.2 Homologous recombination (HR)

Homologous recombination (HR) is the most accurate DSB repair mechanism in mammalian cells. It is also involved in the recovery of damaged replication forks, in the final steps of interstrand crosslink (ICL) repair, telomere maintenance and meiosis [18, 47–49]. HR is an intricate and rather slow DSB repair process [50]. It is the major DSB repair pathway in yeast, but exhibits only low activity in most mammalian cells [21, 51–53]. Mutations in HR genes, notably in Brca1 and Brca2, are associated with increased risks of breast and ovarian cancer [54].

Unlike C-NHEJ, HR uses long homologous sequences of at least 200-500 bp in length as template for error-free repair [55]. These homologous sequences mostly locate in *trans* and preferentially on sister chromatids, allowing to reconstitute the missing sequence with no loss in genetic information [56]. However, HR may also use homologous chromosomes, homologous sequences in non-allelic positions (e.g. repetitive elements) or exogenous donor DNA as illegitimate repair template which may cause loss of heterozygosity (LOH) and chromosomal rearrangements [57, 58]. Hence, to minimize illegitimate recombination, HR is predominately active in late S and G2 cell cycle phase, when sister chromatids are available and paired [59].

The initial step in HR is exposure of long 3' ssDNA by extensive 5'-3' DNA end resection, which commits DSB repair to HR (Fig. 1.3). The transition between end protection and end resection is believed to be mediated by Brca1, a target of the ATM kinase. Brca1,

in concert with the Brca1-associated RING Domain 1 (Bard1), promotes end resection by antagonizing the resection barrier imposed by 53BP1 and by interacting with the MRN-CtIP nuclease complex [54, 60]. End resection occurs in a two-step bidirectional process: First, Mre11, the nuclease of MRN, initiates end resection by endonucleolytic incisions in the 5' strand proximal to the DSB end, a process that is stimulated by CtIP [61, 62]. The incised DNA strand is then digested by the 3'-5' exonuclease activity of Mre11 creating short 3' ssDNA overhangs of approximately 100 nt in length [8]. Second, upon initial resection, long-range resection in the 5'-3' direction generates 3' ssDNA tails of kilobases in length. In mammalian cells, this is mediated by the Exonuclease I (Exo1) alone or by the Dna2 nuclease in conjunction with either the Werner syndrome (Wrn) helicase and/or the Bloom syndrome (Blm)-Topo3α-Rmi1-Rmi2 helicase-topoisomerase complex [63]. Resected ssDNA is rapidly bound by RPA, which protects the exposed ssDNA and contributes to amplifying the DDR via ATR signaling (Fig. 1.1).

Subsequently, the Rad51 recombinase, a key HR protein, replaces RPA bound to the resected ssDNA. Rad51 (RecA in bacteria) forms helical nucleoprotein filaments on the ssDNA, which stretches the bound DNA by 50% [65, 66]. In mammalian cells, the formation of the Rad51-ssDNA complex, termed presynaptic filament, is mediated by the tumor suppressor Brca2 in cooperation with the Brca1-Bard1-Palb2 complex and possibly the Rad51 paralogs (Rad51B, Rad51C, Rad51D, Xrcc2, Xrcc3) [67, 68]. In Brca2-deficient cells, Rad51 loading depends on the Rad52 protein, the main presynaptic filament mediator in yeast cells [69]. To prevent spontaneous and unscheduled HR, Rad51 presynaptic filament formation is regulated by the Kelch-like ECH-associated protein 1 (Keap1) factor, which prevents Brca1 interaction with Brca2-Palb2 during G1 phase [70]. By contrast, Rad51 nucleofilament disassembly prior to strand invasion is controlled by anti-recombinases, including the Proliferating Cell Nuclear Antigen (PCNA)-interacting factor (Pari), F-box protein 18 (Fbxo18; also known as Fbh1), RecQ-like helicase 5 (Recql5) and Blm [71–74].

The presynaptic filament orchestrates homology search and pairing with intact homologous repair templates. Homology search likely occurs through a random three dimension sampling together with a one dimension sliding mechanism with preference for sequences in close proximity to the DSB [75, 76]. The closest template is typically located within sister chromatids, as they are tied together by cohesin molecules [77]. However, it can also be found within non-allelic sequences, if they colocalize with the DSB due to spatial genome organization like chromatin loops or clustered highly-transcribed genome loci
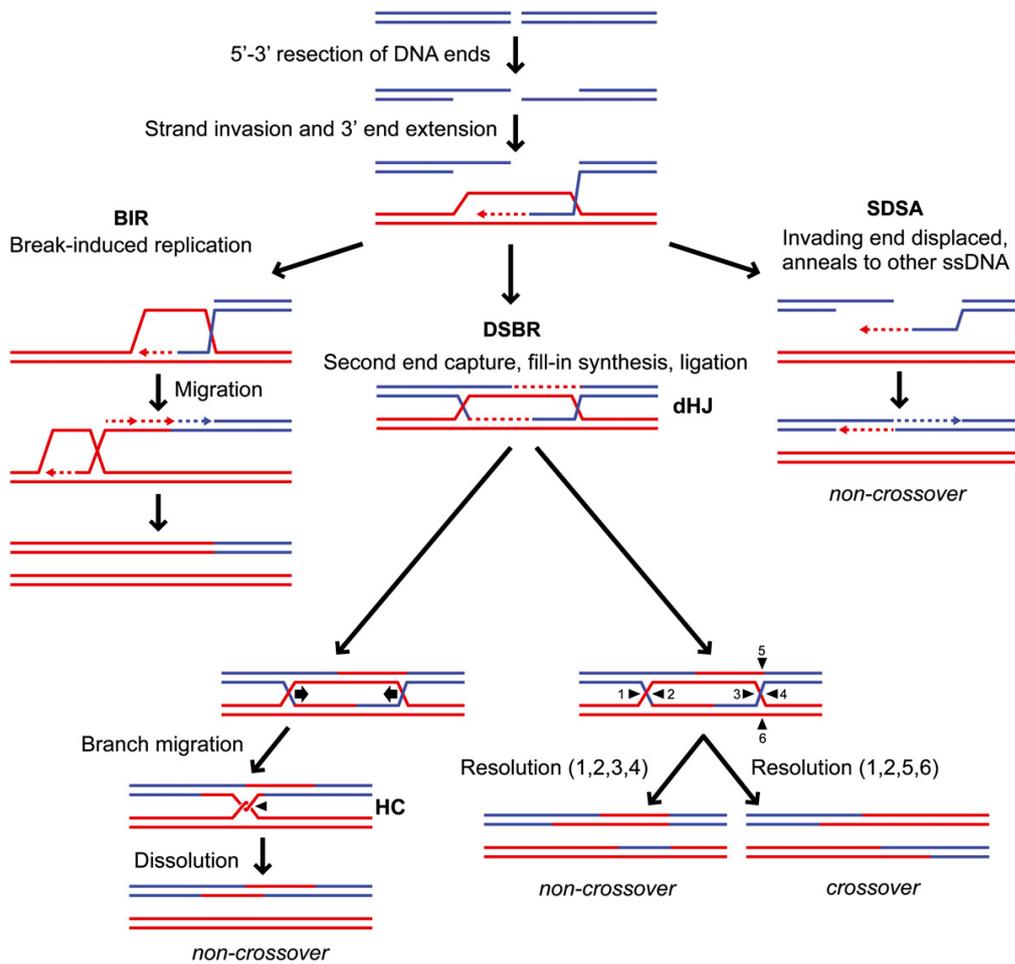
Figure 1.3: Homologous recombination (HR) repair of DNA double strand breaks (DSBs) by break-induced replication (BIR), canonical DSB repair (DSBR) and synthesis-dependent strand annealing (SDSA). The intact homologous sequence (in red) serves as template for the damaged DNA sequence (in blue). Pairing of red and blue DNA sequences generates a heteroduplex DNA. Newly synthesized DNA is depicted with dotted lines. Double Holliday junctions (dHJs) can either be dissolved by helicases into a hemicatenane (HC) producing non-crossover products or resolved by nucleases creating non-crossover (positions 1, 2, 3, and 4) or crossover (positions 1, 2, 5, and 6) products (reproduced from [64]).

[78]. Identification of homology triggers strand invasion of the Rad51 filament into the dsDNA donor template, mediated by Brac1-Bard1, Rad51 paralogs and Rad54 helicase [68, 79, 80], and initiates strand exchange [81]. This leads to the formation of a displacement (D)-loop that contains a heteroduplex as well as a displaced ssDNA strand, which is coated by RPA (Fig. 1.3). If the heteroduplex sequences within the nascent D-loop are similar but too divergent (homeologous) for recombination, mismatch repair (MMR) can abort HR via heteroduplex rejection [82].

In order to reconstitute the missing DNA sequence disrupted by the DSB, DNA syn-

thesis is established within the nascent D-loop. To this end, the Rad54 helicase mediates the dissociation of the Rad51 recombinase [83] allowing the invading 3' end to prime DNA synthesis and extend the D-loop along the homologous template with the help of DNA polymerase δ [84]. Upon D-loop extension, HR repair branches into three distinct pathways, that result in different repair products: synthesis-dependent strand annealing (SDSA), canonical DSB repair (DSBR) and break-induced replication (BIR) (Fig. 1.3).

In case of the DSBR pathway, the displaced strand of the expanded D-loop anneals to ("captures") the second broken 3' ssDNA end [85]. This second-end capture step creates another region of heteroduplex DNA and is followed by gap filling via DNA synthesis and end ligation, leading to the formation of a double Holliday Junction (dHJ). This stable repair intermediate can then be resolved or dissolved. dHJ resolution depends on structure-specific endonucleases, such as the Gen1 and the Slx1-Slx4-Mus81-Eme1 complex, which introduce symmetric and asymmetric nicks, respectively, into the dHJ [86]. Cleaving both inner dHJ strands maintains the original DSB flanking sequences and produces non-crossover products (NCO) while cleaving one inner and one outer dHJ strand switches the DSB flanking sequences and yields crossover products (CO) [64] (Fig. 1.3). In contrast, the Blm-Topo3α-Rmi1-Rmi2 complex mediates dHJ dissolution. In addition to its role in end resection, this complex converges the dHJ via branch migration, leading to the formation of a hemicatenane intermediate followed by its decatenation [64]. Dissolution leads to NCO events only.

In case of the SDSA pathway – the major HR pathway in mitotic cells [87] – the newly synthesized strand becomes dissociated from the extended D-loop and anneals to the complementary tail of the second broken 3' ssDNA end. This creates an additional heteroduplex region that initiates DNA fill-in synthesis and end ligation. Several anti-recombinases contribute to the dissociation of the extended D-loop in mammalian cells, including the Regulator of telomere elongation helicase 1 (Rtel1), Fanconi anemia complementation group M (FancM), and Blm proteins, and, thus, channel HR repair towards SDSA [88–91]. Unlike DSBR, SDSA results exclusively in NCO events.

The D-loop may proceed to the BIR pathway in the absence of a second homologous DSB end, which commonly occurs at stalled and broken replication forks or at eroded telomeres [49]. BIR initiates a long-range extension of the D-loop by DNA replication to complete replication until the end of the chromosome, which is presumably mediated by the non-essential Polδ3 (Pold3) polymerase subunit and Pif1 helicase [92, 93]. BIR

DNA replication is conservative and associated with frequent template switching [94, 95]. This special mode of DNA replication can lead to extensive LOH as well as complex chromosomal rearrangements and is thus more mutagenic than DSBR and SDSA.

HR subpathways are tightly controlled to ensure appropriate balance of CO and NCO products. During meiosis, COs occur frequently and promote allelic exchange between homologous chromosomes and assure their proper segregation [18, 96]. However, they are suppressed during mitosis as they may cause LOH in the DSB flanking sequence and chromosome rearrangements between non-allelic sequences [57, 58]. Thus, during mitosis, NCOs are the preferred repair outcome as they are less mutagenic than COs [97]. Most NCOs arise from SDSA repair.

Both NCO and CO products can result in gene conversion. Gene conversion refers to a non-reciprocal exchange of genetic material between two homologous sequences, which can lead to short-range LOH in the converted region. The gene conversion length, known as gene conversion tract, depends on the extent of the heteroduplex DNA as well as of the DNA sequence copied from the donor strand to fill the gap and is around five-fold longer in CO than in NCO products [98]. More precisely, the heteroduplex DNA sequences formed during HR can consist of mismatched bases when using non-identical homologous donor templates [97, 99]. Repair of these mismatches by MMR result in restoration of the original sequence, gene conversion or a mixture of both [99].

1.2.3   Single strand annealing (SSA)

Single strand annealing (SSA) is a DSB repair pathway that occurs in regions of direct repeat sequences [100]. For instance, SSA has been observed between repetitive elements in tandem orientation, including long terminal repeats (LTRs) of retroelements or homologous segments of Alu elements [101, 102]. SSA requires extensive 3' end resection to expose complementary ssDNA ends of at least 30 nt in length. Subsequently, RPA-coated complementary homologous sequences are annealed by Rad52 [103], a process that resembles the second end annealing during SDSA in HR. Following annealing, protruding non-homologous tails (termed flaps) are removed by Ercc1-Xpf, gaps filled and ends ligated by presumably Lig1 [104–106]. SSA is a highly mutagenic repair pathway as annealing of complementary sequences leads to the deletion of one repeat copy as well as the intervening sequence. As SSA depends on extensive end-resection and the Rad52

epistasis group, SSA is sometimes classified as HR repair mechanism despite the lack of Rad51-mediated strand invasion.

1.2.4 Alternative end-joining (Alt-EJ)

In addition to C-NHEJ, HR and SSA, growing evidence suggests the existence of a fourth incompletely characterized DSB repair pathway family [107–111]. This family of DSB repair pathways is generally referred to as alternative end-joining (alt-EJ, sometimes abbreviated as a-EJ or alt-NHEJ). Occasionally, it is also termed backup non-homologous end-joining (B-NHEJ) or polymerase θ-mediated end-joining (TMEJ). Alt-EJ activity was originally discovered in NHEJ-deficient yeast and mammalian cells that still exhibited robust end-joining and V(D)J recombination [112–114]. Although alt-EJ was initially considered as a backup pathway, current data imply that alt-EJ also operates in cells proficient for both C-NHEJ and HR repair [114, 115]. Interestingly, many cancer and immortalized cells possess upregulated alt-EJ levels, likely to compensate for deficient or insufficient C-NHEJ and HR activities [116–119]. Alt-EJ factors may thus serve as potential therapeutic target for cancer treatment [120].

Alt-EJ pathways rely on microhomologies (MHs) at or near the DSB to anneal broken DNA ends together. These MHs usually range from as little as 1 bp to $\sim$ 25 bp, as opposed to C-NHEJ and HR, which require no and longer homologous donor templates, respectively [121–123]. Annealing of these MH sequences typically results in the deletion of one MH copy and the intervening sequences, but insertions templated from nearby DNA sequences and apparent blunt joins are also common. Thus alt-EJ is often associated with complex and highly mutagenic mutations including chromosomal translocations, that may cause cancerogenesis [121, 124, 125].

Alt-EJ occurs throughout the cell cycle but is enhanced in S and G2 phase and functions independently of core C-NHEJ and HR factors [107, 110, 121, 122, 126, 127]. However, to expose potential MHs, alt-EJ presumably shares the initial end resection step with HR, relying on the MRN-CtIP nuclease complex for limited 5' end resection [128]. The alt-EJ DSB repair family presumably encompasses two distinct subpathways, namely microhomology–mediated end-joining (MMEJ) and synthesis-dependent microhomology-mediated end-joining (SD-MMEJ).

MMEJ, probably the major alt-EJ pathway, depends on pre-existing small (5-25 bp)

MHs and always creates deletions [121] (Fig. 1.4). MMEJ resembles mechanistically the SSA pathway but has distinct genetic requirements. Poly(ADP-ribose) polymerase 1 (Parp1) is an early MMEJ factor that initially competes with Ku proteins for DSB binding and thereby antagonizes C-NHEJ repair [129]. Subsequently, it facilitates MRN-CtIP recruitment which initiates short end resection [130]. In addition, Parp1 is able to anneal complementary MHs and to recruit other alt-EJ factors, including the Ercc1-Xpf complex and the translesion DNA polymerase θ (Pol theta) [108, 131, 132]. Ercc1-Xpf is the major endonuclease to excise flap structures during alt-EJ [107, 133]. However, Artemis, a nuclease involved in resection-dependent C-NHEJ, together with 53BP1-Rif1 may also play a role [134]. Pol theta can dissociate both RPA and Rad51 molecules from ssDNA through its helicase activity and thus counteracts HR [116, 135, 136]. Moreover, Pol theta contributes to the annealing of MHs as well as to the gap-filling by low fidelity DNA synthesis [111, 137]. Besides Pol theta, other translesion DNA polymerases, such as the Pol η and Pol ζ, as well as Pold3 are reported to contribute to alt-EJ repair pathways in yeast [133, 138]. Finally, the Lig3-Xrcc1 complex, possibly assisted by Wrn, mediates end ligation [117, 128, 139]. It remains currently unclear in which repair context Parp1 and Pol theta contribute to MH annealing and whether their activities are redundant.

Unlike MMEJ, SD-MMEJ does not rely on pre-existing MHs but creates them *de novo* by limited DNA synthesis [122, 140] (Fig. 1.4). SD-MMEJ repair junctions are often accompanied by templated insertions and deletions but can also lead to apparent blunt joints. SD-MMEJ is considered to be a salvage pathway for resected ssDNA ends that fail to expose accessible MH sequences [119, 122]. Circumstantial evidence suggests that SD-MMEJ could even rescue Rad51-covered ssDNA ends in hamster and human cells in the absence of long homologous sequences as required for HR [116, 141]. Interestingly, SD-MMEJ pathways are associated with random plasmid integration in human and rodent cell genomes [141–143].

Pol theta is believed to be a key enzyme for SD-MMEJ repair, in addition to its proposed function in MMEJ [122, 137, 144, 145]. Pol theta might mediate *de novo* synthesis of MH sequences either by using transient secondary structures located in *cis* and/or minimally annealed broken DNA ends, possibly starting from as little as 1 bp homology, to prime non-processive DNA synthesis or by using its terminal transferase activity to introduce non-templated nucleotides [111, 122, 140, 145–147]. SD-MMEJ is completed by ligation using Lig3-Xrcc1 and/or Lig1 [148, 149].
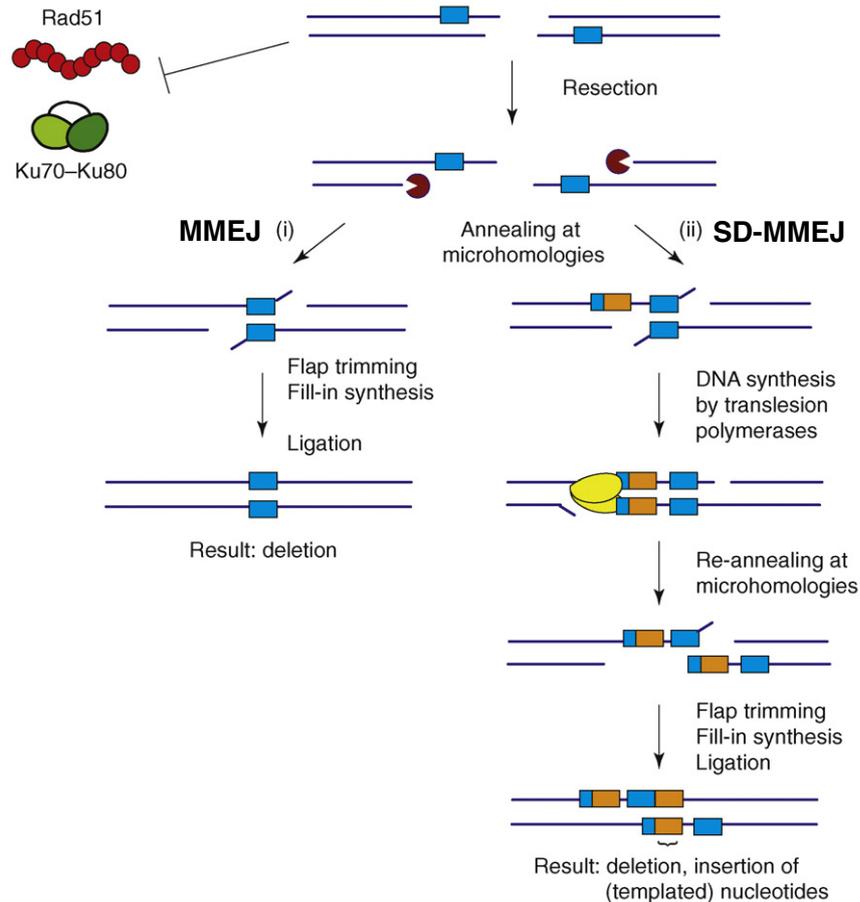
Figure 1.4: Model for alternative end-joining (alt-EJ) by microhomology-mediated end-joining (MMEJ) and synthesis-dependent microhomology-mediated end-joining (SD-MMEJ). (i) MMEJ uses long (5-25 bp) microhomologous sequences (blue boxes) to stably bridge the broken DNA ends, resulting in deletions. (ii) SD-MMEJ uses minimally ($\geq$1 bp) annealed sequences to prime templated synthesis by translesion DNA polymerases (yellow). Following unwinding of the initial priming microhomologies, the inserted nucleotides serve as *de novo* microhomologies for end bridging, resulting in indel repair junctions (adapted from [121]).

1.3 DSB repair pathway choice

To maintain genome integrity and cell viability, C-NHEJ, HR, SSA and alt-EJ DSB repair pathways act complementary. However, these pathways also compete for the same substrate, as deficiency in one pathway usually coincides with upregulation of the remaining DSB repair pathways [150–152]. Hence, DSB repair activities need to be strictly controlled depending on the repair context.

A major factor influencing DSB repair pathway choice is the cell cycle phase. Whereas

C-NHEJ is active throughout the cell cycle, alt-EJ, SSA and HR preferentially occur in S and G2 phase [21, 59, 126]. Among others, the phase of the cell cycle determines the availability of donor sequences. Sister chromatids – the preferred donor template for HR – are only present following replication in S phase and thus restrict HR to the S and G2 phase. Moreover, the cell cycle phase is a key determinant for DSB end resection. For instance, the activity of CtIP, a core factor for end resection, is cell cycle dependent. It is suppressed in G1 but activated by cyclin-dependent kinases in S and G2 phases, constraining extensive end resection to S and G2 phase [153–155]. Interestingly, MRN-CtIP-mediated limited end resection, which might be sufficient for most alt-EJ activities, remains possible in G1 phase [121, 154]. Therefore, DSB end resection represents a major regulation step in channeling DSB repair either towards resection-dependent repair pathways, i.e. HR, SSA and alt-EJ, or towards pathways that depend on no or little end resection, i.e. C-NHEJ.

DSB type and structure are other key factors for DSB repair pathway choice. C-NHEJ is typically used for simple or "clean" DSB ends, as arise upon endonuclease cleavage or treatment with topoisomerase II inhibitors, while HR is mostly deployed for complex or "dirty" DSB ends, as induced upon exposure to high-energy carbon ions [50]. In addition, programmed DSBs are often biased towards a specific repair pathway. For example, DSBs associated with V(D)J recombination are mostly repaired by C-NHEJ, while DSBs associated with meiosis are mainly repaired by HR [156, 157]. DSB structure also governs repair pathway choice, e.g. blunt-ended DSBs tend to C-NHEJ, while overhanging DSBs rather rely on alt-EJ and HR [158–160]. Further work will be needed to refine how overhang polarity and length govern alt-EJ and HR pathway choice.

Chromatin state and nuclear localization of the DSB also modulate DSB repair pathway choice. Unlike transcriptionally active regions (euchromatin), where both HR and C-NHEJ operate [52, 161], DSBs in heterochromatin are primarily repaired by HR [50, 162], possibly as the compact heterochromatin is inaccessible for fast repair by C-NHEJ. It has also been noted that DSB repair is compartmentalized: DSBs at the nuclear center and inner nuclear membrane are predominantly repaired by error-free HR, while DSBs at the nuclear periphery and pores undergo error-prone BIR or alt-EJ repair [163]

Finally, DSB repair pathway choice is influenced by the cell type or organism itself, in spite of the high conservation of repair proteins. As discussed previously, mammalian cells possess substantially lower HR but higher C-NHEJ levels than yeast cells, suggesting

that HR is suppressed in highly complex genomes to mitigate the risk of non-allelic re-combinations leading to deleterious chromosomal rearrangements [51, 53]. Moreover, intrinsic differences in DSB repair pathway activities also exist between mammalian model systems. For instance, rodents possess considerably lower Ku and DNA-PKcs expression and activity levels, and thus reduced C-NHEJ but elevated alt-EJ activities compared to human cells [164–166]. Differences have also been observed between stem cells, differentiated cells and cancer cells with an apparent shift towards more error-prone end-joining pathways along differentiation and immortalization [167–171].

1.4 Involvement of DSB repair proteins in other DNA repair pathways

In addition to the tight interconnection between the different DSB repair pathways, C-NHEJ, HR, SSA and alt-EJ also closely interact with and/or co-opt proteins from a multitude of other DNA damage repair pathways, including base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), interstrand crosslink (ICL) repair and single strand break (SSB) repair.

1.4.1 Base excision repair (BER)

Base excision repair (BER) corrects small base lesions arising from alkylation, oxidation and deamination that typically cause little distortion to the DNA double helix [172, 173]. BER is initiated by several DNA glycosylases that recognize specific types of base damage [174]. Upon recognition of the damaged base, the glycosylase cleaves the N-glycosidic bond that connects the damaged base with the sugar-phosphate backbone leading to the removal of the base and creating an abasic site, also known as apurinic or apyrimidinic (AP) site. An AP endonuclease, mainly Ape1 in mammals, subsequently introduces a nick in the phosphodiester bond 5' of the abasic site, generating a SSB with a 3'-OH and a 5'-deoxyribose phosphate (dRP) group [175]. This SSB, at least for a subset of base lesions, is bound by Parp1 to prevent DSB formation [176, 177].

BER subsequently progresses by either short-patch or long-patch BER pathways. In short-patch BER, DNA polymerase β fills in the single nucleotide gap while it removes

the 5'-dRP via its phosphodiesterase activity to restore a 5'-phosphate at the termini, as required for end ligation mediated by the Lig3-Xrcc1 complex [178, 179]. In long-patch BER, 2-10 nucleotides are replaced and newly inserted by displacing DNA synthesis co-ordinated by the DNA polymerases β, δ, ε and PCNA [180]. This generates a flap, which is cleaved by the Fen1 endonuclease [181], followed by Lig1-mediated end ligation [180].

1.4.2 Nucleotide excision repair (NER)

Unlike BER, nucleotide excision repair (NER) removes bulky, helix-distorting DNA lesions, such as DNA adducts and pyrimidine dimers, that arise from exposure to UV light, environmental mutagens and chemotherapeutic agents [182, 183].

NER consists of two subpathways, called transcription-coupled NER (TC-NER) and global genome NER (GG-NER), which differ in the initial damage recognition but share the same core NER factors to complete repair [184, 185]. TC-NER recognizes DNA lesions that interfere with progression of RNA polymerases. Stalled RNA polymerases initiate TC-NER by recruiting mediator proteins including Cockayne syndrome A and B (CSA and CSB). GG-NER, on the other hand, is initiated independently of transcription by the xeroderma pigmentosum complementation group C (XPC)-RAD23B dimer and the UV-damaged DNA-binding protein (UV-DDB) that scan throughout the genome for DNA helix distortions.

Following NER initiation, the multi-subunit transcription factor TFIIH localizes to the DNA damage along with xeroderma pigmentosum complementation group G (XPG) to unwind the strand DNA around the lesion. This unwinding of the DNA generates a bubble with a ssDNA stretch of approximately 30 nt. RPA binds shortly after TFIIH to protect the unwound and undamaged DNA strand. Finally, the Ercc1-Xpf endonuclease complex is recruited to the TFIIH complex. Ercc1-Xpf and XPG incise the damaged DNA strand upstream and downstream of the lesion, respectively, creating a gap of approximately 24-32 nucleotides [183, 186]. Loading of PCNA facilitates gap-filling by the DNA polymerases Pol δ, Pol ε and/or Pol κ, a translesion DNA polymerase [182]. Lig1 or Lig3-Xrcc1 complex seal the remaining nick [187]

1.4.3 Mismatch repair (MMR)

Mismatch repair (MMR) recognizes mismatches such as insertion/deletion loops or mispaired bases [188, 189]. MMR is mainly active to repair DNA polymerase errors that escape the polymerase proofreading activity, but is also involved in antibody diversification and influences the efficiency and fidelity of DNA recombination. During HR, for instance, MMR is anti-recombinogenic by suppressing recombination between homeologous DNA sequences via heteroduplex rejection but also repairs mismatched bases within heteroduplexes, possibly leading to gene conversion [82, 190]. MMR is usually strand-specific. This applies to both replication- and recombination-associated repair, which are biased towards repair of the newly replicated and the invading DNA strand, respectively [97, 99, 191].

MMR utilizes the Msh2-Msh6 (MutSα) heterodimer for detection of single base mismatches and small insertion/deletion loops but the Msh2-Msh3 (MutSβ) heterodimer for recognition of larger insertion/deletions loops [192]. The Mlh1-Pms2 heterodimer is subsequently recruited to the Msh2 complexes and translocates along the dsDNA to identify nicks, such as a gap between Okazaki fragments in the lagging strand [189]. This facilitates recruitment of PCNA and Exo1 nuclease leading to the nucleolytic degradation of the nicked strand across the insertion/deletion loop or mispair [193]. The resulting ssDNA gap is coated by RPA, filled in by DNA polymerases δ and ε and ligated.

1.4.4 Interstrand crosslink (ICL) repair

Interstrand crosslinks (ICLs) are lesions in which the Watson and Crick strand are covalently interconnected. ICLs are highly cytotoxic as they prevent DNA strand separation and therefore block DNA replication, recombination and RNA transcription. Several chemotherapeutic drugs are potent ICL-inducing agents, including mitomycin C (MMC) or cisplatin [48].

Outside of S phase, ICL repair depends on NER and translesion synthesis (TLS) [194]. ICLs are recognized by helix distortion and/or blocked RNA polymerases and incised by the structure-specific Ercc1-Xpf endonuclease, which creates one nicked DNA strand and one DNA strand with the ICL adduct, also known as ICL unhooking [48]. Translesion DNA polymerase Pol ζ fills in the gap across the unhooked ICL lesion followed by the

removal of the ICL adduct by NER [195].

In contrast, in S phase, ICL repair is replication-dependent involving a cascade of Fanconi anemia (FA), NER, TLS and HR repair mechanisms [194]. Stalled DNA replication forks are recognized by the FancM protein and stabilized by fork regression, possibly forming a so-called "chicken foot" structure. FancM recruits the FA core complex, consisting of at least seven FA proteins, as well as the Slx1-Slx4-Mus81-Eme1 nuclease complex, Ercc1-Xpf, the Blm-Topo3α-Rmi1-Rmi2 helicase-topoisomerase complex and other repair proteins [196]. These factors promote nucleolytic ICL unhooking followed by TLS using Rev1 and Pol ζ, possibly with the help of other translesion DNA polymerases, and removal of the ICL by NER [197, 198]. TLS creates a DSB in the second DNA strand that is repaired by HR using the newly synthesized strand as donor template [199].

1.4.5   Single strand break (SSB) repair

Single strand breaks (SSBs) are DNA lesions in which only one strand of the DNA double helix is broken, while the other remains intact. SSBs are generated directly by ROS and IR or indirectly as intermediates during BER [200]. SSBs can convert into DSBs, either when they cause replication fork stalling and collapse or when two closely-spaced SSBs locate in complementary DNA strands [201].

Parp1 is a key sensor for both SSB and DSB repair [129, 176, 177, 202]. During SSB, it presumably accelerates recruitment of other SSB repair proteins but seems not to be crucial for all BER-induced SSBs [176, 177, 203]. Upon Parp1 binding, Xrcc1 localizes to SSBs in a Parp-dependent manner and acts as a scaffold for end processing factors, including polynucleotide kinase 3-phosphatase (PNKP), Ape1 nuclease, Pol β and aprataxin (APTX) [200]. Following end processing, SSB and BER pathways converge and SSB repair is completed via short- or long-patch BER, as previously described.

1.5   Random and site-specific genome modifications

The ability to genetically modify a genome, also known as genome engineering, is one of the most widely used technique in molecular biology. Genome engineering deploys

the cellular DNA repair machinery (see previous sections) to inactivate, insert or correct genes using both non-targeted (random) and targeted (site-specific) approaches.

1.5.1   Random integration

Traditionally, expression of exogenous DNA sequences, termed transgenes, relies on the integration of vectors encoding the transgene at undefined and thus random genomic loci. Such gene transfer vectors are commonly derived from viruses, transposons or bacterial plasmids [204–206].

The most frequently used vectors for transgene delivery into mammalian cells are bacterial plasmids. They accept large transgene sizes and are easy to handle and produce. Unlike viral vectors and transposons that often integrate as single copies, plasmids integrate as multicopy array at a single or few genomic loci [207–209]. Random plasmid integration relies on the cellular repair machinery and the formation of spontaneous DSBs. Such a recombination event happens on average in one out of 10^3 mammalian cells [210].

Random plasmid integration presumably occurs in two steps: The first step involves the formation of so-called plasmid concatemers – long DNA molecules consisting of an array of multiple plasmid copies [141, 207]. These concatemers are believed to arise upon recombination of DNA plasmids using DSB repair pathways. HR, C-NHEJ as well as alt-EJ repair are associated with plasmid concatemerization, with apparent discrepancies between model systems [207, 208, 211–213]. In hamster cells, this process is independent of C-NHEJ, possibly antagonized by HR but relies on Pol theta and Lig3 repair factors, in line with a MMEJ- and/or SD-MMEJ-like mechanism [141]. The second step is the integration of plasmid concatemers into the host genome. Although C-NHEJ was traditionally considered to mediate chromosomal integration, growing evidence suggests an involvement of SD-MMEJ-like mechanisms with Pol theta as one of the key proteins in human and rodent cells [142, 143, 214–216]. Consistently, concatemer integration in hamster cells depends on Pol δ and Lig1 activities, in line with a SD-MMEJ process, though Rad51 appears also to be involved [141].

Although being relatively efficient, random plasmid integration remains restricted to transgene insertions and suffers from highly variable transgene expression. The observed variability in transgene expression is attributed to several parameters, including the vary-

ing number of integrated transgenes as well as the presence of bacterial sequences, which can lead to transgene silencing [209, 217]. Another important parameter is the transgene integration site, which may cause positional effects, such as transgene silencing in heterochromatin, and insertional mutagenesis, such as altered endogenous gene expression including oncogene activation [218].

1.5.2 Gene targeting

The limitations associated with random integration can be circumvented using gene targeting [219]. Gene targeting is a site-specific genome modification approach, allowing not only to insert but also to inactivate or correct genes in a predefined genetic environment. Hence, it may be particularly valuable for gene therapy approaches that involve the repair of gain-of-function disease-causing mutations. Gene targeting relies on precise HR-mediated recombination between a chromosomal DNA sequence (target site) and a homologous DNA molecule (donor). The donor is often a linearized DNA plasmid that shares extensive homology, typically at least 1 kb, with the target sequence [220].

Gene targeting is highly efficient in organisms with naturally high HR activity levels. This includes yeast, fungi as well as certain mammalian cells, such as mouse embryonic stem cells, chicken DT40 cells and human Nalm-6 cells [221–224]. However, due to low HR activity, gene targeting is inherently inefficient in most other mammalian cells with typical frequencies of 10^{-6} per cell [210]. Therefore, targeted genome modifications are usually outnumbered by random integrations by three orders of magnitude, hampering its broad application. Several strategies have thus been established to augment gene targeting. So far the most promising strategy is genome editing.

1.5.3 Genome editing

Genome editing describes genome engineering techniques that utilize synthetic endonucleases to induce site-specific DSBs at the locus to be modified [225, 226]. These DSBs trigger the cellular DSB repair pathways and thereby boost targeted genome modifications by several orders of magnitude compared to traditional HR-based gene targeting [226–229]. The most frequently used endonucleases are meganucleases, zinc finger nucleases (ZFNs), transcriptional activator-like effector nucleases (TALENs) and the RNA-guided

clustered regularly interspaced short palindromic repeat (CRISPR)-CRISPR-associated (Cas) nuclease systems [230–233].

1.5.3.1 Meganucleases, ZFNs and TALENs

Meganucleases, also known as homing endonucleases, are sequence-specific endonucleases that recognize DNA target sites, typically between 14 to 40 bp in length [230, 234]. The HO and I-SceI are the most frequently used meganucleases, though only I-SceI has been extensively tested in metazoans [235]. They were both discovered in budding yeast, where they mediate mating-type switching and mitochondrial intron mobility, respectively, by introducing DSBs with 4 nt 3' overhangs [236, 237]. HO and I-SceI recognize specific 24 bp and 18 bp sequences, respectively. However, both of these meganucleases are difficult to engineer so that they detect non-natural DNA recognition sites, and they have, thus, primarily been used as a tool to study DNA DSB repair mechanisms [235, 238].

Unlike meganucleases, ZFNs and TALENs are protein-guided endonucleases that possess programmable sequence-specific DNA recognition domains linked to non-specific FokI cleavage domains [226]. The FokI nuclease cleaves 4 nt 5' overhanging DSBs and is obligate dimeric. Formation of a functional FokI nuclease thus requires the binding of two ZFN or TALEN monomers, respectively, on opposite DNA strands in correct orientation and distance (Figs. 1.5A and 1.5B). Each ZFN monomer consists of a DNA binding domain composed of 3-4 zinc finger modules [231]. Each zinc finger recognizes 3 bp yielding a total DNA recognition length of 18-24 bp. On the other hand, the DNA binding domain of a TALEN monomer consists of 15-18 TALE repeat modules [239]. Each TALE recognizes a single bp summing up to a 30-36 bp recognition site. Individual zinc fingers and TALE repeats can be modularly assembled to create customized gene-specific recognition domains. However, the construction of such engineered recognition domains remains technically challenging and time-consuming [240].

1.5.3.2 CRISPR-Cas nuclease systems

As alternative to ZFNs and TALENs, the RNA-guided CRISPR-Cas nuclease systems offer superior ease of design combined with high cleavage efficiency [233, 242, 243]. These nucleases evolved as RNA-based adaptive immune system in bacteria and archaea
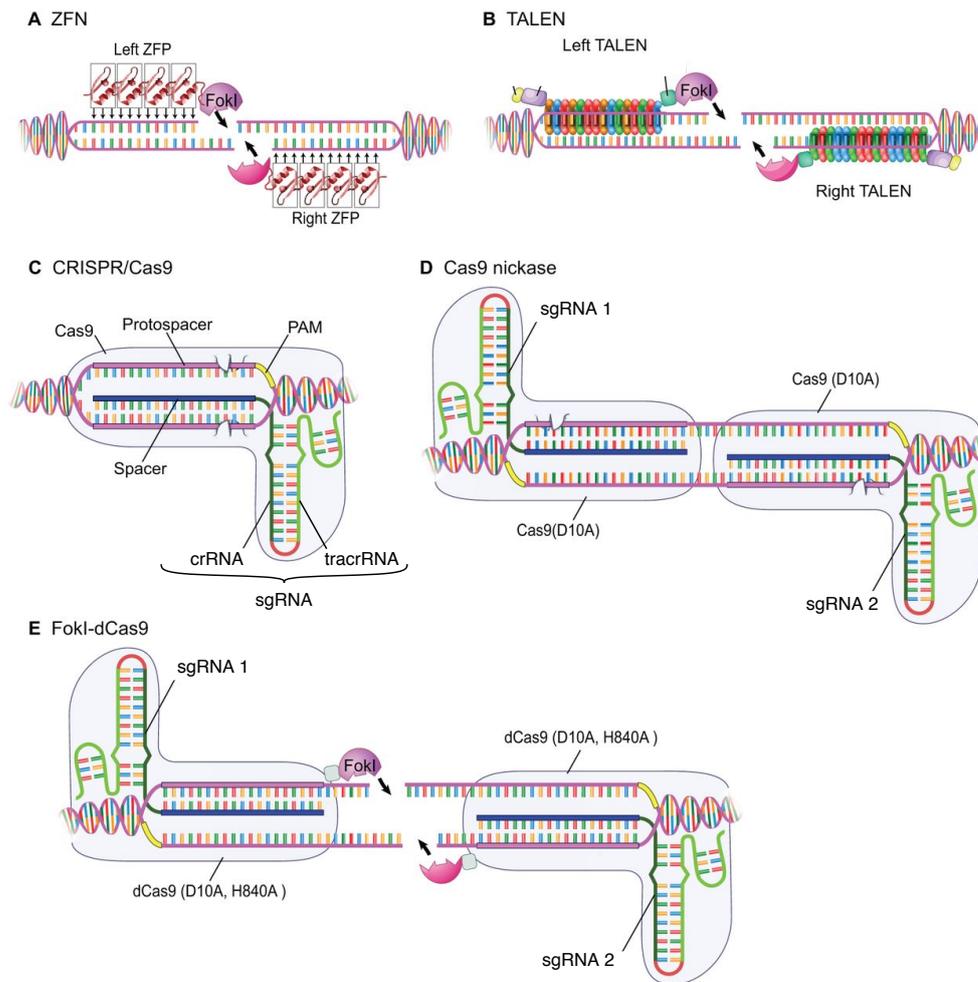
Figure 1.5: Schematic overview of commonly used genome editing nucleases. (A) Zinc finger nucleases (ZFNs) consist of left and right zinc finger subunits as DNA recognition sites linked to the obligate dimeric FokI endonuclease. (B) Transcription activator-like effector nucleases (TALENs) employ TALE repeats for DNA recognition and the dimeric FokI for cleavage. (C) The CRISPR-Cas9 system recognizes and cleaves a target sequence using a chimeric single guide RNA (sgRNA) bound to a Cas9 protein in the presence of a protospacer adjacent motif (PAM) at the 3' of the 20 nt sgRNA recognition site. The sgRNA consists of a site-specific CRISPR RNA (crRNA, in blue and dark green) and a trans-activating crRNA (tracrRNA, in light green) fused by a linker loop (in red). The two Cas9 nuclease domains (HNH and RuvC) are marked with a tooth-shaped structure. (D + E) Paired Cas9 nickases or the obligate dimeric catalytic inactive Cas9 fused to FokI (FokI-dCas9) enhance Cas9 target specificity (adapted from [241]).

to protect against invading viruses, plasmids and mobile elements and can cleave DNA as well as RNA sequences [244]. The basis of the CRISPR-mediated immunity is the CRISPR array consisting of spacer sequences interspaced by identical repeats. These spacers emerged from foreign nucleic acids acquired during previous infections. Array transcription produces long precursor CRISPR RNA (pre-crRNA) that can be processed

by cleaving within the repeat sequence to produce short mature crRNAs containing a single sequence-specific spacer flanked by repetitive sequences. Assembly of a crRNA with a Cas endonuclease forms a functionally active ribonucleoprotein complex, that is able to recognize and cleave complementary target DNA sequences, known as protospacers, guided by the crRNA. In comparison to ZFNs and TALENs that require cumbersome protein engineering, CRISPR-Cas target specificity depends on a short sequence-specific RNA molecule, which is relatively easy to customize and produce. Despite its prokaryotic origin, CRISPR-Cas has been successfully used to introduce site-specific modifications in a wide range of eukaryotes, including rodent and human cells, and is currently the most frequently used programmable nuclease system for genome editing applications [243, 245–247].

CRISPR-Cas systems are categorized into class 1 and class 2 depending on whether they employ a multi-subunit Cas complex or a single Cas protein, respectively, for crRNA binding and target cleavage [248, 249]. The simpler class 2 is mainly used for genome editing and consists of types II, V and VI. Type V includes DNA-targeting Cas12 nucleases, which are best suited for A-T rich genomes and introduce 5 and 7 nt 5' overhanging DSBs [250]. Type VI contains RNA-targeting Cas13 nucleases, which holds great potential for gene knockdown, RNA repair and RNA visualization studies [251, 252].

Class 2 Type II CRISPR systems rely on the Cas9 DNA endonucleases to interfere with invading foreign DNA sequences. Type II Cas9 requires a trans-activating crRNA (tracrRNA) to process pre-crRNAs into mature crRNAs and to form a crRNA:tracrRNA duplex which binds and guides the Cas9 nuclease to the target site [253] (Fig. 1.5C). For genome editing, the crRNA:tracrRNA duplex is often replaced by a chimeric single guide RNA (sgRNA) that can be easily expressed from a guide RNA expression plasmid with a customized crRNA and a constant tracrRNA [254]. The Cas9 protein from *Streptococcus pyogenes* (SpCas9) is the most thoroughly characterized and most frequently used CRISPR-Cas9 system for genome editing studies. Nevertheless, other Cas9 orthologs have also been tested successfully [255–257].

SpCas9 target recognition depends on a crRNA:tracrRNA duplex or an engineered sgRNA that includes a 20 nt guide sequence complementary to a target DNA sequence (protospacer) as well as a protospacer adjacent motif (PAM) that immediately follows the 3' end of the protospacer (Fig. 1.5C). The canonical PAM sequence in SpCas9 is 5'-NGG-3', but cleavage at non-canonical NAG and NGA PAM sites has also been detected [258,

259]. The Cas9:sgRNA ribonucleoprotein utilizes the PAM sequence to quickly probe throughout the genome to identify suitable target sites [260]. The presence of a PAM sequence initiates complementarity search between the crRNA and protospacer sequence and can lead to dsDNA unwinding upon formation of a stable RNA-DNA heteroduplex [260]. This conformational change activates the HNH and RuvC nuclease domains of SpCas9 to cut the target and non-target DNA strand, respectively, typically generating blunt-ended but sporadically also 1 nt 5' DSBs 3 bp upstream of the PAM sequence [261–263]. The Cas9 protein remains associated with the cleaved target site and only slowly dissociates, with the help of yet to be uncovered factors [261, 264–267].

A common concern in CRISPR-Cas9 genome editing is mutagenesis at unintended sites, known as off-target sites, due to the relatively short 20 nt target recognition length. Although reports are conflicting, off-target mutagenesis is typically considered rare ($<$ 0.5%) but was detected at sites with up to five mismatches and small indels [268–270]. Several new Cas9 variants have thus been engineered to improve Cas9 specificity. This includes high-fidelity Cas9 variants [271, 272], Cas9 nickases (Cas9n) [247, 273] as well as catalytically inactive Cas9 (dCas9) fused to the obligate dimeric FokI nuclease (FokI-dCas9) [274, 275] (Figs. 1.5D and 1.5E). Cas9n and FokI-dCas9 nucleases contain mutations in either the HNH (H840A) or the RuvC (D10A) nuclease domains and in both domains simultaneously, respectively, and thus required the pairing of two Cas9 molecules with proper orientation and spacing for DSB induction. All these methods impose more stringent constraints on-target recognition and thus substantially reduce Cas9 off-target mutagenesis, however, often at the expense of on-target cleavage frequency.

1.5.4 Genome editing outcomes and DSB repair pathways

Genome editing may result in gene inactivations, gene insertions or gene corrections [276]. However, its outcome considerably depends on the DSB repair pathways that become activated upon the nuclease-mediated DNA cleavage.

Gene inactivations, also known as gene knock-outs, most efficiently arise as a result of error-prone C-NHEJ and alt-EJ repair (Fig. 1.6A). These repair pathways often introduce variable indel mutations of various length at the DSB site, which may disrupt the coding sequence. Growing evidence suggests that the sequence flanking the DSB greatly influences indel mutagenesis and that appropriate sgRNA design might help to obtain

predictable out-of-frame mutations at frequencies as high as 80% [262, 263, 277–279].

Site-specific gene insertions, termed gene knock-ins, are mediated either through precise HR or imprecise C-NHEJ and alt-EJ repair pathways. HR-assisted gene insertion
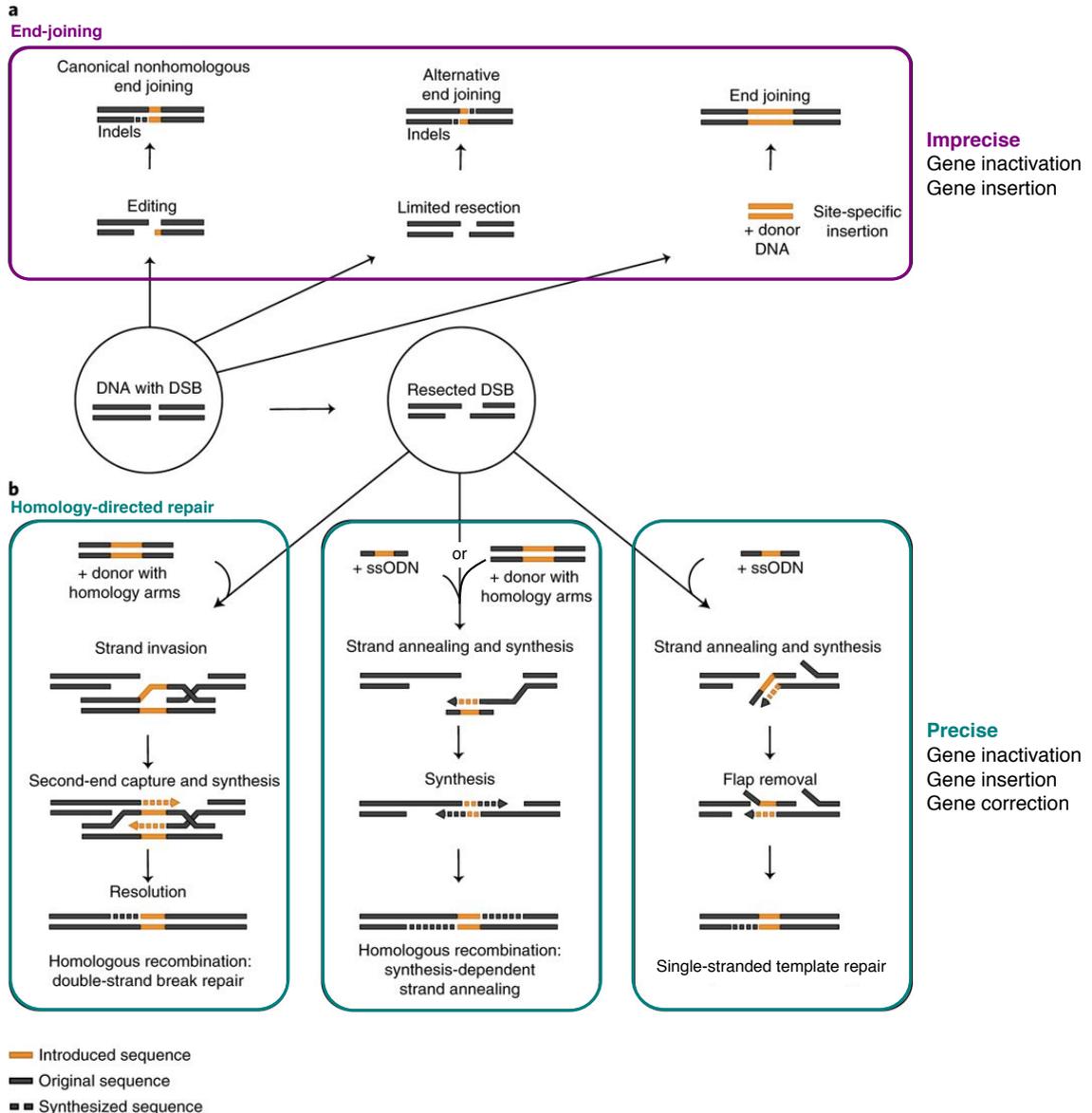
Figure 1.6: DSB repair pathways control genome editing outcomes. (A) End-joining pathways (i.e. classical non-homologous end-joining (C-NHEJ) and alternative end-joining (alt-EJ)) mediate imprecise gene inactivations and gene insertions. (B) Homology-directed repair (HDR) pathways (i.e. homologous recombination (HR) and single-stranded template repair (SSTR)) mediate precise gene inactivations, gene insertions and gene corrections. HR can proceed either via the double-strand break repair (DSBR) or the synthesis-dependent strand annealing (SDSA) pathway. Donor DNA consists either of a double-stranded DNA sequence (plasmid or PCR product) or a single-stranded oligodeoxyribonucleotide (ssODN) (adapted from [249]).

is the classical approach using targeting vectors that share substantial homology to the target site, analogously to gene targeting (Fig. 1.6B). However, the inherently low HR activity in most mammalian cells hampers efficient HR-mediated insertions [21, 23, 171]. Alternatively, the Obligate Ligation-Gated Recombination (ObLiGaRe), the Homology-Independent Targeted Integration (HITI) or the Precise Integration into Target Chromosomes (PITCh) methods exploit the more active but mutagenic C-NHEJ and MMEJ repair pathways to integrate transgene sequences that share no or ≤ 40 bp MHs, respectively, to the target site [280–283] (Fig. 1.6A). Although these non-HR methods elevate targeted insertions, integration is typically orientation-independent and imprecise with frequent indel mutations at the integration site.

Gene corrections rely exclusively on gene conversions mediated by homology-directed repair (HDR) pathways, an umbrella term encompassing HR and single-stranded template repair (SSTR) pathways [284] (Fig. 1.6B). HR uses dsDNA donors (plasmids or PCR products) with long homology arms to introduce large and/or multiple gene corrections [285]. Mechanistically, gene conversion using dsDNA donors is likely mediated by the SDSA repair pathway, but DSBR might also be active depending on the repair context and/or vector structure [286–289]. HR-mediated gene correction using dsDNA donors is inefficient in most cell types with typical frequencies of less than 5%, though frequencies can be considerably lower in cells recalcitrant to HR [171, 246, 247, 290, 291]. In contrast, single-stranded oligodeoxyribonucleotides (ssODNs) that possess only short homology arms (30-40 nt) are able to efficiently introduce point mutations or small genetic changes [292, 293]. Gene correction by ssODN donors presumably proceeds either via SDSA HR repair or SSTR (Fig. 1.6B). SSTR is a novel and poorly defined repair mechanism, which resembles the SSA pathway and is likely Rad51- and Brca2-independent [158, 294–296]. Although ssODNs are easier to produce and less likely to cause random integrations than dsDNA donors, they are limited in size (< 200 nt) precluding their application for larger genetic modifications and can as well cause imprecise HDR edits [293, 297].

1.6 Limitations of genome editing

Although programmable nucleases, notably the CRISPR-Cas9 nuclease system, have greatly advanced targeted genome engineering, some limitations remain. I discuss be-

low the limitations of inefficient HR-mediated genome editing and multi-locus genome editing. Besides the limitations discussed here, other limitations such as mosaicism, large deletions and rearrangements, *in vivo* delivery of engineered nucleases as well as ethical and regulatory concerns remain also to be addressed prior to clinical translation [249, 298–300]. However, these limitations are beyond the scope of this thesis.

1.6.1 HR-mediated precise genome editing

One major limitation is the overall low frequency of HR-mediated genome editing in mammalian cells. The most likely reasons for limited HR-mediated gene conversion are the relatively high activity of C-NHEJ and alt-EJ and the inherently low activity of HR in most cell types [51, 53, 171]. This imbalance in repair activity not only leads to a predominance of indel mutations at the target site but also favors illegitimate donor plasmid integrations into the target locus or any random genomic site instead of precise gene conversion [298, 301]. Thus, various strategies have been tested to overcome the bottlenecks of HR-mediated targeted genome modifications, both in the context of classical gene targeting and genome editing.

Some strategies have aimed at optimizing donor template design and availability for HR-mediated repair. Donor design strategies focused on the length and symmetry of homology arms to increase homology search and pairing, the linearization of donor plasmids to promote recombination by the free plasmid DNA ends, as well as the presence of positive and/or negative selection markers to enrich for correctly HR-targeted cells and to counter-select for illegitimate integrations [171, 222, 264, 268, 293, 302, 303]. Moreover, DNA repair templates were linked to the Cas9 ribonucleotide complex to locally augment the donor availability at the DSB site [304, 305].

Other strategies have assessed the effect of DSB structure and timing on HR-mediated repair. For instance, Cas9n-mediated DNA nicks, which are a poor substrate for C-NHEJ and alt-EJ but a substrate for HR, considerably reduced indel frequency at the target site, even if this approach also lowered HR frequency [247]. In addition, DSBs with long 5' overhangs created by two staggered Cas9n favored HR despite an overall low HR frequency [273, 306]. Furthermore, it was attempted to coordinate delivery and/or expression of the Cas9 protein so to obtain highest nuclease activity in S/G2 phase, when HR is active. However, the results were conflicting [307–309]. Finally, base editors were devel-

oped that circumvent DSB intermediates and rather use BER to catalyze specific single base pair corrections [290].

Other promising strategies have focused on interfering with DSB repair activities. The discovery of mammalian cells with high HR activity as well as the apparent competition between DSB repair mechanisms (see Sections 1.5.2 and 1.3) nurtured the idea that HR repair may be boosted by altering DBS repair pathway choice. To do so, several studies tried to overexpressed HR factors, in particular Rad51, or enhance HR using small molecules leading to modestly elevated intrachromosomal repair and/or gene conversion frequencies [310–315]. However, one study noticed reduced HR frequencies following Rad51 or Rad52 overexpression [316]. Many other studies transiently knocked down and/or knocked out C-NHEJ factors, e.g. Ku and Lig4 proteins, or inhibited C-NHEJ using small molecules [141, 317–320]. Interestingly, while some studies observed an increase in gene conversion, others did not, suggesting that alt-EJ rather than HR mechanisms compensate for deficiency in C-NHEJ repair. Moreover, activating HR and/or suppressing C-NHEJ typically did not alter random donor integration levels, in line with it being alt-EJ-dependent [141, 311, 317]. Comprehensive studies about the contribution of alt-EJ pathways to HR-mediate precise gene targeting are missing with only two studies trying to bypass random donor integrations by inactivating C-NHEJ and alt-EJ in parallel [142, 143].

In sum, these strategies typically increased HR-mediated repair by approximately 2- to 10-fold, which might still be too low for clinical applications. Furthermore, the effect of HR stimulation, notably when altering DSB repair activities, was highly variable among distinct cell types and/or organisms [309, 321], highlighting the intrinsic differences in activity and/or regulation of DSB repair pathways between model systems.

1.6.2 Multi-locus genome editing

Another challenge is the simultaneous generation of genome modifications at multiple genomic loci, also termed multiplexing. Multi-locus genome editing comprises mutagenesis of multiple alleles in multiploid organisms, multi-copy genes (e.g. ribosomal RNA genes) and highly repetitive elements (e.g. endogenous viral sequences) using a single nuclease as well as mutagenesis of multiple distinct genes using several nucleases [322]. Due to the abundance of target sequences, mutagenesis efficiency of an individual locus is

typically reduced compared to single locus genome editing [323–325]. Stable expression of the nuclease system helped to partially overcome the inefficient multi-locus editing of endogenous viral sequences in porcine cells [325]. However, the resulting high nuclease activity introduces a multitude of DSBs in parallel. Therefore, it induces a strong DDR which possibly provokes cellular stress responses, growth arrest, cytotoxicity and chromosomal rearrangements [325–328].

1.7 Cells as expression host for biopharmaceuticals

In addition to traditional small-molecule drugs produced by chemical synthesis, protein-based drugs – also known as biopharmaceuticals – are manufactured in living cells. Typcial biopharmaceutical products include monoclonal antibodies, hormones and cytokines [329]. There are currently more than 300 biopharmaceuticals on the market and many more are currently tested in clinical trials [330]. While non-mammalian cells (e.g. bacteria and yeast) are used to produce simple therapeutic proteins, mammalian cells are required to produce large and complex therapeutic proteins with post-translational modifications (PTMs). PTMs, including glycosylation and disulfide bonds, are required to ensure proper protein folding, stability and activity of most secreted proteins, such as antibodies [331].

Chinese hamster ovary (CHO) cells are the most frequently used mammalian cell expression system [329]. Besides their ability to produce therapeutic proteins with human-like PTMs, these immortalized cells show high transfection efficiencies, high levels of random plasmid integrations and are able to grow in suspension culture as well as in serum-free chemically defined media, allowing for reproducible high-density large-scale protein production [329]. Moreover, CHO cells seem to be less susceptible to certain viral infections compared to other mammalian cells [332, 333].

1.7.1 Viral contaminations

Although mammalian cell expression systems, notably CHO cells, allow the production of therapeutic proteins with human-like PTMs at high yields, they are susceptible to contaminations with adventitious agents, in particular viruses. Until today, some incidences of viral contaminations during biomanufacturing have been reported, with every incidence

potentially threatening patient safety and/or drug supply [334]. Thus, to ensure absence of viral contaminations during biomanufacturing as well as in the final biopharmaceutical product, regulatory bodies require ample precautions, time-consuming and cost-intensive viral removal and inactivation steps followed by comprehensive testing [335].

Viral contaminations usually arise from exogenous sources. This includes contaminated raw material, notably animal-derived components, serum and cell culture medium, but also inappropriate handling of raw material, cells and bioreactors [336, 337]. However, the cell itself is also a potential source of viral contamination. More precisely, all mammalian cells harbor large reservoirs of stably integrated endogenous viruses. These viruses may produce viral-like particles (VLPs), which can subsequently contaminate the biopharmaceutical product [338].

1.7.2 Endogenous retroviruses

Endogenous retroviruses (ERVs) are the most abundant endogenous viral elements in mammals. They can be found in thousands of copies and are estimated to make up to 8% and 10% of the human and mouse genomes, respectively [339–342]. ERVs are remnants of previous retroviral integrations into the host germline and therefore transmitted vertically to the offspring. Unlike DNA and other RNA viruses, retroviruses are (+)ssRNA viruses that naturally integrate into the host genome as part of their viral life cycle [343]. They are able to reverse transcribe their RNA genome into DNA followed by the integration of this viral DNA intermediate into the host genome using the virally encoded reverse transcriptase and integrase enzymes, respectively [344] (Fig. 1.7A). The integrated virus is called provirus.

The size of an ERV genome is ∼7-11 kb and is composed of the *gag*, *pol* and *env* genes in that order flanked by two long terminal repeats (LTRs) [345]. The *gag* gene codes for the matrix, capsid and nucleocapsid that form the structural proteins of the viral core. The *pol* gene codes for the protease, reverse transcriptase and integrase, all of which fulfill key functions during the retroviral life cycle and belong to the most conserved viral genes [346]. Finally, the *env* gene encodes the surface and transmembrane proteins located in the viral envelope that determine viral tropism. Gag, Pol and Env expression is controlled by the LTRs which encompass enhancer and promoter sequences and the polyadenylation signal [347]. Gag and Pol proteins are translated from a single full-length Gag-Pol-Env
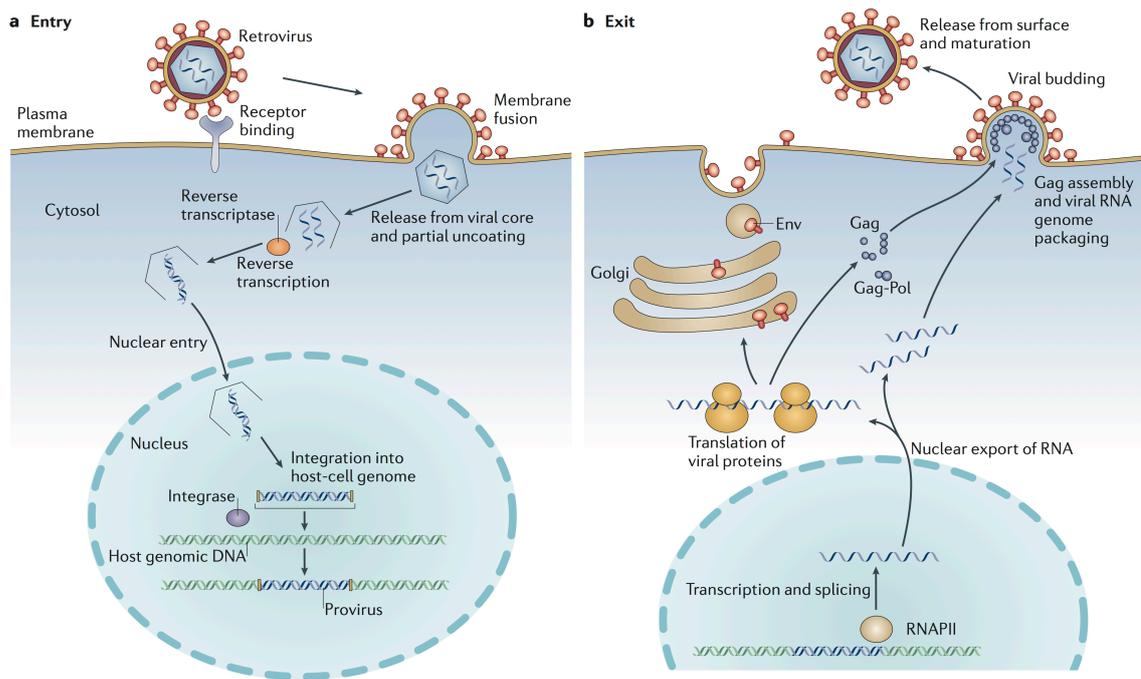
Figure 1.7: Retroviral life cycle. (A) Formation of stably integrated retroviruses (proviruses) requires the virally encoded reverse transcriptase and integrase to produce a DNA intermediate that can integrate into the DNA of the host. Integration into the germline yields endogenous retroviruses (ERVs), which may be fixed in the population (B) Production of infective viral particles depends on expression of *gag*, *pol* and *env* genes by the host cell (reproduced from [344]).

precursor mRNA. While the Gag protein is highly expressed, Pol RNA is only translated following termination suppression or ribosomal frameshifting occurring in ∼5% of the transcripts [348]. Env translation occurs from a spliced mRNA.

Despite the large number of ERV sequences in mammalian genomes, ERVs usually do not produce infective viral particles. This is due to two reasons. First, most ERVs are epigenetically silenced, which prevents ERV transcription and thus suppresses further infections [349]. Second, ERVs often accumulate deleterious mutations and recombination events [341, 350]. For instance, the two LTR sequences frequently recombine by SSA repair, which deletes the entire ERV coding sequence and creates so-called solo LTRs [101].

Interestingly, most mammalian genomes, including mouse, pig and hamster, still contain full-length proviruses and/or transcribed ERVs [341, 344, 351–353]. For example, CHO cells possess one or few ERVs able to form VLPs, which are released in the cell culture supernatant. However, CHO-derived VLPs have never been shown to be infective [333, 354]. Unlike CHO cells, porcine cells still contain ERVs that can produce particles

capable of infecting human cells *in vitro* [326]. Such infective particles are a major concern for pig-to-human xenotransplantation [326]. To prevent porcine ERV transmission to humans, genome editing strategies were developed to inactivate the numerous proviral sequences [325–327]. However, these strategies suffered from the limitations associated with multi-locus genome editing described in detail in Section 1.6.2.

The uncountable ERV sequences in mammalian cells are a latent danger for the production of biopharmaceuticals. Although most of them are currently inactive and/or produce defective VLPs, ERVs have been shown to be able to form again infective viral particles. For instance, epigenetically silenced ERVs may awake in response to chemical treatments, changes in cell culture conditions or novel viral infections [355–358]. Moreover, defective ERVs can acquire gain-of-function mutations or recombine with other non-allelic ERV sequences, forming chimeric functional ERV sequences [338, 346]. Finally, reactivation of virus expression may also result from *trans* complementation of individually defective ERVs [359].

Chapter 2

Aim of the thesis

The CRISPR nuclease system is currently revolutionizing the field of genome editing. It can be easily programmed to introduce site-specific DNA double strand breaks (DSBs) and thereby considerably boosts targeted genome modifications in mammalian cells. While CRISPR-mediated mutagenesis is unprecedentedly efficient in inactivating single genes, precise genome editing mediated by homologous recombination (HR) DSB repair has remained inefficient, possibly due to competing DNA repair pathways. As of today, no universal strategy has been discovered to augment HR-mediated repair. This may result from incomplete characterization of DSB repair pathways as well as possible intrinsic differences in DSB repair activity and/or regulation in various model systems.

Therefore, the main aim of the thesis was to better understand DSB repair and recombination pathways in Chinese hamster ovary (CHO) cells, so as to provide novel strategies for improving genome editing. Unlike primary cells, CHO cells are believed to have reduced classical non-homologous end-joining (C-NHEJ) as well as low HR repair activities [164, 171, 291]. In addition, recent data from our lab indicated that a family of as yet poorly described alternative end-joining (alt-EJ) DSB repair pathways is highly active and a key mediator for random genomic integrations of plasmids in CHO cells [119, 141]. Thus, the predominant alt-EJ activity may explain the moderate increase in gene targeting previously observed in C-NHEJ-deficient CHO background [318]. Taken together, this raised the hypothesis that high alt-EJ activities, as observed in these cells, compete with and ultimately restrict precise HR-mediated genome editing. Interestingly, comprehensive studies about the potential function of alt-EJ and other competing factors on HR-mediated genome editing were missing in CHO as well as in other mammalian

33

cells.

Hence, this thesis aimed at identifying rate-limiting factors for HR-mediated precise genome editing in CHO cells. The initial steps towards this aim were to design a fluorescence-based HR assay and to evaluate the effect of transient depletion of DSB repair factors on HR frequency. These steps also allowed us to test the hypothesis that the predominance of alt-EJ activities is one of the rate-limiting factors for HR. To further broaden the understanding of CHO DSB repair pathways, we explored how simultaneous depletion of multiple repair factors as well as overexpression of rate-limiting factors influence HR frequency. We also analyzed whether the contribution of repair factors depended on the CRISPR nuclease type, the HR assay or the cell line.

The secondary goal of this thesis was to establish a genome editing strategy to inactivate repetitive endogenous retroviruses (ERVs) in CHO cells and to prevent them from producing viral-like particles (VLPs). We hypothesized that the improved understanding of CHO DSB repair mechanisms may be useful to achieve gene inactivations at multi-locus ERV sites, thereby allowing to generate CHO cells with an increased safety profile for biopharmaceutical production. Overall, this thesis constitutes a further step to understand DSB repair mechanisms and to mediate more efficient genome editing in CHO and possibly other mammalian cells.

Chapter 3

A role for alternative end-joining factors in homologous recombination and genome editing

This chapter is based on a manuscript submitted for publication entitled "A role for alternative end-joining factors in homologous recombination and genome editing" by **Bosshard S.** and Mermod N.

3.1 Abstract

CRISPR technologies greatly foster genome editing in mammalian cells through site-directed DNA double strand breaks (DSBs). However, precise editing outcomes, as mediated by homologous recombination (HR) repair, are typically infrequent and outnumbered by undesired genome alterations. By using knockdown and overexpression studies in Chinese hamster ovary cells as well as characterizing repaired DNA junctions, we found that efficient HR-mediated genome editing depends on alternative end-joining (alt-EJ) DNA repair activities, a family of incompletely characterized DNA repair pathways traditionally considered to oppose HR. This dependency was influenced by the CRISPR nuclease type and the DSB-to-mutation distance, but not by the DNA sequence surrounding the DSBs or reporter cell line. We also identified elevated Mre11 and Pari, and low Rad51 expression levels as the most rate-limiting factors for HR, and counteracting these three bottlenecks improved precise genome editing by up to 75%. Altogether, our study provides novel insights into the complex interplay of alt-EJ and HR repair pathways, highlighting their relevance for developing improved genome editing strategies.

3.2 Introduction

Programmable nucleases, particularly the CRISPR RNA-guided nuclease system, greatly facilitate targeted genome modifications and thereby revolutionize genome editing in mammalian cells [233]. While CRISPR-mediated indel mutagenesis and resulting gene inactivation is highly efficient in most cells lines, precise gene insertion and gene correction have remained inefficient and highly variable, with typical frequencies of less than 5% [246, 247, 290]. However, the growing interest in accurate and predictable genome editing for research and translational applications makes the development of strategies for improved precise genome editing highly desirable.

Genome editing outcomes critically depend on DNA double strand break (DSB) repair pathways, which are activated in response to the CRISPR-mediated cleavage. Gene inactivation mostly arises from the mutagenic classical non-homologous end-joining (C-NHEJ) and alternative end-joining (alt-EJ) DSB repair pathways. C-NHEJ, the predominant DSB repair pathway in most mammalian cells, is a fast repair mechanism that ligates DSB ends together using no or minimal end processing and generally leads to small 1-4 nt insertions or deletions at the repair site due to non-compatible DSB ends [44]. Core C-NHEJ proteins include the Ku70-Ku80 heterodimer and the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) that protect and bridge the DNA strand extremities prior to end ligation by the Ligase4-Xrcc4 complex.

In addition to C-NHEJ, evidence suggests that many cancer and immortalized cell lines possess supplementary end-joining pathways, collectively termed alt-EJ, to manage increased DSB levels [116, 118, 119]. The alt-EJ repair family is incompletely characterized but likely comprises pathways such as microhomology-mediated end-joining (MMEJ) and synthesis-dependent MMEJ (SD-MMEJ), which utilize microhomologies at or near the DSB to anneal complementary bases at broken DNA ends [121, 122]. Annealing of these microhomologies often results in the deletion of one microhomology copy and intervening sequences. However, the insertion of bases replicated from nearby DNA sequences, termed templated inserts, are also commonly observed at alt-EJ repair junctions, for instance following SD-MMEJ [122]. Moreover, alt-EJ pathways were associated with random plasmid integration in human and rodent cell genomes [141–143].

Microhomologies required for alt-EJ become exposed after DSB end processing and

single strand resection, presumably initiated by the Mre11-Rad50-Nbs1 (MRN)-CtIP complex [128]. These microhomologies can range from as little as 1 bp to 25 bp, but may also be created de novo by limited DNA synthesis of DNA Polymerase θ (Pol theta) during SD-MMEJ, if resection fails to expose accessible microhomologies [121–123]. Following end resection, Poly(ADP-ribose) polymerase 1 (Parp1) and Pol theta mediate the annealing of microhomologies, which typically produces protruding non-homologous DNA tails, termed flaps [108, 111]. Subsequently, the endonuclease complex Ercc1-Xpf can excise these flap structures and gaps can be filled by the non-processive polymerase activity of Pol theta and Polymerase δ subunit 3 (Pold3) [107, 111, 133]. Alt-EJ is completed by end ligation mediated by Ligase I (Lig1) and the Ligase III (Lig3)-Xrcc1 complex [128, 149].

On the other hand, accurate but infrequent targeted gene insertion and gene correction rely on the homologous recombination (HR) DSB repair pathway. HR is a relatively error-free but more complex and tightly regulated DSB repair pathway. It uses long homologous sequences as template for repair, preferentially from the sister chromatid, and is therefore mainly active in late S/G2 phase. However, HR accounts for only 15-20% of DSB repair in G2 phase human cells [52]. This implies that mutagenic end-joining pathways are more efficient even in the presence of active HR repair. To reveal homologous sequences, HR relies on end resection initiated by the MRN-CtIP complex – a step shared with the alt-EJ repair family. However, unlike the limited resection of alt-EJ, HR requires extensive end resection to expose extended 3' single-stranded DNA (ssDNA) ends. These ssDNA ends are coated by the Rad51 recombinase, which mediates homology search and strand invasion of homologous sequences, as a prerequisite to DNA synthesis. Interestingly, circumstantial evidence suggested that Rad51-coated ssDNA ends can be directed into alt-EJ repair pathways in the absence of sufficiently long homologies [116, 141]. Canonical HR is most frequently completed either by synthesis-dependent strand annealing (SDSA) or canonical DSB repair (DSBR) pathways, both of which can lead to a nonreciprocal transfer of the donor template sequence to the damaged chromosome and hence may give rise to gene conversion events.

The overall low frequency of HR-mediated genome editing in mammalian cells has been associated with two main factors: one being the prevalence of C-NHEJ- and alt-EJ-mediated mutations at the CRISPR target site [298], and the other being illegitimate donor integrations into the target locus or any random genomic site instead of gene conversion [301]. Thus, numerous studies have described strategies to improve HR-directed

repair. Some strategies have focused on improving HR donor design and availability [264, 268, 304], as well as DSB structure and the timing of DSB formation relative to the cell cycle [247, 307]. Other promising strategies have focused on interfering with DSB repair activities, either by overexpressing HR factors [310, 313], or by the transient knockdown or knock-out of C-NHEJ factors [318, 319]. So far, few studies tested the combination of interfering with specific activities of both alt-EJ and C-NHEJ factors to bypass random donor integrations [142, 143], and comprehensive analysis of the roles of these alternative DSB repair activities is lacking. Finally, small molecules were used to bias DSB repair in favor of HR [314]. While all these strategies improved HR frequency by approximately 2- to 10-fold, the DSB repair outcome was still limited by the persistent predominance of imprecise editing mechanisms, as well as by an overall low reproducibility when comparing distinct cell types [309, 321]. The lack of universal strategies to boost HR-mediated repair highlights intrinsic differences between model systems, an incomplete understanding of DSB repair mechanisms, and a possible interplay of the alt-EJ and HR pathways. Despite recent progress, improving HR-mediated genome editing thus remains a major challenge.

Here, we aimed at systematically characterizing the effects and possible interconnections of alt-EJ and HR repair pathways during precise genome editing in Chinese hamster ovary (CHO) cells, the most widely used host for recombinant therapeutic protein production. CHO cells, as rodent cells in general, show reduced C-NHEJ activities compared to human cells [164]. In addition, recent evidence suggested that CHO cells have elevated alt-EJ activities [119, 141], which consequently results in high random integration frequencies, but low HR activities [171, 268, 291]. CHO cells are therefore an ideal model system to study the hypothesis of whether high alt-EJ activities, as observed in these cells, compete with and ultimately restrict precise HR-mediated genome editing. Surprisingly, we find in this study that several alt-EJ factors critically contribute to efficient HR, as measured by different chromosomal gene correction assays. This contribution was most evident at staggered DSBs induced by the highly specific FokI-dCas9 nuclease, but it was also detected from blunt-ended DSBs induced by the wild-type Cas9 nuclease when the DSB is distant from the mutation. These findings reveal an interplay between alt-EJ and HR pathways for the repair of various DSB structures that may extend beyond sharing the initial end resection and Rad51 nucleofilaments. Moreover, this study led to the identification of specific rate-limiting factors for HR in CHO cells. Counteracting these limitations increased gene correction by up to 75%, offering a translatable strategy to increase precise

genome editing.

3.3   Material and Methods

siRNA and plasmids

Small interfering RNA (siRNA) duplexes were designed to target CHO DNA repair protein mRNAs and were provided by Microsynth AG (Balgach, Switzerland). For each target, three siRNAs were designed to increase the efficiency and specificity of the target mRNA depletion. Three negative non-targeting siRNAs (siNeg) were designed as controls. The sequences of siRNAs used in this study are listed in Table S3.1 or were previously published [141].

The mammalian codon optimized *Streptococcus pyogenes* Cas9 nuclease (Addgene plasmid # 43861) [360] was used as wild-type Cas9 nuclease, and the FokI nuclease fusion to the catalytically inactive Cas9 (FokI-dCas9; Addgene plasmid # 52970) [275] as high-fidelity nuclease. The GFP-specific single sgRNA (G1, G3, G7) and paired sgRNA (G1G5, G3G7) expression plasmids were kindly provided by David Liu [275] (Fig. S3.1A). To construct the GFP gene correction vector (HR reporter), two stop codons and a 101 bp deletion were introduced into the GFP coding sequence by overlapping PCR as previously described [361]. In brief, primers HRassay_F1 and HRassay_R1, and primers HRassay_F2 and HRassay_R2, respectively, were used for the first and primers HRassay_F1 and HRassay_R2 for the second PCR round (Table S3.2). The insertion of the second stop codons creates an AflII recognition site (CTTAAG), which is absent in the corrected GFP sequence. This PCR product was inserted in the HindIII/XbaI sites from a PiggyBac (PB) transposon vector [362], to contain a puromycin and the mutated GFP expression cassette surrounded by the PB inverted terminal repeats. Transient transfection of this HR reporter construct does not result in GFP positive (GFP+) cells, even in the presence of a CRISPR nuclease.

To generate GFP donor vectors, two silent PAM mutations per donor construct were first introduced into the PB transposon vector using overlapping PCR [361], to prevent CRISPR cleavage of corrected GFP sequences. Briefly, primers GFP_F3 and G1_R, and primers G1_F and GFP_R4h were used to mutate the G1 sgRNA PAM site (GGC>TGC)

in the first PCR round. These PCR products were then combined using primers GFP_F3 and GFP_R4h in a second PCR round and used to replace the wild-type GFP sequence using HindIII/XbaI sites in the PB transposon vector. This mutagenesis procedure was repeated to introduce the second silent PAM mutation for the G5 sgRNA (AGG>AAG) and to create a second GFP donor sequence containing G3 and G7 sgRNA-specific PAM mutations (GGG>GTG and TGG>TCG, respectively) (Fig. S3.1B). These silently mutated GFP sequences were then digested with HindIII and AgeI, end-filled using Klenow fragment and blunt cloned into the SmaI site of the pUC19 vector. To create a translational trap, the translation initiation codon was removed by PCR amplification of the PAM-mutated GFP sequence using GFP_Amut_KpnI_F and GFP_Amut_R primers followed by KpnI/NotI digestion to replace the previously inserted GFP sequence in the pUC19 vector.

Protein overexpression vectors were constructed as follows. The Rad51, Ercc1, Pold3, Lig3, Ku70 and Ku80 CHO cDNAs were inserted instead of the GFP sequence of a previously described mammalian expression vector that contains a GAPDH promoter upstream of the coding sequence [363]. The Neomycin expression plasmid was similarly generated by PCR amplification of the Neomycin cassette from the pCMV-DsRed-Express plasmid (Clonetech). The dominant-negative Rad51-K133A (AAG>GCC) and Rad51-K133R (AAG>AGG) expression plasmids were derived from the Rad51 overexpression vector using overlapping PCR [361]. The empty vector control was generated by GFP coding sequence excision and consists of the plasmid backbone only. All primers (purchased from Microsynth AG, Balgach, Switzerland) and plasmids used in this study are listed in Tables S3.2 and S3.3.

Cell lines

Suspension-adapted CHO-K1 cells were maintained in serum-free HyClone SFM4CHO medium supplemented with HyClone Cell boost 5 supplement (GE Healthcare), with addition of L-glutamine and HT supplement (Gibco). Antibiotic-antimycotic solution (Gibco) was added for recovery after cell sorting experiments.

The CHO-K1 HR reporter cell line was established as follows. CHO-K1 cells were transfected with the HR reporter PB transposon and a PB transposase expression vector [362] using the Neon electroporation system (Thermo Fisher Scientific), as per the manufacturer's instructions. Two days after transfection, cells were transferred to cul-

ture medium containing 4 μg/ml puromycin and selected for two weeks. 64 cell clones containing stable HR reporter integrations were isolated by limiting dilution, expanded and assessed for HR reporter copy number. A cell clone containing approximately four copies of the HR reporter per diploid genome, as assessed by quantitative PCR (qPCR), was selected for further experiments and cultured under continuous puromycin selection.

Adherent CHO-DG44 cells were cultivated in DMEM/F-12+GlutaMAX supplemented with 10% FBS, HT supplement and antibiotic-antimycotic solution (Gibco). The CHO-DG44 HR reporter cell line was derived from a GFP+ CHO-DG44 clone expressing a low copy number of the described PB transposon vector (a gift from Solenne Bire). GFP negative (GFP-) cells were isolated using single cell sorting to obtain GFP loss-of-function mutations resulting from natural mutagenesis of the GFP sequence. A GFP- cell clone containing a 215 bp GFP deletion compatible with the wild-type Cas9 and the FokI-dCas9 nuclease systems was selected for further experiments.

PB copy number

To analyze stably integrated PB copy number, total genomic DNA was extracted from CHO cells using the DNeasy Bood & Tissue Kit (Qiagen). 15 ng of genomic DNA was used for qPCR assays using the SYBR Green I Master mix for the Roche LightCycler 480 instrument, and each sample was analyzed in triplicate. PB transposon integrations were quantified using puromycin-specific primers (Table S3.2). Beta-2-microglobulin (B2M), a housekeeping gene reported to be present in two copies per CHO diploid genome [362], was used as reference gene. PB copy number was calculated using a previously described relative quantification method that accounts for different primer efficiencies [364].

sgRNA cleavage efficiency

To assess the cleavage efficiency of the previously published GFP-specific sgRNA [275] in CHO cells, 700,000 GFP+ CHO-DG44 cells were electroporated with either 690 ng of a single sgRNA expression plasmid (G1, G3, G7) and 2300 ng of wild-type Cas9 expression plasmid, or 900 ng of paired sgRNA expression plasmids (G1G5, G3G7) and 2500 ng of FokI-dCas9 expression plasmid, together with 200 ng of pCMV-DsRed-Express plasmid as transfection control. All CRISPR plasmids were used at equimolar ratio and

samples were complemented with inert pUC19 plasmid to reach a total plasmid amount of 5000 ng. For negative CRISPR control experiments, the sgRNA and/or nuclease expression plasmids were substituted with pUC19 plasmid. Disruption of the GFP coding sequence, a read-out for efficient CRISPR cleavage, was quantified three days after transfection by flow cytometry and fluorescence microscopy. The sgRNA cleavage efficiencies were higher using the wild-type Cas9 than the FokI-dCas9 nuclease system, in line with previously reported frequencies in human cells [275].

HR assays

To characterize HR-mediated genome editing, CHO-K1 HR reporter cells were seeded at 350,000 cells/ml in puromycin-containing medium one day prior to transfection, and transferred to puromycin-free medium after transfection. For siRNA transfection, 400,000 cells were electroporated with 100 nM of pooled siRNAs (mix of three siRNA, each siRNA at 33 nM) for the initial siRNA screen or with 100 nM of single siRNAs for deconvolution experiments. For combined knockdowns of several targets, 33 nM of the most efficient siRNAs were used and complemented with siNeg to 100 nM, if needed. For double siRNA transfections, cells were re-transfected with 100 nM of the previously transfected siRNA mix two days after the first siRNA transfection. For siRNA knockdown in CHO-DG44 cells, 200,000 CHO-DG44 HR reporter cells were reverse transfected with 50 nM of pooled siRNAs using Lipofectamine RNAiMAX (Thermo Fisher Scientific), as per the manufacturer's instructions. Typical siRNA transfection efficiency in both CHO cell lines was above 90% as assessed by fluorescently labeled negative siRNA (AllStars Neg. siRNA AF 647, Qiagen). siRNA-treated cells were used for further experiments two days after transfection. To induce GFP-specific DSBs, untreated or siRNA-treated CHO-K1 and CHO-DG44 HR reporter cells were transfected with 690 ng of single sgRNA expression plasmid (G1, G3, G7) and 2300 ng of wild-type Cas9 expression plasmid, or 900 ng of paired sgRNA expression plasmid (G1G5, G3G7) and 2500 ng of FokI-dCas9 expression plasmid, 1400 ng of GFP donor plasmid, 200 ng of pCMV-DsRed-Express expression plasmid and complemented with pUC19 to 5000 ng, if necessary. For double siRNA transfections, sgRNA, nuclease, donor and dsRed expression plasmids were co-transfected with the second siRNA load. For negative CRISPR control experiments, the sgRNA, nuclease and/or donor expression plasmids were substituted with pUC19 plasmid, or parental CHO-K1 cells were used instead of the CHO-K1 HR reporter cells. GFP

reconstitution and dsRed expression were analyzed two and three days after CRISPR transfection in CHO-DG44 and CHO-K1 cells, respectively. Only dsRed+ cells were analyzed to quantify HR efficiency in order to account for differences in transfection efficiency. Typical dsRed positive (dsRed+) cell frequencies were 40-60% in CHO-DG44 and 15-30% in CHO-K1 cells.

For protein overexpression experiments, 0.1 to 10 μg of protein expression vectors were co-transfected with the G3G7 sgRNA, FokI-dCas9 nuclease, GFP donor, and dsRed expression plasmids into CHO-K1 HR reporter cells as described above, and cells were analyzed for GFP and dsRed fluorescence three days after transfection. Samples were either compared to the corresponding control cells treated with the same amount of titrated empty vector, or alternatively, all samples were complemented with empty vector to reach the same total plasmid DNA amount and compared to one control sample. The obtained results were comparable using either method.

HR reporter sequence analysis

To analyze repaired GFP sequences, CHO-K1 reporter cells were bulk-sorted for dsRed+ GFP- and GFP+ cells at least four days after the transfection of the G3G7 sgRNA, FokI-dCas9 nuclease, GFP donor, and dsRed expression plasmids. 5,000 cells were sorted per condition, except for Rad51-depleted GFP+ cells, for which only 2000 GFP+ were sorted due to an overall low GFP+ frequency. Polyclonal populations were expanded for a minimum of two weeks, to dilute out episomal donor plasmids, prior to DNA extraction using the DNeasy Blood & Tissue Kit (Qiagen). 100 ng of isolated genomic DNA was used to PCR amplify the target loci using HotStart HiFidelity Polymerase (Qiagen), according to standard protocols, with primers located outside of the donor homology arms (Table S3.2). The expected amplicon sizes derived from the CHO-K1 HR assay were 2.6 kb and 2.7 kb for the unrepaired and repaired GFP sequences, respectively. To assay the HR reporter sequence by restriction digest, these PCR amplicons were column purified, digested with AflII for 1 h at 37 °C, and the digestion products were separated on an agarose gel. Unrepaired GFP sequences are cleaved by AflII yielding a 1.1 kb and a 1.5 kb band, while correctly repaired GFP sequences lack an AflII recognition site. To analyze the HR reporter sequence by Sanger sequencing, PCR products were either directly sequenced (polyclonal PCR sequencing) or cloned into plasmid vectors using the TOPO TA Cloning Kit for Sequencing (Invitrogen). Briefly, PCR products were gel pu-

rified with the Wizard SV Gel and PCR Clean-Up System (Promega), A-tailed using Taq Polymerase (Qiagen) followed by cloning into the TOPO vector. Single transformed E. coli colonies were screened for TOPO vectors with inserts by colony PCR yielding a total of 228 positive colonies, of which 116 colonies were analyzed by Sanger sequencing. In CHO-DG44 cells, GFP+ and GFP- cell clones were isolated after G3G7 sgRNA, FokI-dCas9 nuclease, GFP donor, and dsRed expression plasmid transfections using limiting dilution. The HR reporter sequence flanking the CRISPR target site were PCR amplified and amplicons of 21 clones were Sanger sequenced and analyzed. The expected amplicon sizes derived from the CHO-DG44 HR assay were 2.5 kb and 2.7 kb for unrepaired and repaired GFP sequences, respectively. For each condition, transfection and sequence analysis were performed once.

Gene correction, illegitimate integration and unrepaired GFP deletion quantifications

The frequency of illegitimate integration of the donor plasmid and unrepaired HR reporter sequences was quantified by qPCR. For this, 10 ng of DNA extracted from GFP+ and GFP- CHO-K1 polyclonal populations was used to amplify the GFP donor-derived insert sequence (GFP_Insert), the GFP donor-derived plasmid backbone (Donor_Backbone) or the deleted GFP sequence (GFP_Deletion) (Table S3.2). The GFP_Insert amplicon quantifies gene corrections and illegitimate integrations, namely plasmid knock-in and random integration, while the Donor_Backbone amplicon quantifies illegitimate integrations only. The GFP_Deletion amplicon quantifies uncorrected GFP deletions, while large CRISPR-derived deletions may also prevent amplification with the GFP_Deletion primers. Puromycin-specific primers were used to quantify PB HR reporter copies and B2M-specific primers served as normalization control. qPCR experiments were performed using Power SYBR Green PCR Master Mix for the QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems) and analyzed by relative quantification [364]. The relative gene correction frequency was calculated by subtracting the relative Donor_Backbone value from the relative GFP_Insert value in the same sample.

siRNA knockdown and overexpression quantifications

For real-time quantitative PCR (RT-qPCR), total RNA was extracted from CHO cells using the NucleoSpin RNA kit (Macherey Nagel). To assess siRNA-mediated knockdown efficiency, RNA was extracted two days after transfection, of which 1 μg was reverse transcribed into cDNA using a mix of oligo(dT)15 and random hexamer primers and the GoScript Reverse Transcription System (Promega). To estimate repair protein overexpression levels, RNA was extracted one day after transfection and 400 ng of RNA was reverse transcribed into cDNA using the QuantiTect Reverse Transcription Kit (Qiagen) containing an additional genomic DNA elimination step. The mRNA levels of siRNA targets, overexpressed cDNAs, FokI-dCas9 nuclease, and the B2M housekeeping gene were determined in triplicates using the SYBR Green I Master mix for the Roche LightCycler 480 instrument or the Power SYBR Green PCR Master Mix for the QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems) with primers listed in Table S3.2. For measuring mRNA overexpression levels, RT-minus controls were tested in parallel to confirm a difference of at least 5 CT between samples and RT-minus controls. Knockdown efficiency and overexpression levels were assessed by relative quantification as previously described [364], and they are expressed relative to each siNeg or empty vector control. siRNA knockdown efficiency in CHO-DG44 cells was previously validated [141].

Cell cycle analysis

To analyze CHO cell cycle progression after siRNA knockdown, 2 x 10^6 cells were harvested two days after transfection, fixed by adding dropwise 500 μl ice-cold 70% EtOH while vortexing and stained using 500 μl staining solution comprised of 0.5 μg/ml DAPI and 0.1% Triton X-100 in PBS. Stained cells were stored at 4 °C until flow cytometry analysis.

Flow cytometry and microscopy

To determine the frequency of GFP+ cells (to assess gene correction by HR) and dsRed+ cells (transfection control), CHO-K1 cell suspensions were centrifuged and resuspended in normal growth medium, while adherent CHO-DG44 cells were trypsinized and resuspended in PBS with 2% FBS. Single cell flow cytometry was performed on CyAN

and Gallios instruments (Beckman Coulter) and analyzed using FlowJo software v10.4.2. Cells were first gated using side scatter (SSC) versus forward scatter (FSC) to separate the intact cell population from debris, and then gated for single cells using SSC height versus SSC width or area and using FSC height vs FSC width or area. This cell population was then gated for dsRed+ cells, and the transfected dsRed+ cells were further gated for GFP+ cells using the appropriate fluorescent channels with dsRed+ or GFP+ only cells as gating control. At least 50,000 transfected dsRed+ single cells were analyzed per sample. For bulk cell sorting, CHO-K1 reporter cells were sorted into 96well plates with prewarmed medium using the MoFlo Astrios EQ cell sorter (Beckman Coulter), briefly centrifuged to exchange medium and further cultured for one month. For cell cycle analysis, a minimum of 40,000 single cells was acquired at a low acquisition rate yielding typical coefficient of variation mean values below 6%. Cell cycle distribution was analyzed by fitting the histogram of DNA content, as measured by DAPI intercalation, to the Watson Pragmatic Model using FlowJo software. Cells with cell contents below the G1/G0 phase and above the G2/M phase were grouped as apoptotic and polyploid cells, respectively. Fluorescence microscopy analysis was carried out on an Axio Observer.A1 microscope (Zeiss, Germany) and images were analyzed using ImageJ software.

Statistical analysis

R (https://www.r-project.org) and Microsoft Excel were used for statistical analysis of the data. Graphs show mean value, error bars represent standard error of the mean (s.e.m) and the number of biological replicates is stated per graph. Statistical significance was calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction (* $P < 0.05$, ** $P < 0.01$). To test for a nonlinear, inverse u-shaped relationship between the amount of the overexpression protein and the relative HR frequency (GFP+), a non-linear regression model was used with the following specification:

$$relativeHR_i = \beta_0 + \beta_1 * \mu gProtein_i + \beta_2 * \mu gProtein_i^2 + \varepsilon_i,$$

where $i = 1,\dots,N$ is the i-th measurement, $relativeHR_i$ indicates the relative HR frequency relative to the mean of the empty vector control, $\mu gProtein_i$ is the amount in μg of the corresponding protein applied, and ε_i is a random error with mean zero compris-

ing all factors that influence the relative HR frequency but are not taken into account by the model. The unknown parameters β_0, β_1 and β_2 govern the relationship between *relativeHR_i* and $\mu gProtein_i$, and need to be estimated with the method of Ordinary Least Squares (OLS). OLS estimates the best fitting parameters by minimizing the sum of the squared residuals, that is the part of the variation in *relativeHR_i* which the model cannot explain. The relationship between the relative transfection efficiency (dsRed+) and the amount of protein was calculated analogously using the non-linear model

$$relativeTransfectionEfficiency_i = \beta_0 + \beta_1 * \mu gProtein_i + \beta_2 * \mu gProtein_i^2 + \varepsilon_i$$

For the cluster analysis, the HR repair dissimilarity between different sgRNA target sites and nuclease types was calculated using the sum of the absolute differences method (Manhattan distance).

3.4 Results

3.4.1 Construction of a CHO HR reporter compatible with the FokI-dCas9 nuclease system

The direct repeat GFP (DR-GFP) assay has been the gold standard to understand intra-chromosomal HR repair mechanisms in mammalian cells [312]. However, this HR assay does not rely on an exogenous donor template, as it was not designed to study precise gene insertions and gene corrections. Alternatively, several gene targeting assays were established to quantify HR-mediated in-frame insertions of fluorescent marker genes in human cells [314, 365]. Here, we wished to establish a GFP-based HR assay to specifically measure targeted gene corrections in CHO cells. In addition, this assay was designed to allow comparison of the performance of both the highly sequence-specific FokI-dCas9 and the efficient wild-type Cas9 CRISPR nuclease systems.

The established HR assay is based on a chromosomally integrated GFP loss-of-function mutation and a non-functional GFP donor plasmid. HR-mediated gene conversion can lead to the restoration of a functional GFP coding sequence which is traceable by GFP fluorescence (Figs. 3.1A, S3.1A and S3.1B). The genomic non-functional GFP sequence consists of two in-frame stop codons followed by a 101 bp frame-shifting deletion. This

truncated GFP sequence is adjacent to previously described GFP-specific sgRNA recognition sites (G1, G3, G5, G7) [275]. These sgRNAs can be used either individually together with the wild-type Cas9 nuclease, or as sgRNA pairs (G1G5 or G3G7) together with the highly specific FokI-dCas9 nuclease, to efficiently cleave the GFP target sequence in CHO cells (Figs. S3.1C and S3.1D). It also harbors an AflII recognition site overlapping the second stop codon, to screen GFP coding sequences by restriction digest. It is further flanked by a puromycin resistance selection gene and by two inverted terminal repeats for PiggyBac (PB) transposase-mediated genomic integration, and it is hereafter referred to as the HR reporter. The homologous donor template is a non-functional translational trap plasmid that contains a GFP coding sequence bearing extensive homology to the truncated GFP target site (382 bp left and 1252 bp right homology arms), but that lacks a promoter and a translational initiation codon (Fig. 3.1A). GFP expression from episomal or genome-integrated donor plasmids is further hampered by two in-frame stop codons upstream of the GFP coding sequence. To avoid CRISPR cleavage of the repaired product, the donor DNA contains additional silent mutations at the sgRNA PAM sites (Fig. S3.1B). Targeted HR-mediated GFP gene correction should thus remove the AflII restriction site and reconstitute GFP fluorescence.

To generate a chromosomal HR reporter cell line, suspension-adapted CHO-K1 cells were transfected with the HR reporter construct and a PB transposase expression vector. Cells having stably integrated the HR reporter in their genome were selected for puromycin resistance. PB transposon vectors preferably integrate as single copies in separate loci and at a low copy number per genome [362]. After isolating single cells, we quantified the number of stably integrated HR reporter constructs relative to a two-copy CHO cell gene using quantitative PCR [362]. A clone containing four integrated HR reporter copies was selected, allowing to mimic genome editing of multi-locus sites (Figs. 3.1B and S3.2A). Reconstitution of a functional GFP coding sequence in the HR reporter cell line required the co-transfection of the G3G7 sgRNA, FokI-dCas9 nuclease and GFP donor expression plasmids. This yielded approximately 0.25% of GFP positive (GFP+) cells among the entire cell populations, and 0.74% of GFP+ cells among the transfected dsRed+ cell population three days after transfection (Figs. 3.1C and S3.3), whereas dsRed+ cell frequencies ranged around 15-30%. A HR frequency under 1% is within the expected range for CHO cells, which have inherently low HR activity [171, 268, 291]. GFP+ cells were below 0.01% when transfecting the parental CHO-K1 cells devoid of the reporter construct, implying that the majority of GFP fluorescence arises from repair
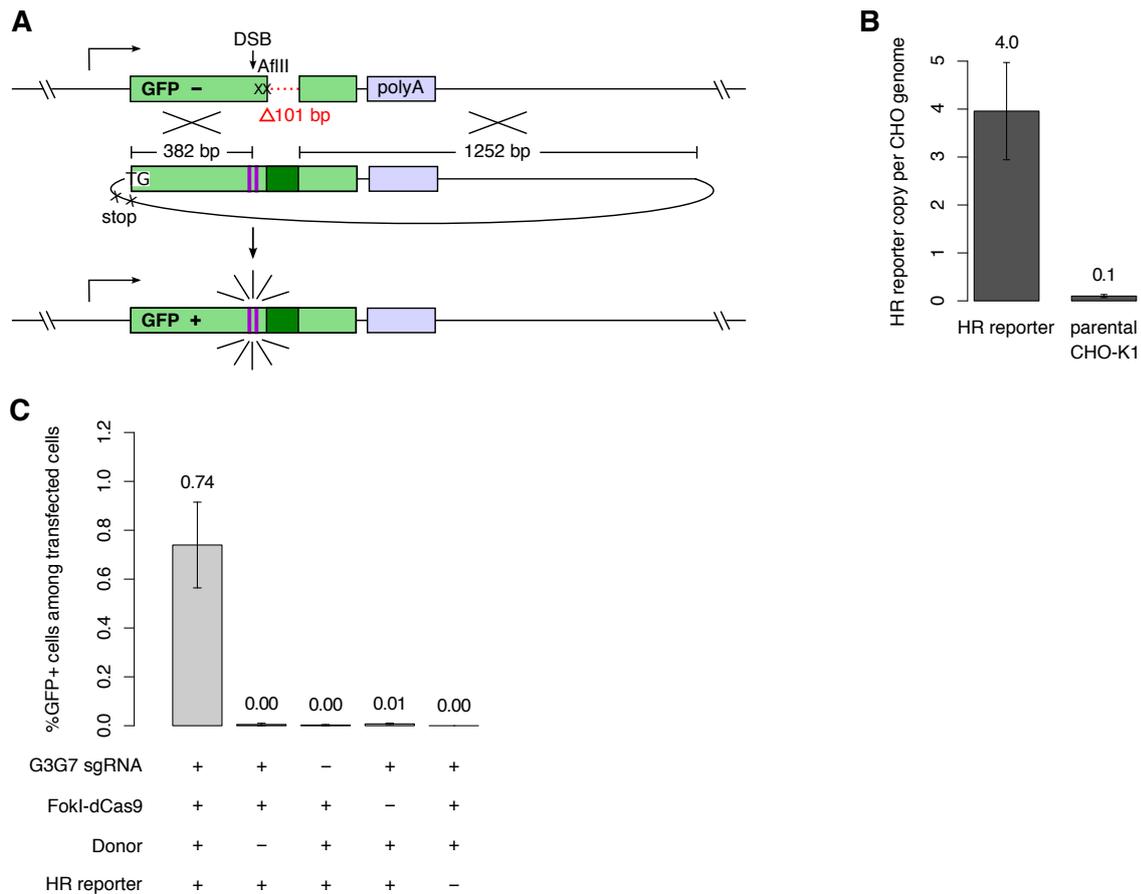
Figure 3.1: Construction of a CHO-K1 GFP gene correction assay. (A) Schematic overview of the newly established HR assay. The chromosomal HR reporter construct, stably integrated into the CHO-K1 cell genome using PiggyBac (PB) transposition, contains two-in frame stop codons (X signs) followed by a 101 bp GFP frame-shifting deletion (red dotted line) downstream of several CRISPR target sites. An AflII restriction site overlaps with the second stop codon in the truncated GFP sequence. The homologous donor template is a circular plasmid containing a non-functional GFP coding sequence that provides the missing GFP sequence (dark green) but lacks a promoter and translation initiation codon, and that shares homology arms to the deleted chromosomal GFP sequence on either side (horizontal bracket lines with homologous sequence length indicated). Two silent PAM mutations (purple bars) as well as two upstream stop codons (X signs) are present in the donor plasmid vector. Following CRISPR-mediated DSB induction, HR-mediated GFP correction results in a functional GFP coding sequence and in cellular GFP fluorescence. (B) qPCR assay of the number of HR reporter copies per genome chromosomally integrated in the selected CHO-K1 reporter cell line using primers specific to the puromycin selection gene located on the PB transposon, as shown in Figure S3.2A (n = 3, error bars represent s.e.m). (C) Flow cytometry quantification of the percentage of GFP expressing cells among the transfected dsRed+ cell population of either HR reporter construct-containing cells (HR reporter) or parental CHO-K1 cells transfected with different combinations of the GFP-specific G3G7 sgRNA, the FokI-dCas9 nuclease, and the GFP donor expression plasmids, as indicated (n=2, error bars represent s.e.m).

50

of the chromosomal HR reporter sequence rather than episomal repair or donor plasmid integration.

3.4.2   GFP reconstitution relies on the HR repair pathway

To confirm that GFP fluorescence reconstitution quantifies canonical HR repair, we next evaluated the GFP expression stability. HR repair is expected to result in permanent GFP correction of the chromosomal HR reporter and hence it should lead to sustained GFP expression. In agreement with HR-mediated GFP repair, unsorted and bulk-sorted GFP+ cell populations retained stable GFP expression levels (Fig. 3.2A and data not shown). Among GFP+ sorted cells, less than 0.03% co-expressed dsRed and GFP, implying that co-integration of the dsRed transfection control plasmid and GFP sequence restoration are infrequent events (data not shown). We further assessed that GFP+ cells were devoid of the AflII recognition site in the HR reporter sequence. The HR reporter cassette was PCR amplified using primers located outside of the homologous donor arms and digested with AflII (Fig. S3.2A). Whereas the amplicons from untreated and GFP- cells were cleaved by AflII, they were not when assessing GFP+ cells. This implies that the anticipated chromosomal GFP sequence correction causes GFP fluorescence and that most or all HR reporter alleles were corrected (Fig. S3.2B).

To assess whether HR pathway components are required to generate GFP+ cells, we transiently knocked down various DSB repair factors in the CHO-K1 HR assay. For siRNA knockdown assays, each target gene was silenced with a pool of three siRNAs, and proper knockdown of the mRNA level was validated experimentally by RT-qPCR (Fig. S3.4). To assess siRNA specificity, siRNA pools were deconvoluted to substantiate the knockdown efficiency, consistency and reproducibility of the individual siRNAs (Fig. S3.5). Transfecting the HR assay with siRNAs targeting Rad51 – a key recombinase for HR repair – decreased the occurrence of GFP+ cells to the background levels observed from non-targeting negative control siRNA (siNeg), whereas knockdown of other repair pathway components (Ku80, Lig3) had little effect (Fig. 3.2A). In contrast, knockdown of Pari, which opposes Rad51 nucleation and HR [71], rather increased the occurrence of GFP+ cells. This data further corroborated an HR-dependent reconstitution of the GFP coding sequence in this assay.
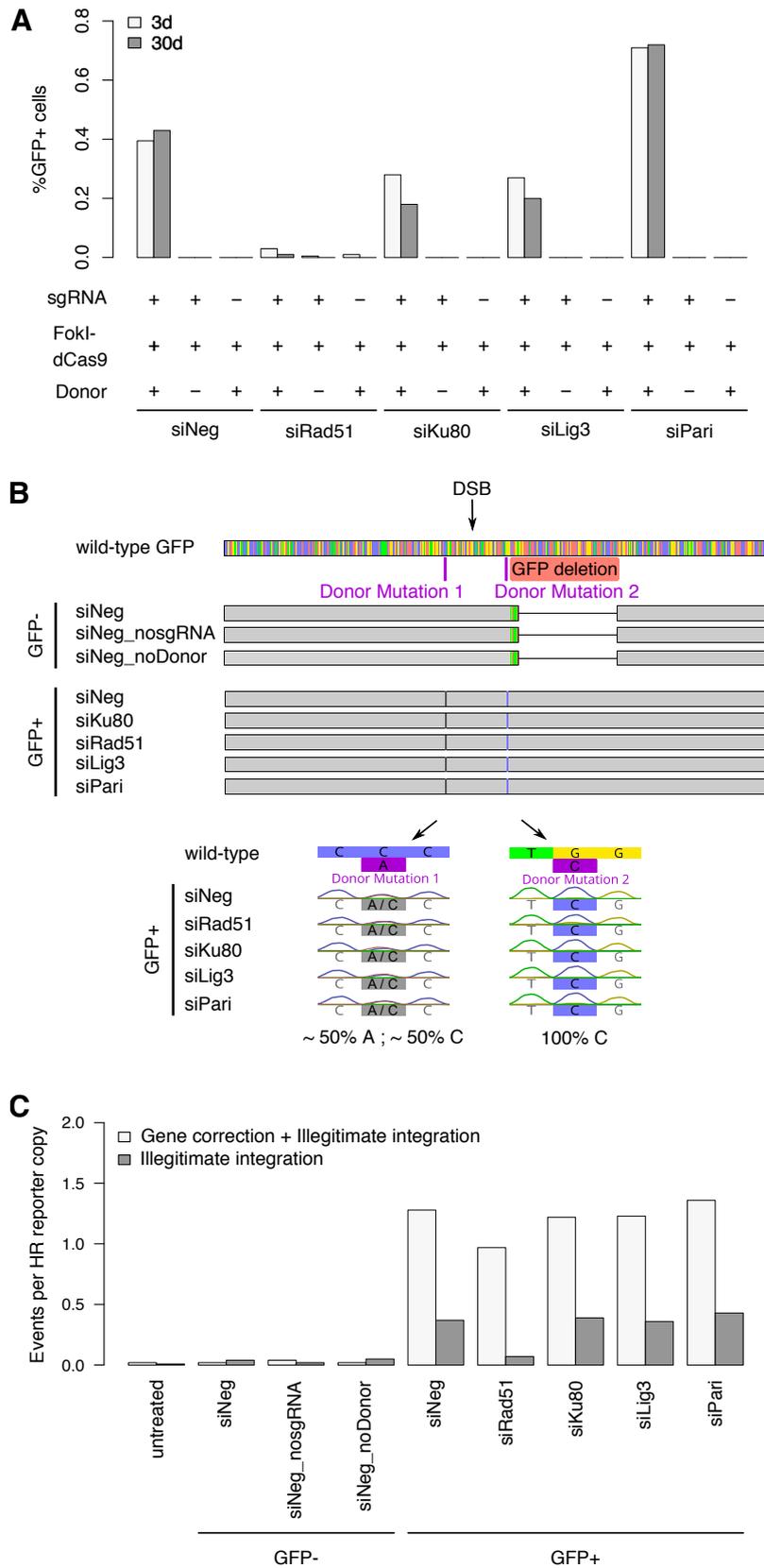
Figure 3.2: Characterization of GFP reconstitution events in the CHO-K1 HR reporter cells. CHO-K1 cells containing chromosomal HR reporter constructs were treated with a selection of validated siRNA pools to knockdown the expression of specific DNA repair activities or with a pool of negative control siRNAs (siNeg). Caption continued next page.

Figure 3.2: This was followed by the transfection of a second load of siRNA pools together with G3G7 sgRNA, FokI-dCas9 nuclease and/or donor expression plasmids, as indicated. (A) Quantification of cellular GFP fluorescence was performed by flow cytometry 3 d and 30 d after cell transfection and is expressed as percentage of the entire cell population (n=1, see Figure 3.3A for extended data). (B) Polyclonal populations of siRNA-treated cells expressing (GFP+) or not (GFP-) a functional GFP sequence following the HR assay were sorted by flow cytometry. The chromosomal HR reporter sequences were amplified using primers located outside of the donor plasmid homology arms (Fig. S3.2A) and processed by Sanger sequencing. Mutations present in the donor GFP sequence, depicted by purple lines, consist of a C to A substitution (Donor Mutation 1) and a G to C substitution (Donor Mutation 2). The schemes depict the most frequently detected outcome of the assays as deduced from DNA sequence analysis, where a grey line indicates a mix of the donor-derived mutation and wild-type sequence, and a blue line indicates the most frequent occurrence of the donor-derived G to C substitution. The sequencing electropherograms corresponding to the Donor Mutation 1 and 2 loci of GFP+ cells are shown for each experimental condition. (C) Primers specific to the GFP donor sequence (GFP_Insert) and donor plasmid backbone (Donor_Backbone), as depicted in Figure S3.2A, were used to quantify the co-occurrence of correctly repaired GFP sequences and illegitimate integration events (open bars), or of illegitimate integration events solely (closed bars), respectively. The number of events was normalized to the copy number of the chromosomal HR reporter constructs.

To further substantiate the occurrence of HR-mediated repair, we sequenced PCR amplicons of the HR reporter sequences from GFP- and GFP+ cell populations. PCR products from polyclonal GFP- cells maintained the original GFP deletion but GFP+ cells displayed the corrected GFP sequence and donor-derived PAM mutations (Fig. 3.2B). Intriguingly, all sequencing reads obtained from GFP+ cells contained the deletion-proximal downstream donor-derived mutation C (Donor mutation 2), whereas, at the position of the distal upstream Donor mutation 1, we observed a mix between the donor-derived A and the chromosomal GFP-derived C allele. In addition to polyclonal PCR product sequencing, we sequenced 116 HR reporter sequences cloned into plasmid vectors. Among 24 GFP- colonies, all possessed the original chromosomal GFP deletion, whereas five had additional indel mutations around the CRISPR DNA cleavage site (Fig. S3.6). As expected, we did not observe such CRISPR-derived indel mutations in the absence of a functional CRISPR system (siNeg_nosgRNA), but they occurred in the absence of a GFP donor plasmid (siNeg_noDonor). The majority of these additional mutations showed MMEJ- or SD-MMEJ-specific repair signatures, substantiating high alt-EJ activities in CHO cells (Table S3.4). Among the 88 DNA sequences derived from GFP+ cells, 83 no longer displayed the GFP deletion and were converted to the donor sequence, among which 75% and 100% contained the upstream and downstream donor-specific mutations, respectively,

consistent with the percentages observed from polyclonal PCR sequencing.

This particular mutation pattern observed among the GFP+ repair junctions may be explained by DNA mismatch repair (MMR) and protruding DNA sequences occurring during HR. Indeed, following extensive 5'-3' end resection, HR repair requires that at least one of the resected 3' ssDNA chromosomal ends invades a homologous donor template. In the HR assay described here, the GFP donor plasmid contains two silent PAM mutations. Thus, invasion of the upstream 3' ssDNA end into the donor plasmid, for instance, creates a one base pair mismatch at the PAM site upstream of the DSB site (Fig. S3.7, boxed green cone and purple bar). The mismatched bases activate MMR, which can lead to the incorporation of the upstream donor-derived PAM mutation in the repaired product (Fig. S3.7, grey bar). The observation that more than 50% of the repaired GFP sequences obtained from GFP+ cells contained the upstream donor mutation indicates a preference of the MMR to correct the invading DNA strand, as noted previously [190]. Following MMR, GFP gene conversion is then mediated by the SDSA or DSBR pathway. Notably, during both the SDSA or DSBR repair pathways, a protruding non-homologous 3' flap arises at the downstream 3' ssDNA end, as a consequence of the distance between the CRISPR cleavage site and the GFP deletion (Fig. S3.7, yellow box). To complete HR repair, this protruding flap must be removed. However, since this flap derives from the chromosomal GFP sequence, and thus lacks a PAM mutation, the downstream donor-derived PAM mutation should become fixed in all GFP repair products (Fig. S3.7, purple bar), as observed here.

Among the GFP+ colonies, five out of 88 (6%) showed no evidence of gene correction. Thus, we further quantified the amount of unrepaired GFP alleles by qPCR using primers surrounding the deletion site (Fig. S3.2A). Among GFP- cells, around 80-90% of the chromosomal GFP alleles still contained the original deletion, while on average only 8% of the GFP+ cells did (Fig. S3.2C). The finding that 6-8% of the GFP alleles remain unrepaired in the polyclonal GFP+ cell population indicated that most cells must have successfully corrected all their GFP alleles, possibly due to co-conversion.

When screening a total of 228 bacterial colonies for donor plasmid-containing HR reporter sequences, we noted that two GFP- and two GFP+ cell-derived sequences (2%) contained visibly shorter inserts than expected. Indeed, these inserts harbored large deletions ranging from 333 to 678 bp, likely resulting from alt-EJ-mediated repair (Table S3.4). While these large deletions were located close to the predicted DSB site observed

from GFP- cells, they occurred more than 50 bp up- or downstream of the DSB site and even outside of the GFP coding sequence in GFP+ colonies. These large mutations indicate that resection length and homology search may extend over several hundred base pairs and that HR-mediated GFP reconstitution may co-occur with other repair mechanisms at the same DSB end in CHO cells.

Given that GFP reconstitution may be accompanied by illegitimate donor plasmid integrations, we sought to quantify their occurrence. Illegitimate integration events include donor plasmid knock-ins at chromosomal GFP alleles (Fig. S3.7), as well as random integrations. Such integration events are considered as undesired by-products of gene correction, as they introduce unwanted sequences, often from bacterial origin, leading to possible transgene silencing [217]. We first quantified the total number of gene correction and illegitimate integration events using primers specific to the GFP sequences provided by the donor plasmid and required for reconstitution of a functional GFP coding sequence by HR (Fig. S3.2A). Donor-derived GFP inserts were detected at background levels in polyclonal GFP- cell populations, as expected for unrepaired GFP sequences, but they were present at levels slightly above the estimated HR reporter copy number in polyclonal GFP+ cell populations (Fig. 3.2C, open bars). Next, we quantified the amount of illegitimate recombination events solely, using donor backbone-specific primers (Figs. 3.2C, closed bars, and S3.2A). For instance, in a polyclonal cell population treated with negative control siRNA, we observed 1.28 copies of the donor-derived GFP insertion per chromosomal HR reporter copy, among which 0.37 correspond to illegitimate plasmid integrations (Fig. 3.2C, siNeg column, open vs. closed bars, respectively). Similar frequencies were also observed in samples depleted in Ku80, Lig3 and Pari. However, we measured only 0.1 backbone integrations per HR reporter copy in the Rad51 knockdown samples, in line with the previously reported Rad51-dependent mechanism for illegitimate plasmid integration in CHO cells [141]. We infer that the difference between the total number of donor-derived insertions and illegitimate integrations corresponds to GFP gene correction events only. Hence, among the tested siRNA conditions, at least 70% of GFP inserts are true gene corrections events, while the remaining inserts result from knock-ins or random integrations. This identifies gene correction as the predominant recombination mechanism in various polyclonal GFP+ cell populations. Taken together, we concluded that this novel chromosomal HR assay primarily detects HR-mediated GFP gene correction.

3.4.3 Knockdown of alt-EJ factors reduces HR frequency

We hypothesized that efficient HR repair in CHO cells may be hindered by its competition with highly active DSB repair pathways, especially alt-EJ pathways. In line with this hypothesis, alt-EJ factors were found to contribute to plasmid end-joining and random genomic integration [119, 141]. Moreover, mRNA levels of alt-EJ factors are elevated in CHO cells compared to those of other DSB repair factors, particularly HR factors (Fig. S3.8A). To systematically identify possibly rate-limiting or competing activities for HR in CHO-K1 cells, we transiently knocked down DNA repair factors involved in the various DSB repair pathways using validated siRNAs (Figs. S3.4 and S3.5). Targeted DSB repair was subsequently induced by transfecting the GFP-specific G3G7 sgRNA, FokI-dCas9 nuclease, GFP donor and dsRed expression plasmids.

In line with previous assays, the knockdown of the Rad51 recombinase yielded a decrease in the GFP reconstitution frequency of more than 50% compared to the non-targeting negative control siRNA (Fig. 3.3A). Depleting the MRN-CtIP complex components Nbs1 and CtIP resulted in a moderately lower HR frequency, consistent with the role of the MRN-CtIP in DSB recognition and initial end resection [8]. Unexpectedly, knockdown of the Mre11 nuclease, another component of the MRN-CtIP complex, significantly increased HR frequency. This implied that MRN-CtIP constituents might be required at a proper ratio for optimal DSB end processing during HR, and that the knockdown of Mre11 may restore a stoichiometric balance among MRN components, leading to improved HR frequency.

As the knockdown of Rad51 did not fully prevent GFP reconstitution, although being essential for HR, we wondered whether the Rad51 siRNA effect on GFP gene correction could be enhanced. Indeed, when a second load of Rad51 siRNAs was transfected together with the G3G7 sgRNA, FokI-dCas9 nuclease, GFP donor and dsRed expression plasmids, GFP reconstitution was nearly abolished (Fig. 3.3B). Conversely, HR was increased 2-fold after two Mre11 siRNA transfections, while it was increased 1.4-fold after a single siRNA transfection. Therefore, double siRNA transfection offers a simple approach to extend siRNA-mediated effects and to further elevate HR efficiency.
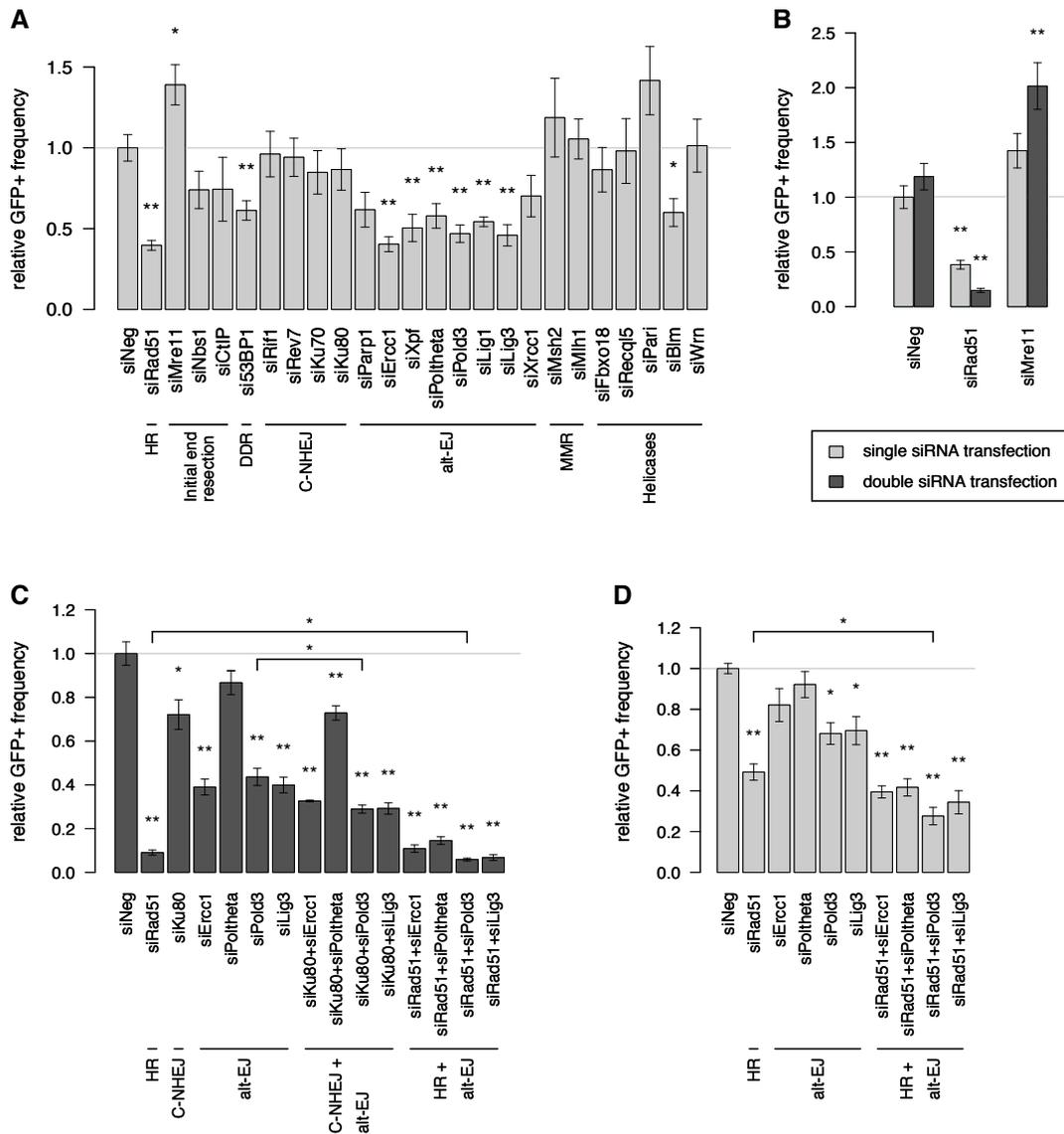
Figure 3.3: Effect of repair protein knockdowns on GFP reconstitution. (A) CHO-K1 HR reporter cells treated with validated siRNA pools targeting various DNA repair proteins were re-transfected with the G3G7 sgRNA, FokI-dCas9 nuclease, donor and dsRed expression plasmids. Cells were then analyzed by flow cytometry for GFP fluorescence reconstitution and the frequency of GFP+ cells among transfected dsRed+ cells was normalized to the frequency obtained with cells treated with negative non-targeting control siRNA pool (siNeg). (B) Single siRNA transfections were performed as described in panel A. For double siRNA transfections, the cells were transfected again with the same siRNA pools together with the G3G7 sgRNA, FokI-dCas9 nuclease, donor and dsRed expression plasmids. (C and D) The effect of combined knockdown of HR (Rad51), C-NHEJ (Ku80) and several alt-EJ (Ercc1, Pol theta, Pold3, Lig3) factors was tested either using a double (C) or a single (D) siRNA transfection using the most efficient singular siRNA from each siRNA pool. Single protein knockdowns were complemented with negative siRNA to maintain a constant total siRNA amount. The frequency of GFP+ cells among the transfected dsRed+ cells are shown relative to the non-specific siNeg negative control ($n \geq 2$, error bars represent s.e.m.). For Figure 3.3C and 3.3D, brackets indicate one-to-one comparison performed in addition to the statistics done against siNeg. P values (* $P < 0.05$, ** $P < 0.01$) were calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction.

Knockdown of C-NHEJ factors did not significantly affect HR frequency. This suggests that C-NHEJ does not substantially compete with HR, in agreement with its overall low activity in CHO cells [119, 141]. However, depletion of the DNA damage response (DDR) protein 53BP1 diminished HR frequency. This could be explained by 53BP1's ability to protect DSB ends from hyper-resection, which is considered to oppose productive HR repair [366].

Interestingly, HR reporter cells deficient in alt-EJ factors showed on average 40% reduced HR frequency levels. This drop was most significant for the Ercc1, Xpf, Pold3, Lig1, and Lig3 knockdowns. Pol theta-depleted cells showed also reduced HR frequencies, despite a relatively low and variable siRNA knockdown efficiency (Fig. S3.4). These data suggest that alt-EJ factors, although generally being considered as HR competitors, are limiting for – and positively contribute to – efficient HR-mediated gene correction in CHO cells.

To ascertain that siRNA effects do not result from clone-specific biases, we knocked down Rad51 and Lig3 in three independent CHO-K1 clones bearing the chromosomal HR reporter construct at similar copy number. We noticed that deficiency in Rad51 and Lig3 yielded consistent results between the different HR reporter clones, indicating that the observed siRNA effects are likely not linked to the particular HR reporter cell line used (Fig. S3.8B).

Proteins that oppose HR, such as MMR factors and helicases, were also tested for their effect on HR activity [124]. Knockdown of the MMR factors Msh2 or Mlh1 as well as the helicases Fbxo18, Recql5 or Wrn exhibited no detectable effect on HR activity (Fig. 3.3A). However, depletion of the helicase Pari moderately enhanced HR by 1.4-fold, as seen before (Fig. 3.2A). Depletion of the Blm exonuclease and helicase reduced HR frequency, which may result from diminished end resection and/or from fewer DSBR dissolution events (Fig. S3.7).

Given that single C-NHEJ and alt-EJ knockdowns did not elevate HR, we hypothesized that multiple concomitant and possibly redundant repair pathways might prevent efficient HR-mediated gene correction. To test this hypothesis, we simultaneously knocked down Ku80, a representative C-NHEJ protein, and various alt-EJ factors in the CHO-K1 HR assay using a double siRNA transfection approach, to inhibit these end-joining pathways altogether. Surprisingly, we first noticed that a double siRNA transfection of Ku80, unlike a single siRNA transfection, led to a weak reduction in the frequency of GFP+

cells, suggesting that early C-NHEJ factors may play a role in HR in CHO cells (Fig. 3.3C). Second, GFP reconstitution did not increase for all combined C-NHEJ and alt-EJ knockdown conditions. HR repair was even significantly decreased upon combined Ku80 and Pold3 knockdowns compared to Pold3 knockdown alone (Fig. 3.3C). Similarly, HR frequency did not increase when Ku70 was knocked down in combination with alt-EJ factors (data not shown). Overall, we concluded that the knockdown of C-NHEJ or alt-EJ factors, either alone or in combination, did not increase HR frequency, suggesting the occurrence of other competing repair pathways and/or insufficient overall HR activity.

3.4.4 Overexpression of rate-limiting alt-EJ factors increases HR efficiency

Having shown that alt-EJ components may positively contribute to HR, we aimed to strengthen this finding by overexpressing some of the most HR rate-limiting factors, in order to reverse the effect elicited by their knockdown. Rad51, Ercc1, Pold3 and Lig3 were selected to be overexpressed, as their knockdown lowered HR frequency the most (Fig. 3.3A). Although previous studies have already attempted to improve HR by Rad51 overexpression, their results were contrasting or inconclusive, possibly due to variable Rad51 overexpression levels [310, 316].

To identify the optimal expression level to leverage HR, we transfected the HR reporter cells with increasing amounts of Rad51, Ercc1, Pold3 and Lig3 expression plasmids. Under these conditions, mRNA levels were increased between 1.3- to 299-fold on average for the lowest and highest plasmid amounts, respectively, as assessed by RT-qPCR (Fig. S3.9). Overexpression conditions tended to affect HR in a dose-dependent manner: lower expression plasmid amounts moderately stimulated HR (with an increase of up to 1.5-fold for Rad51), while higher expression plasmid amounts decreased HR (Fig. S3.10A). Hence, the HR frequency upon overexpression of HR rate-limiting factors seemed to follow an inverse u-shaped curve. Indeed, a non-linear regression model pooling together all HR rate-limiting factors provided statistical support for such a non-linear, inverse u-shaped relationship between the amount of protein and the relative HR frequency (Table S3.5). Further analysis revealed that this inverse u-shaped relationship is not caused by a change in the relative frequency of dsRed+ transfected cells (Fig. S3.10B, Table S3.6).

To exclude that this effect may be non-specific, we used the neomycin resistance protein as a control, and indeed, its overexpression did not elicit such a dose-dependent increase of HR frequency (Figs. S3.10C and S3.10D). Nevertheless, we cannot rule out a toxic effect of the highest levels of overexpression of some of these proteins, as we noted a decrease of dsRed+ cells upon Rad51 and Ercc1 overexpression. Interestingly, Ku70 and Ku80 overexpression appeared to leverage HR repair at specific doses without affecting the frequency of dsRed+ cells, implying that unbalanced levels of the two subunits of the Ku70-Ku80 heterodimer might favor HR rather than C-NHEJ. Cells overexpressing the dominant-negative Rad51 mutants Rad51-K133R and Rad51-K133A, which are deficient in ATP hydrolysis and ATP binding, respectively [367], showed reduced HR frequencies even at the lowest plasmid amount, providing additional evidence that GFP reconstitution requires Rad51-mediated HR repair. Altogether, moderate overexpression of specific alt-EJ factors, but not control proteins, tended to stimulate HR in a dose-dependent manner, in line with alt-EJ factors being one of the bottlenecks for efficient HR-mediated precise genome editing in CHO cells.

3.4.5   Alt-EJ contribution to HR is cell line independent

We next assessed whether the observed HR dependency on alt-EJ activities could be reproduced using another CHO cell line. For this purpose, we established a HR assay based on the adherent CHO-DG44 cell line, a CHO cell line that diverged from the suspension-adapted CHO-K1 cell line more than 40 years ago, and thus differs genetically [368]. The CHO-DG44 HR assay consists of a non-fluorescent GFP sequence lacking a 215 bp fragment of the GFP C-terminus coding sequence and the polyA signal and is present at low copy number (Figs. S3.11A and S3.11B). The GFP deletion is located 270 bp downstream of the CRISPR target sites and it shares around 400-500 bp homology on either side with the donor template. Similarly to the CHO-K1 HR reporter cell line, GFP reconstitution in CHO-DG44 cells required co-transfection of G3G7 sgRNA, the FokI-dCas9 nuclease, and GFP donor expression plasmids. This yielded 0.28% and 0.40% of GFP+ cells among all and transfected dsRed+ cells, respectively (Fig. S3.11C).

For HR reporter sequence analysis, we isolated GFP+ and GFP- CHO-DG44 HR reporter cell clones following treatment with different siRNAs, to avoid a potential bias generated by PCR amplification of polyclonal cell populations. Subsequently, the GFP sequences of single clones were PCR amplified and sequenced, as performed before. Out of

32 clones (6x GFP+, 2x GFP- per siRNA condition), 21 yielded single amplicons whose sequence could be determined, although with variable amplicon sizes (Fig. S3.11D). This indicated that the clones with non-amplifiable HR reporter sequences may contain plasmid knock-ins or other CRISPR-mediated large chromosomal rearrangements. As expected, among eight GFP- clones, none yielded a corrected chromosomal GFP sequence. However, two clones showed mutations at the CRISPR target site compatible with an alt-EJ DNA repair mechanism (Fig. S3.11E, Table S3.7). Among the fourteen GFP+ clones, eight clones displayed only the longer corrected GFP alleles containing the donor-derived missing GFP sequence, while six clones still contained the shorter GFP deletion sequence (Fig. S3.11F). Since illegitimate donor integrations do not lead to GFP expression at a detectable frequency (Fig. S3.11C), we infer that the shorter amplicons are preferentially PCR amplified when only a fraction of the alleles are corrected. Similar to the CHO-K1 repair junctions, 75% and 100% of the correctly repaired clones displayed the upstream and downstream PAM donor mutations, respectively. Furthermore, we observed in three clones that a 800 bp deletion occurred between two direct repeats downstream of the GFP coding sequence, independently of the repair status of the 270 bp GFP deletion, suggesting a single strand annealing (SSA)-like repair mechanism (Fig. S3.11F, Table S3.7). Overall, we concluded that this CHO-DG44 assay measures mainly HR, but that large chromosomal rearrangements occur more frequently than in the CHO-K1 reporter cells.

Knocking down DNA repair factors in CHO-DG44 HR reporter cells had similar effects on GFP reconstitution as in CHO-K1 cells. As before, the strongest inhibitory effect on HR frequency occurred upon Rad51 depletion (Figs. S3.11G). Unlike in CHO-K1 cells, knockdown of the MRN-CtIP proteins Mre11, Rad50, and CtIP, but not Nbs1, reduced HR frequency (Figs. 3.3A and S3.11G). Thus, there were some different responses to knockdowns when comparing the two cell lines, which may relate to possible differences in the levels or activities of their DNA repair factors. However, as seen in CHO-K1 HR reporter cells, knockdown of C-NHEJ factors had little effect, while knockdown of 53BP1 as well as most alt-EJ factors prevented GFP reconstitution. Interestingly, Pol theta-depleted CHO-DG44 cells showed HR levels similar to the control cells despite an efficient protein knockdown in this cell line [141]. Altogether, we concluded that the interconnection between HR and alt-EJ factors is mostly conserved and not limited to a given reporter cell line or to a particular truncated GFP HR reporter construct. As the interplay between the repair pathways cannot be explained by a simple competition for DSB structures, we conclude that Alt-EJ components may rather promote efficient HR.

3.4.6 Alt-EJ factors do not modulate cell cycle distribution and act distinctly from the HR pathway

One possible mechanism by which alt-EJ factors could influence HR may be by altering cell cycle progression. For instance, their knockdown might arrest cells in G1 phase, when HR is not favored. Based on DNA content analysis, the percentage of control CHO-K1 cells in the G1/G0, S, and G2/M phase was on average 32%, 40%, and 17%, respectively, and only few cells were apoptotic or polyploid (Fig. S3.12 and data not shown). Rad51 knockdown cells showed an increased proportion of G2 phase as well as of apoptotic cells, as expected for a repair factor involved in the G2/M checkpoint [369]. While a moderate increase of G2 phase cells was also detected following Mre11 knockdown, especially following Mre11 knockdown using the most efficient siRNA (siMre11_1), the cell cycle distribution was not altered in alt-EJ knockdowns. Hence modulation of the cell cycle cannot explain the observed contribution of alt-EJ factors to HR repair.

The inhibition of GFP reconstitution by alt-EJ and HR factor knockdowns might result from partially overlapping or fully distinct alt-EJ and HR repair pathways. While partially overlapping pathways typically have epistatic interactions, distinct repair pathways often show additive effects upon combined knockdowns. Combining Rad51 with Pold3 knockdown further diminished HR activity as compared to the knockdown of Rad51 or Pold3 alone, using both single and double siRNA transfections (Figs. 3.3C and 3.3D). Combined knockdown of Rad51 with other alt-EJ factors such as Ercc1, Pol theta or Lig3 followed the same trend, notably for single siRNA transfections, although the reduction in HR activity was not statistically significant when compared to Rad51 knockdown alone (Figs. 3.3C and 3.3D). Thus, the presence of additive effects for siRad51+siPold3 knockdowns as well as the absence of identifiable epistasis in the other tested conditions suggest that alt-EJ and HR factors have at least partially independent functions in HR repair.

3.4.7 Alt-EJ factor contribution to HR depends on the nuclease type and DSB-to-mutation distance

Previous studies have shown that the choice of sgRNA can significantly impact the repair pathway activities and outcomes following CRISPR-induced DSBs [262, 277]. We thus tested whether other sgRNAs as well as the CRISPR nuclease type might influence HR

repair and the interplay between the HR and alt-EJ pathways in CHO-K1 HR reporter cells. To do so, we combined various GFP-targeting sgRNAs with either the wild-type Cas9 or the FokI-dCas9 CRISPR nuclease systems. Depending on the DSB location, these sgRNAs should generate protruding flaps ranging between 17 nt (G7 sgRNA) to 94 nt (G1 sgRNA) during HR (Figs. S3.1 and S3.7, Table S3.8). These different sgRNAs and nucleases mediated HR repair in around 0.4-1.0% of the cells, irrespective of whether we used the wild-type Cas9 or FokI-dCas9 nuclease, although the latter exhibited an overall lower nuclease activity (Figs. S3.1 and S3.13A). This implies that short 5' overhanging DSBs, as generated by FokI-dCas9, are better substrates for HR repair than the blunt-ended DSBs of wild-type Cas9, as previously noted for long 5' overhanging DSBs [306, 370].

To assess whether the HR-mediated repair of staggered DSBs introduced by FokI-dCas9 depends on alt-EJ factors regardless of the sgRNAs used, the G1G5 and G3G7 sgRNA pairs were tested in cells transfected twice with a selection of siRNAs. Overall, the responses to specific siRNAs were similar using either sgRNA pair, i.e. a strong reduction in HR frequency for Rad51 and most alt-EJ knockdowns and an increase upon Mre11 and Pari knockdowns (Figs. S3.13B and S3.13C).

Given that alt-EJ factors appeared to stimulate HR repair from two distinct FokI-dCas9 DNA cleavage sites, we assessed whether this could reflect a nuclease- and/or DSB structure-specific effect. Therefore, we compared HR repair induced by the wild-type Cas9 and G3 sgRNA with the FokI-dCas9 and G1G5 sgRNA. These two nuclease systems mediate DNA cleavage at the same DNA site but generate either blunt or staggered DSBs, respectively (Fig. S3.1A). Intriguingly, HR repair elicited the wild-type Cas9 and G3 sgRNA did not rely on alt-EJ factors, unlike FokI-dCas9-mediated staggered ends (Figs. S3.13C and S3.13D). The response to most other knockdowns remained unchanged. This indicated that the alt-EJ implication in HR does not depend on the DNA sequence surrounding the DSB, but that it may rather relate to the nuclease and/or to the DSB structure.

To further assess whether the HR repair of wild-type Cas9-mediated DSBs depends on alt-EJ factors in other DSB contexts, the HR reporter cell line was transfected with two other sgRNAs, namely G1 or G7, and the wild-type Cas9 expression vector. These sgRNAs target the non-template and template GFP strand, respectively, and introduce DSBs located at different distances from the GFP deletion (Fig. S3.1A, Table S3.8).

Using the deletion-proximal G7 sgRNA, knockdown of alt-EJ factors did not impede GFP reconstitution, as observed with the G3 sgRNA (Fig. S3.13E). Using the deletion-distal G1 sgRNA, which creates a 94 nt long protruding flap, deficiency of alt-EJ factors restricted HR frequency to levels similar to those observed from the FokI-dCas9 nuclease (Figs. S3.13B, S3.13C and S3.13G). Hence, it appears that the distance between the DSB and mutation site, which correspondingly affects the length of the protruding flap, modulates the involvement of alt-EJ factors in wild-type Cas9-mediated HR repair.

To better visualize similarities and differences between all tested sgRNA, nuclease and cell line conditions, we performed a heat map and clustering analysis. This analysis highlighted conserved knockdown responses, as observed for instance in Rad51, Mre11 and Pari (Fig. 3.4A). However, cluster analysis divided the experimental conditions into two distinct groups, which correlated with the observed HR dependency on alt-EJ factors. The first group consists of wild-type Cas9 and G3 as well as G7 sgRNA-mediated DSBs which did not rely on alt-EJ factors for gene correction, or were even antagonized by the alt-EJ Pol theta factor. The second group comprises all other DSBs which strongly depended on the alt-EJ activities, e.g. staggered DSBs induced by the FokI-dCas9 with paired sgRNA G3G7 or blunt-ended DSBs induced by the wild-type Cas9 with G1 sgRNA. We therefore concluded that alt-EJ contribution to efficient HR repair may be influenced by both the CRISPR nuclease system as well as by the DSB-to-mutation distance, and that specific favorable effects – as obtained from Mre11 or Pari knockdown – can be observed across a wide range of experimental conditions.
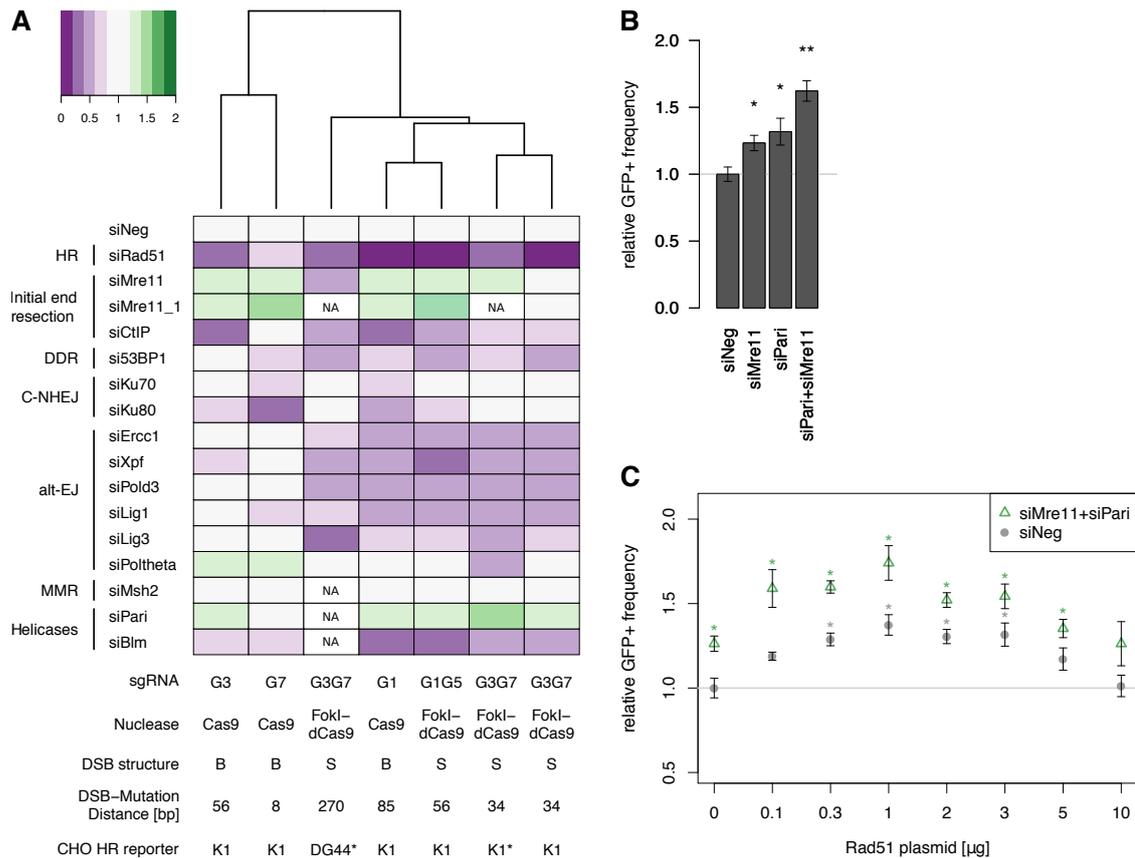
Figure 3.4: siRNA knockdown effects on GFP reconstitution using different sgRNAs, CRISPR nucleases, and HR reporter cell lines, and amelioration of GFP reconstitution by HR. The heat map colors indicate a decrease (purple) or increase (green) in GFP reconstitution among the transfected dsRed+ cells as a consequence of a specific knockdown relative to the corresponding siNeg control. G1, G3, G7 sgRNAs were used with the wild-type Cas9 nuclease, which creates mostly blunt-ended DSBs (abbreviated as B), whereas the G1G5 and G3G7 sgRNAs were used with the FokI-dCas9 nuclease, which creates 4 nt 5' staggered DSBs (abbreviated as S). The DSB-to-mutation distance varies and depends on the sgRNA, as indicated in Figure S3.1A. Stars denote knockdown experiments using a single siRNA treatment only in CHO-K1 and CHO-DG44 reporter cells (Figs. 3.3A and S3.11G). The remaining conditions were obtained from double siRNA transfections, for which the raw data plots are shown in Figs. S3.13B-S3.13F. siMre11 and siMre11_1 indicate transfections using the siRNA pool and the most efficient singular siRNA targeting Mre11, respectively. All other samples were treated with siRNA pools. Clustering analysis uses Manhattan distance metric (NA: not available). (B-C) GFP reconstitution frequency after transfection of the G3G7 sgRNA, FokI-dCas9 nuclease, donor and dsRed expression plasmids in the CHO-K1 reporter cell line. The reporter cells were doubly transfected with siRNA targeting Mre11 and Pari without (B) or with varying amount of Rad51 expression plasmids (C). Data show frequency of GFP+ cells among the transfected dsRed+ cells expressed relative to siNeg control. For panel C, data is normalized to the siNeg control without Rad51 overexpression (0 μg). siRNA transfections were performed using the most efficient singular siRNA from each siRNA pool and complemented with negative siRNA, while overexpression samples were complemented with empty vector, to keep a constant siRNA and plasmid amount (n $\geq$ 3, error bars represent s.e.m.). P values (* P < 0.05, ** P < 0.01) were calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction.

3.4.8 Counteracting HR rate-limiting steps improves GFP reconstitution

The identification of various limiting or competing activities that may prevent efficient GFP reconstitution implies that several bottlenecks need to be removed simultaneously to ameliorate HR efficiency in CHO cells. In CHO-K1 cells, limited Rad51 and elevated Mre11 and Pari activities appeared to antagonize productive HR, thus raising the question of their possible interplay or synergism. While the knockdown of Mre11 or Pari individually increased the frequency of GFP+ cells by 25-35%, combined knockdowns led to an additive increase in GFP reconstitution by 60%, thus further optimizing HR activity (Fig. 3.4B). As the rate-limiting activity of HR proteins could also constitute a bottleneck, we overexpressed Rad51, the most rate-limiting HR factor among all tested conditions, in combination with Mre11 and Pari knockdowns, to assess GFP reconstitution frequency. As shown in Figure 3.4C, HR repair efficiency increased up to 75% when combining Mre11 and Pari knockdown with low to medium Rad51 overexpression, substantiating the previously described dose-dependent effect of overexpressing rate-limiting HR factors (Fig. S3.10A). Cells exposed to this optimized HR booster treatment showed no sign of increased cellular toxicity, with similar frequencies of dsRed+ cells than the control sample (data not shown). Moreover, the higher HR frequency obtained upon combined knockdowns and Rad51 overexpression did not increase the illegitimate integration frequency, as based on the occurrence of less than 0.01% of dsRed+GFP+ cells after one month. This implies that the established tailored treatment specifically increased HR but not other repair pathways that mediate illegitimate recombination.

3.5   Discussion

The advent of programmable site-specific nucleases has opened up unique opportunities to simplify HR-mediated precise genome editing, notably in cell lines with low endogenous HR levels. However, the supremacy of DNA repair pathways mediating mutagenic end-joining repair creates a major bottleneck for precise editing in most cells. In this study, we investigated the role of alt-EJ factors as potential modulators of HR-mediated gene correction in CHO cells. To do so, we designed novel chromosomal GFP HR assays allowing the comparison of the repair of FokI-dCas9 and wild-type Cas9 generated DSBs

next to GFP loss-of-function deletions. We also assessed the effect of varying the distance between the CRISPR target sites and the GFP mutation, as this leads to the formation of protruding non-homologous flaps of variable length during HR repair.

Interestingly, the knockdown of several alt-EJ factors, such as the Ercc1-Xpf endonuclease, the Pold3 DNA polymerase as well as the Lig3 DNA ligase, diminished HR in both CHO-K1 and CHO-DG44 HR assays, contrary to what one would expect from competing pathways. Furthermore, we observed reduced HR efficiency in alt-EJ factors-depleted cells when using various sgRNA target sequences and CRISPR nuclease systems. Taken together, our results imply that efficient HR repair in CHO cells relies on subsets of alt-EJ factors in specific experimental settings. This evokes the intervention of non-canonical alt-EJ activities in HR pathways, in addition to their well-established function in microhomology-driven repair.

These CHO cell-based findings that certain alt-EJ factors may be involved in HR repair may explain or support previous observations made using other model systems. For instance, Parp1 was observed to mediate gene conversion of the immunoglobulin genes in chicken cells, besides its well-documented activities in base excision repair and single strand break repair [371]. The Ercc1-Xpf endonuclease (Rad1-Rad10 in yeast) was also associated with removing flap structures during HR and SSA in both yeast and mammalian cells [372–374]. Although the Ercc1-Xpf complex is as well a key nuclease during nucleotide excision repair, no other nucleotide excision repair factor seems to contribute to alt-EJ and/or HR repair [373]. Moreover, Pold3 is not only critical in break-induced replication, a HR subpathway at one-ended DSBs, but it may also play a role in gene conversion in yeast [92]. Finally, Lig1, Lig3 and Xrcc1 may also contribute to the ligation of DNA ends during HR repair, in addition to their roles in excision and single strand break repair [375].

Therefore, overall evidence suggests that alt-EJ factors non-canonically engage in processing HR intermediates. More specifically, alt-EJ factors may favor the pairing and annealing of short homologous regions, contribute to removing the flaps at both the invading and non-invading DNA ends, promote strand extension and end ligation using a MMEJ-like mechanism, until bona fide HR mechanisms can resume (Fig. 3.5, yellow ellipses). Such an alt-EJ mechanism would make DSB repair more efficient at donor sequences with short homologies and/or more tolerant to long non-homologous flaps, as seen here for G1 sgRNA-mediated wild-type Cas9 cleavage. These alt-EJ activities may be less important

for the repair of blunt-ended DSBs with short protruding flaps. Ultimately, the interplay between alt-EJ and HR factors would favor conservative SDSA and DSBR pathways over mutagenic end-joining pathways.
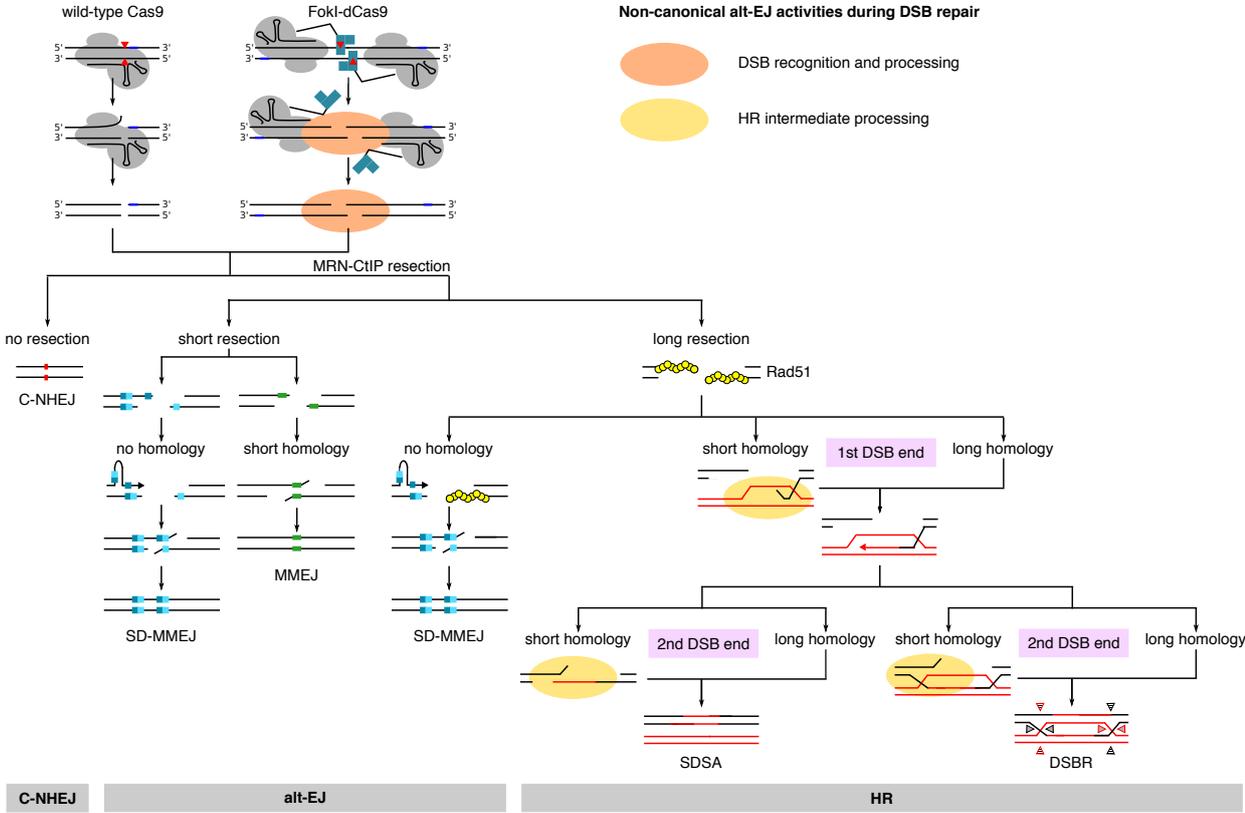
Figure 3.5: Model for alt-EJ factor contribution to HR in CHO cells. Besides the previously suggested functions of alt-EJ factors in MMEJ and SD-MMEJ end-joining pathways, alt-EJ factors may have additional non-canonical alt-EJ activities during DSB repair which stimulate productive HR repair at two stages. First, at an early stage, alt-EJ factors may be required to promote DSB recognition and processing at overhanging DSB ends, notably at staggered FokI-dCas9-induced DSBs (orange ellipses). Alt-EJ factors are not necessarily required for the repair of blunt DSBs induced by wild-type Cas9 nuclease at this early stage. Second, at a later stage, DNA repair may involve the alt-EJ family of DNA repair mechanisms when Rad51 filaments are formed upon long strand resection but fail to locate long homologous sequences (yellow ellipses). In the presence of a short microhomology and/or a protruding non-homologous flap, alt-EJ factors may then contribute to end annealing, flap removal, strand extension and end ligation in a MMEJ-like mechanism to allow DSB repair completion via bona fide HR mechanisms, such as SDSA and DSBR. In the absence of even a short homology, the DNA polymerase-dependent SD-MMEJ pathway may duplicate a short sequence to allow for a sufficiently long microhomologous sequence at the broken DNA end, resulting in the templated insertion of the duplicated sequence at the repaired junction [116, 141]. PAM and CRISPR cleavage sites are shown in blue rectangles and red triangles, respectively. Triangles in the DSBR pathway represent nucleases.

As such, this model of the alt-EJ contribution to HR does not suffice to explain the finding that HR repair arising from FokI-dCas9-mediated DSBs consistently depended on Ercc1-Xpf, Pold3 and Lig3 alt-EJ factors, unlike wild-type Cas9-induced DSBs. This suggests that alt-EJ factors may have other non-canonical activities at an early DSB repair stage. They could, for instance, assist in DSB recognition and processing, i.e. before C-NHEJ, alt-EJ and HR repair pathways diverge (Fig. 3.5, orange ellipses). The alt-EJ factors may be recruited specifically to FokI-dCas9, but not wild-type Cas9 cleavage sites, possibly because of the distinct nuclease dynamics and/or DSB end structures elicited by these two nuclease types.

In contrast to the obligate dimeric FokI-dCas9 nuclease, where weak FokI-FokI interactions between two catalytically dead Cas9 proteins mediate DNA cleavage [275], the monomeric wild-type Cas9 remains tightly bound to the cleaved DNA sequence. It also releases the DSB ends asymmetrically, starting with the PAM-distal site of the non-target DNA strand [264]. We therefore hypothesize that, when compared to wild-type Cas9, the nuclease dynamic of FokI-dCas9 may facilitate DSB end release, making the cleaved DNA sequence more readily accessible to DSB repair proteins, and notably to alt-EJ factors.

In addition to distinct nuclease kinetics, elevated alt-EJ factor recruitment to FokI-dCas9 cleavage sites could be associated with the DSB end structure. While the wild-type Cas9 nuclease creates mostly blunt-ended and occasionally 1 nt 5' overhanging DSBs [262], the FokI-dCas9 nuclease produces 4 nt 5' overhanging DSBs. A preference of alt-EJ factors for 5' overhanging versus blunt-ended DSBs is consistent with previous observations that DSBs with ssDNA overhangs seem to attract overall more types of repair proteins than blunt-ended DSBs [376] and that 5' overhangs appear to favor alt-EJ repair at chromosomal DSBs [159]. In support of an augmented alt-EJ activity at DSBs cleaved by the FokI-dCas9 nuclease, non-HR GFP repair junctions obtained using this nuclease often displayed alt-EJ-specific repair signatures, while they were devoid of C-NHEJ-characteristic 1-4 bp indel mutations typically observed after wild-type Cas9 cleavage [377]. Together with the finding that the more specific, albeit less active, FokI-dCas9 induces HR as efficiently as wild-type Cas9, these observations indicate that alt-EJ and HR pathways may preferentially repair FokI-dCas9-mediated DSBs, while C-NHEJ repair is mainly active at wild-type Cas9-mediated DSBs. However, it remains to be established whether the different DSB pathway activities observed between blunt and overhanging DBSs might also depend on the formation of local secondary structures and thereby in-

fluence recruitment of repair factors [140]. The proposed involvement of alt-EJ factors in DSB recognition and processing further suggests that alt-EJ factors may directly or indirectly contribute to the DDR signal, for instance by recruiting upstream signaling factors [378], and/or that they may assist in dislodging the dCas9 protein from the target DNA sequence [261].

Consistent with this model, evidence is mounting that a priori competitive DSB repair pathways may need to be coupled to mediate timely and proper repair, and hence to ensure conservation of the genome integrity. For instance, when homologous sequences are too short to accommodate efficient HR, then alt-EJ factors may direct DSB repair to salvage SD-MMEJ or MMEJ pathways, some of which being Rad51-dependent [116, 141]. Moreover, previous studies proposed that end-joining factors assist in completing HR events, especially during long-tract gene conversion, by annealing the invading DNA end to the second non-invading chromosomal end [293, 379]. Coupling of HR and end-joining pathways is commonly believed to cause mutations and complex inversions/duplications [379]. However, here we find that a collaboration between alt-EJ and HR activities can be crucial to achieve more efficient precise genome editing in CHO cells.

DSB repair pathway activity and/or regulation can vary between different cell types and/or organisms [164, 166, 167]. Indeed, we and others have noticed that CHO cells inherently differ in their DSB repair capacities as compared to primary and non-malignant cells, as they possess high alt-EJ but low C-NHEJ and HR activities [141, 171, 291]. Distinct C-NHEJ mechanisms in CHO cells may possibly explain the unexpected observation that knockdown of C-NHEJ factors did not alter and under some conditions even reduced HR activity. Although these findings contrast other studies in mouse and human cells, they are coherent with other CHO cell-based studies [141, 317–320]. In addition to their canonical C-NHEJ function, Ku70 and Ku80 are associated with apoptotic signaling, base excision repair and interaction with other repair proteins, such as Parp1 and Wrn [380, 381]. Whether the lack of increased HR efficiency in C-NHEJ-deficient background relates to such additional roles of Ku70 and Ku80 and/or to the overall low C-NHEJ activity in CHO cells remains to be established.

Another interesting observation is that Mre11 knockdown increased HR frequency in the CHO-K1 assays. Mre11 together with Rad50 forms the heterotetramer core of the MRN-CtIP complex – a pivotal complex responsible for tethering broken ends as well as initiating end resection [8]. Therefore, Mre11 knockdown is expected to block initial end

resection and strongly reduce HR frequency [382], which agrees with the CHO-DG44 model but is inconsistent with the CHO-K1 data. This suggests that a proper ratio of the MRN subunits might be required for optimal HR activity and that Mre11 may be expressed at higher levels than other MRN subunits in CHO-K1 cells. Indeed, CHO-K1 RNA sequencing experiments revealed that the mRNA levels of Mre11 are higher than those of Rad50 and Nbs1. Excess Mre11 protein could sequester Rad50 and Nbs1 proteins and thereby prevent proper MRN complex formation. This stoichiometric imbalance may be restored by knocking down Mre11, leading to improved HR activity. In support for this hypothesis, yeast cells expressing a nuclease-deficient Mre11 mutant were not impaired in HR [383], unlike Mre11 knockout cells, suggesting that MRN complex formation rather than Mre11 nuclease activity facilitates HR repair. However, we cannot exclude that Mre11 knockdown might also affect the stability and localization of other DNA repair factors, as Mre11 depletion can lead to the co-degradation of Rad50 and Nbs1, as well as Fancd2, a factor involved in interstrand crosslink repair [384].

Of special interest is the finding that most if not all four GFP reporter allelic copies underwent concomitant HR-mediated GFP correction in the CHO-K1 cell line. Such an "all-or-nothing" GFP reconstitution mechanism was not anticipated, given the low efficiency of HR in CHO cells. One explanation is that cells being in a HR-compatible state are more likely to deploy HR repair at multiple distinct CRISPR target sites in the genome, leading to GFP co-conversion. These co-conversions presumably result from independent repair events and not from one corrected allele acting as donor template for other unrepaired alleles [385]. Consistently with the frequent co-conversion observed, the percentage of GFP+ cells did not correlate with the number of HR reporter copies when using various CHO-K1 HR reporter clones (data not shown).

Overall, this work contributes to revealing the factors governing the tight regulation of DSB repair pathways. It uncovers alt-EJ factors as a potential bottleneck for efficient precise genome editing, thus providing novel avenues to more efficient, predictable and less mutagenic gene correction approaches. This highlights the importance of the functional characterization of as yet poorly understood alt-EJ DNA repair pathways and deciphering how they might promote genome stability in CHO and possibly other cancer or immortalized cells lines.

3.6 Supplementary Data
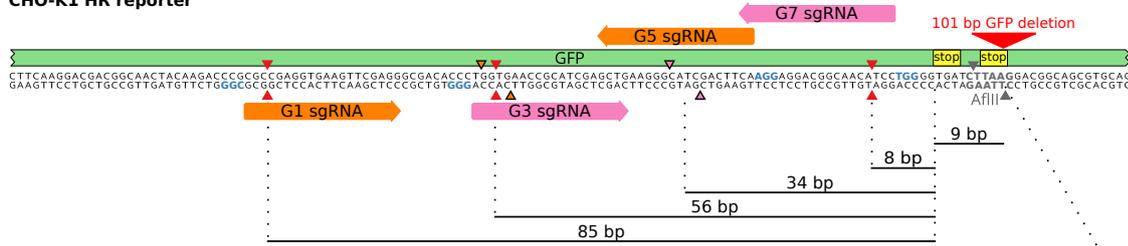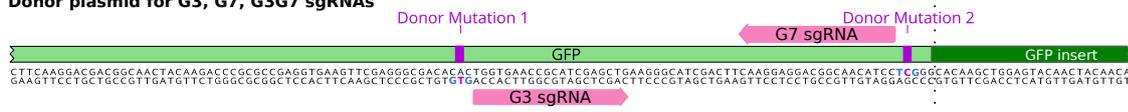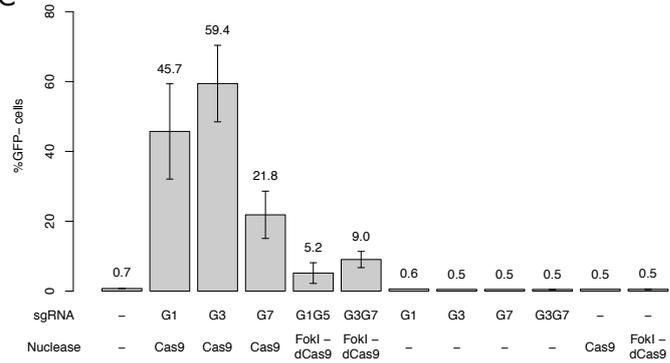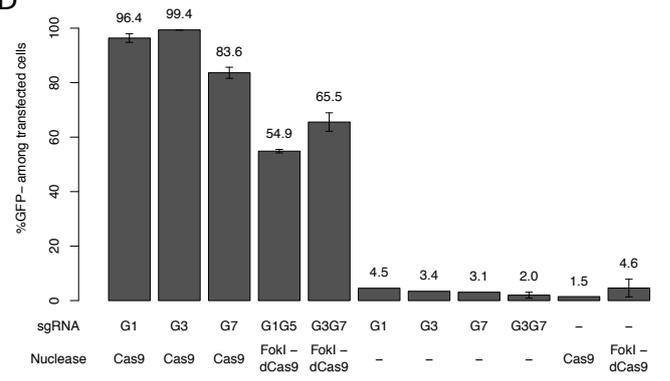
Figure S3.1: Recognition sites and GFP cleavage efficiency of different sgRNAs and nuclease systems. (A) Sequence, location, and orientation of the tested sgRNAs targeting the GFP sequence of the CHO-K1 HR reporter construct. All listed sgRNAs can be used individually with the wild-type Cas9 nuclease or in combination (sgRNA G1 with sgRNA G5 as depicted with orange arrows or sgRNA G3 with sgRNA G7 as pastel pink arrows) with the FokI-dCas9 nuclease [275]. PAM sites are indicated by blue color-lettered bases. Predicted CRISPR cleavage sites are marked with red (wild-type Cas9), orange (G1G5 sgRNA + FokI-dCas9) or pastel pink (G3G7 sgRNA + FokI-dCas9) triangles. Caption continued next page.

Figure S3.1: The AflII recognition and cleavage sites are shown in grey letters and triangles, respectively. GFP truncation results from two in-frame stop codons (yellow boxes) and a 101 bp GFP deletion (red), and is used as chromosomal HR reporter. The indicated black lines denote the distance in base pairs between the GFP deletion start site and the predicted sgRNA cleavage sites (upstream) and as well as the additional 9 bp mismatch arising from the stop codon insertions (downstream). (B) Detailed view of the donor sequences used to repair different sgRNA-induced DSBs. The two silent PAM mutations are shown in purple (Donor Mutation 1 and Donor Mutation 2) and the portion of the GFP sequence required to correct the GFP deletion in the HR reporter is depicted in dark green (GFP insert). (C-D) CRISPR cleavage efficiency was assessed by flow cytometry analysis of CHO-DG44 cells bearing functional chromosomal GFP coding sequences following the transfection of the indicated sgRNA and nuclease expression plasmids. It was quantified as the frequency of loss-of-function GFP mutations among the entire cell population (C) or among the transfected dsRed+ cells only (D) (n=1 or n=2, error bars represent s.e.m.).
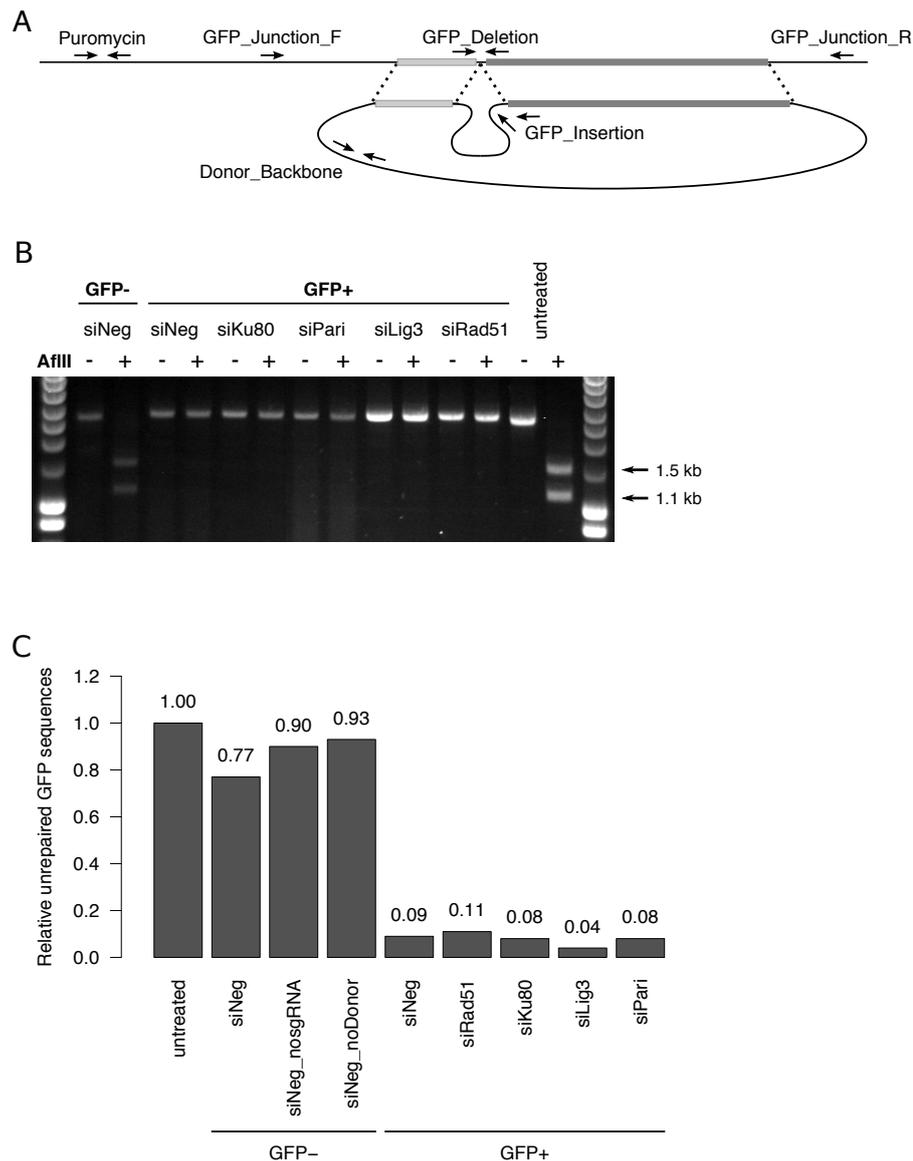
Figure S3.2: Detection and quantification of repaired and unrepaired GFP sequences. (A) End-point and qPCR primer strategies used to amplify the GFP coding sequence of the HR reporter (GFP_Junction amplicon, as generated using the GFP_Junction_F and GFP_Junction_R primers) and to quantify total donor plasmid integration plus proper GFP correction events (GFP_Insertion primer pair), donor plasmid integration only (Donor_Backbone primer pair), and unrepaired chromosomal GFP sequences (GFP_Deletion primer pair). Homology arms are shown as grey rectangles. (B) The GFP coding sequence of polyclonal cell populations sorted to display (GFP+) or not (GFP-) GFP fluorescence were amplified using primers that hybridize to the chromosomal HR reporter, but outside of the donor plasmid homology arms (GFP_Junction_F and GFP_Junction_R primers), as shown in panel A. PCR products were digested without (-) or with (+) the AflII restriction enzyme. The AflII recognition site locates 38 bp downstream of the G3G7 sgRNA-induced DSB (Fig. S3.1A) and AflII-specific cleavage products of 1.1 kb and 1.5 kb size are marked with arrows. The correctly repaired donor-derived chromosomal GFP sequence is expected to be devoid of an AflII recognition site. (C) The amount of unrepaired GFP sequences in GFP- and GFP+ samples relative to the untreated CHO-K1 HR reporter was estimated by qPCR using the GFP_Deletion primer pair, as depicted in panel A. Repair protein knockdown was performed using siRNA pools.
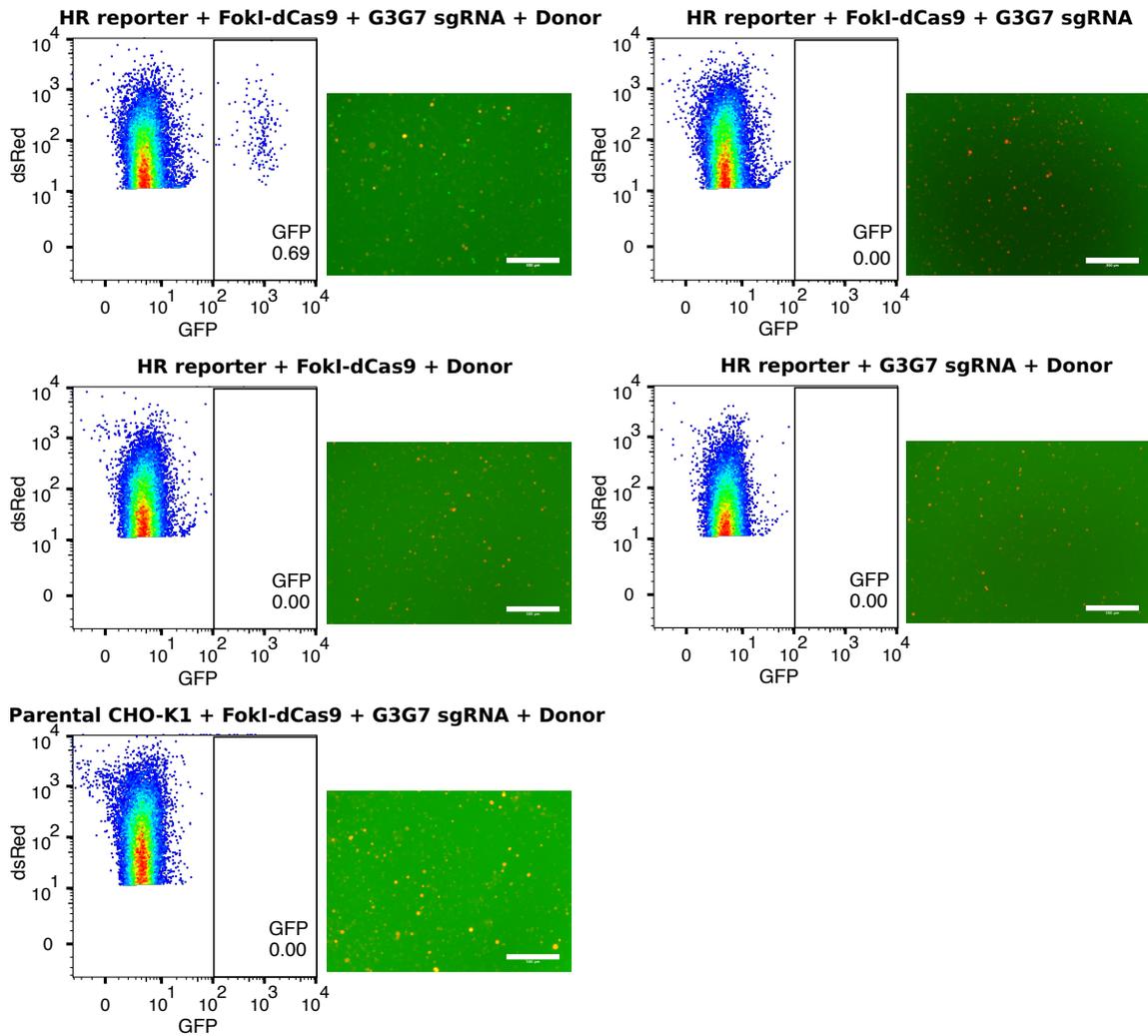
Figure S3.3: Assessment of GFP reconstitution in CHO-K1 HR reporter and parental CHO-K1 cells. Flow cytometry profiles and microscopy images represent GFP reconstitution in either HR reporter construct-containing cells (HR reporter) or parental CHO-K1 cells transfected with different combinations of the GFP-specific G3G7 sgRNA, the FokI-dCas9 nuclease, and the GFP donor expression plasmids, as indicated (Scale bar 300 μm). A co-transfected dsRed expression plasmid serves as transfection control. Flow cytometry profiles show GFP expression among the transfected dsRed+ cells only.
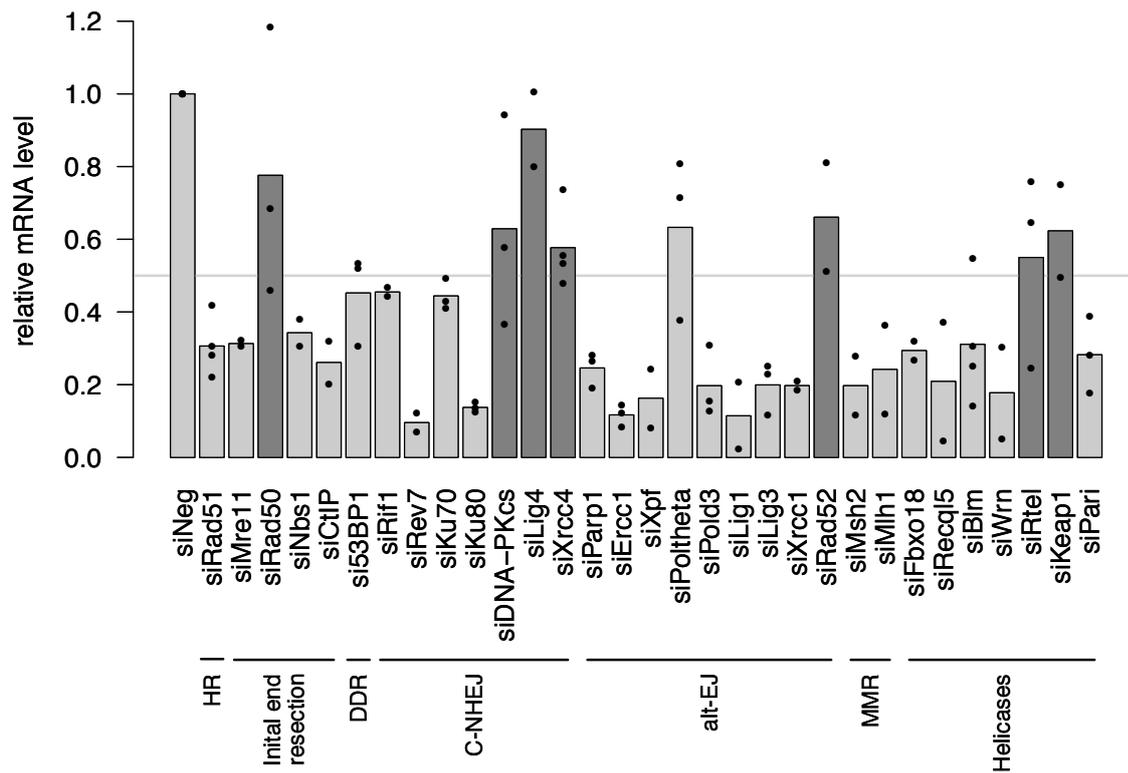
Figure S3.4: siRNA knockdown effect on target mRNA level. siRNA knockdown efficiency of siRNA pools as quantified by RT-qPCR with target-specific primers. Data represent mRNA levels of indicated siRNA-treated cells relative to those of siNeg-treated control cells. Black points show data from independent experiments, whereas bars depict mean values (n ≥ 2). siRNA pools targeting Rad50, DNA-PKcs, Lig4, Xrcc4, Rad52, Rtel1, and Keap1 were excluded from further knockdown assays, as they did not reduce their target mRNA level by 50% in CHO-K1 cells and/or yielded inconsistent inhibitory effects (dark grey bars), whereas the most efficient Pol theta-targeting siRNA and/or the siRNA pool were used in subsequent assays (Fig. S3.5).
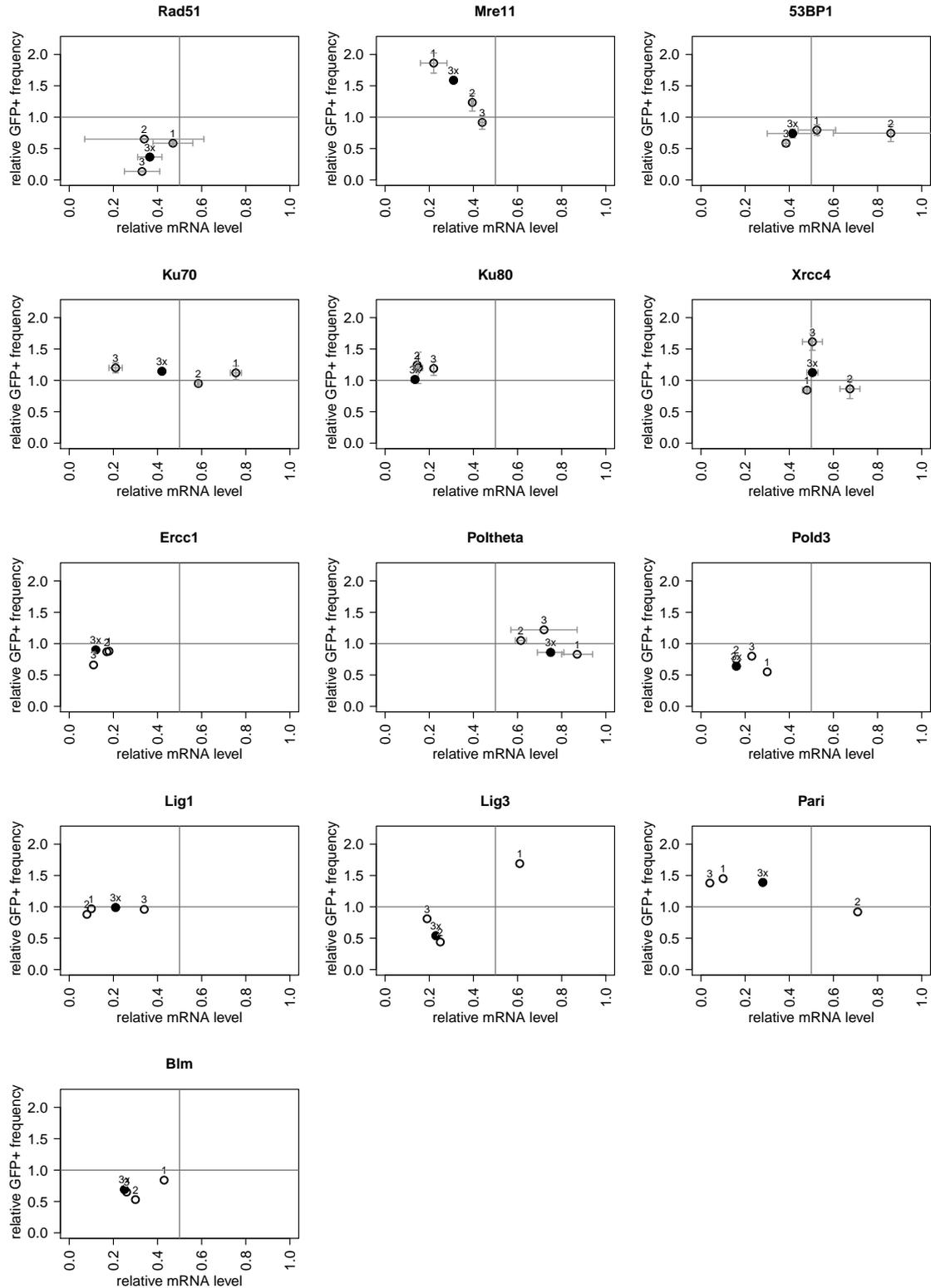
Figure S3.5: Deconvolution of siRNA pools. Single siRNAs (labeled 1, 2, 3, open circles) were assessed for their individual knockdown efficiency by target mRNA RT-qPCR quantification and for their effect on GFP reconstitution by flow cytometry in comparison to their corresponding three siRNA pools (3x, closed circles). siRNAs were used at 33 nM each for the siRNA pool and at 100 nM for single siRNA assays (n=1 or n=2, error bars represent s.e.m.).
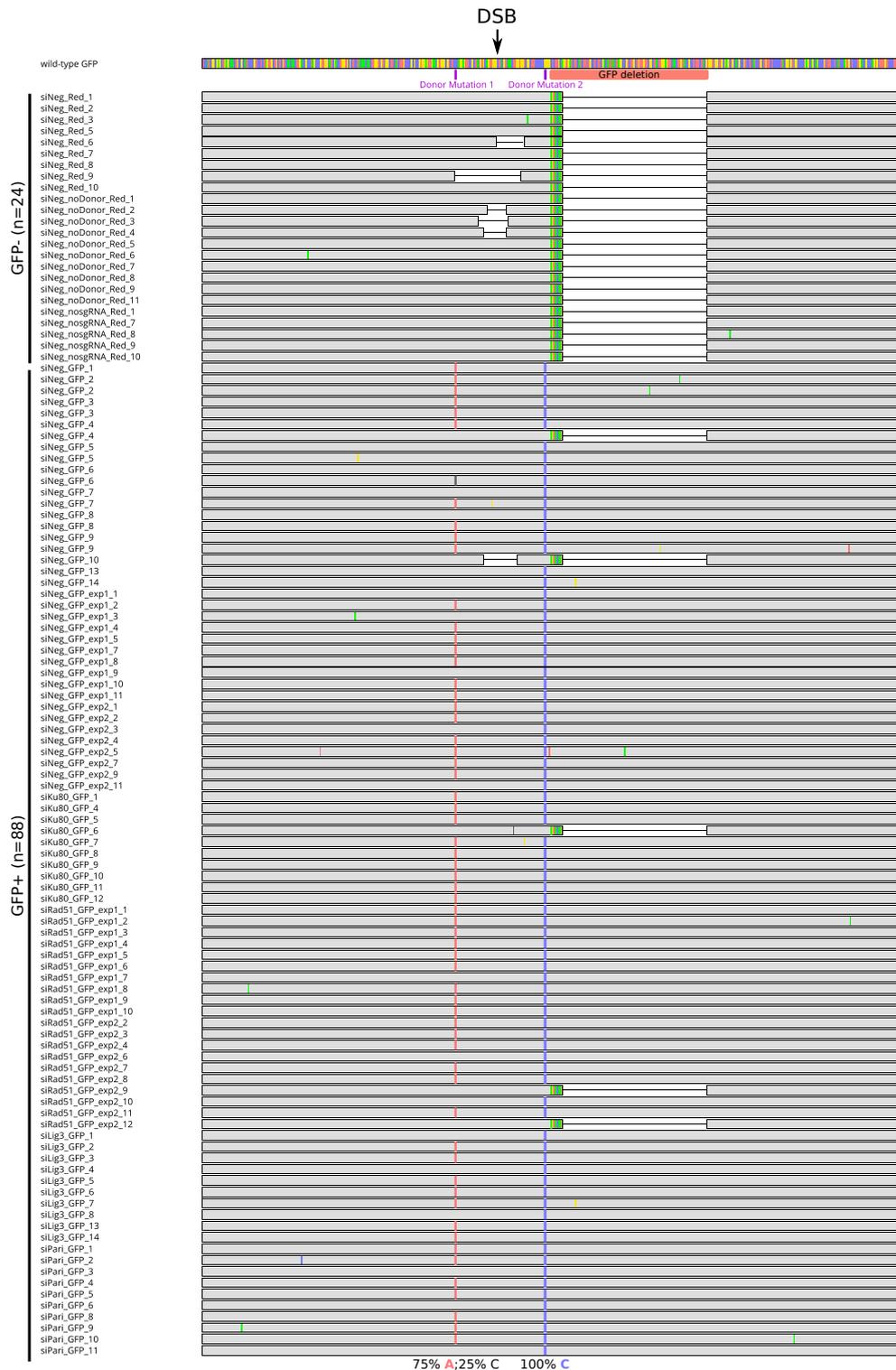
Figure S3.6: Sequence alignment of the chromosomal GFP coding sequence of GFP- or GFP+ sorted cells. GFP coding sequence products of GFP- and GFP+ sorted CHO-K1 HR reporter cells were amplified using primers located outside of the homology arms (Fig. S3.2A), cloned into plasmid vectors, transformed into bacteria, and individual bacterial colonies were Sanger sequenced. A total of 116 sequences were analyzed, of which 24 GFP- and 88 GFP+ sequences are shown. Four sequences with large genomic rearrangements are not included in this alignment but listed in Table S3.4. Donor-derived PAM mutations are marked with red (C to A substitution, Donor Mutation 1) and blue bars (G to C substitution, Donor Mutation 2). siRNA pools were used for repair protein knockdown.

78

Figure S3.7: Proposed HR repair mechanisms leading to GFP gene correction at DSB site with mismatch sequences. DSBs can initiate extensive DNA end resection of more than 100 bp that is followed by strand invasion of the single-stranded DNA ends into homologous sequences, here a homologous donor plasmid. As the donor plasmid contains two silent PAM mutations (purple bars), the pairing between the invading strand and the donor plasmid leads to mismatched bases (boxed green cone and purple bar) 29 bp away from the DSB site, which activates the DNA mismatch repair and can lead to the incorporation of the upstream donor mutation in the repaired product (grey bar). DNA synthesis along the homologous template will copy possible new, non-homologous sequences from the donor sequence (dotted dark green). HR repair can then proceed via SDSA and DSBR. Caption continued next page.

Figure S3.7: During SDSA the newly synthesized strand of the invading DSB end anneals to the complementary sequence at the second DSB end. This creates a protruding non-homologous 3' flap (yellow bar) – as a consequence of the distance between the DSB site and GFP deletion. Removal of this protruding flap is required to complete SDSA repair. During DSBR, the second DSB end also pairs with the donor plasmid, which creates a protruding flap at this end, as during SDSA. Flap excision is followed by DNA synthesis and double Holliday Junction (dHJ) formation. dHJ can then either be branch-migrated and dissolved, or it can be resolved by nuclease cleavages (shown in triangles) to produce gene correction as well as plasmid knock-in repair products. Unlike the upstream donor mutation, 100% of the repair products should incorporate the donor mutation located downstream of the DSB, as complete flap removal (which contains no PAM mutation) circumvents random base conversions during mismatch repair. The drawing portrays the genomic situation in the CHO-K1 HR reporter upon DSB induction with the G3G7 sgRNA and FokI-dCas9 nuclease, followed by strand invasion of the upstream ("left") ssDNA end. Although upstream ssDNA end should be the favored end for invasion due to the absence of an immediate protruding flap, similar HR repair mechanisms should occur for the downstream ('right") ssDNA end as well as other DSBs involving protruding non-homologous 3' tails.

80
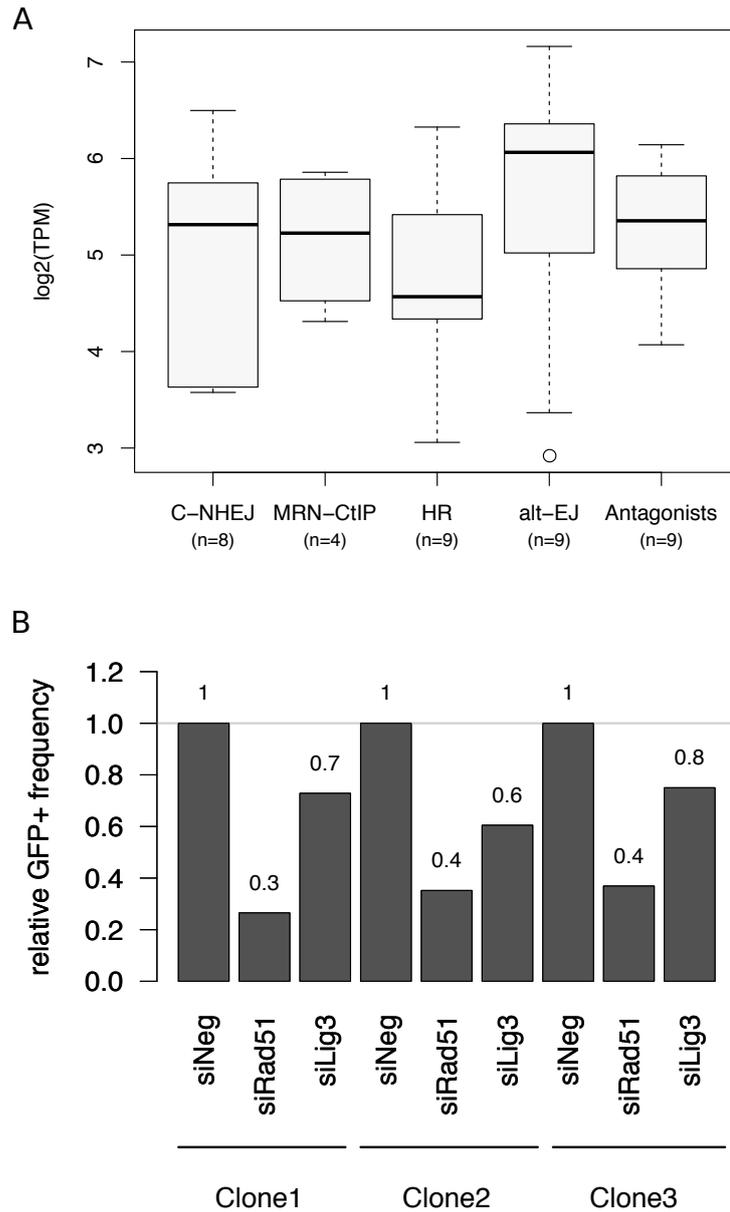
Figure S3.8: Repair protein gene expression in CHO-K1 cells and HR assay clonal reproducibility. (A) Boxplot showing mRNA expression levels of repair protein-encoding genes categorized by DNA repair pathway or function. mRNA expression is shown as log2 of the transcripts' read number per kilobase million (TPM) from a CHO-K1 RNA sequencing assay. The mRNAs of the following proteins were considered for analysis: 53BP1, Rif1, Rev7, Ku70, Ku80, DNA-PKcs, Lig4, Xrcc4 for C-NHEJ-associated activities; Mre11, Rad50, Nbs1, CtIP are grouped under the term MRN-CtIP for the strand resection functions involved in HR and alt-EJ pathways; Rad51, Brca1, Brca2, Rad51B, Rad51C, Rad51D, Xrcc2, Xrcc3, Rad54 for HR-involved activities, and Parp1, Ercc1, Xpf, Pol theta, Pold3, Lig1, Lig3, Xrcc1, Rad52 for alt-EJ pathways. Activities that may antagonize homology-directed repair mechanisms, such as the Msh2 and Mlh1 mismatch repair factors, as well as the Blm, Wrn, Fbxo18, Recql5, Rtel, Keap1, Pari helicases are grouped under the term Antagonists. (B) Effect of Rad51 and Lig3 knockdown on GFP reconstitution of three independent CHO-K1 HR reporter cell lines containing comparable copy numbers of the chromosomal HR reporter PB transposon. Knockdown of Rad51 and Lig3 was performed using siRNA pools. Clone1 was selected for subsequent experiments. Data shows GFP reconstitution frequency among the transfected dsRed+ cells from a single experiment.

81
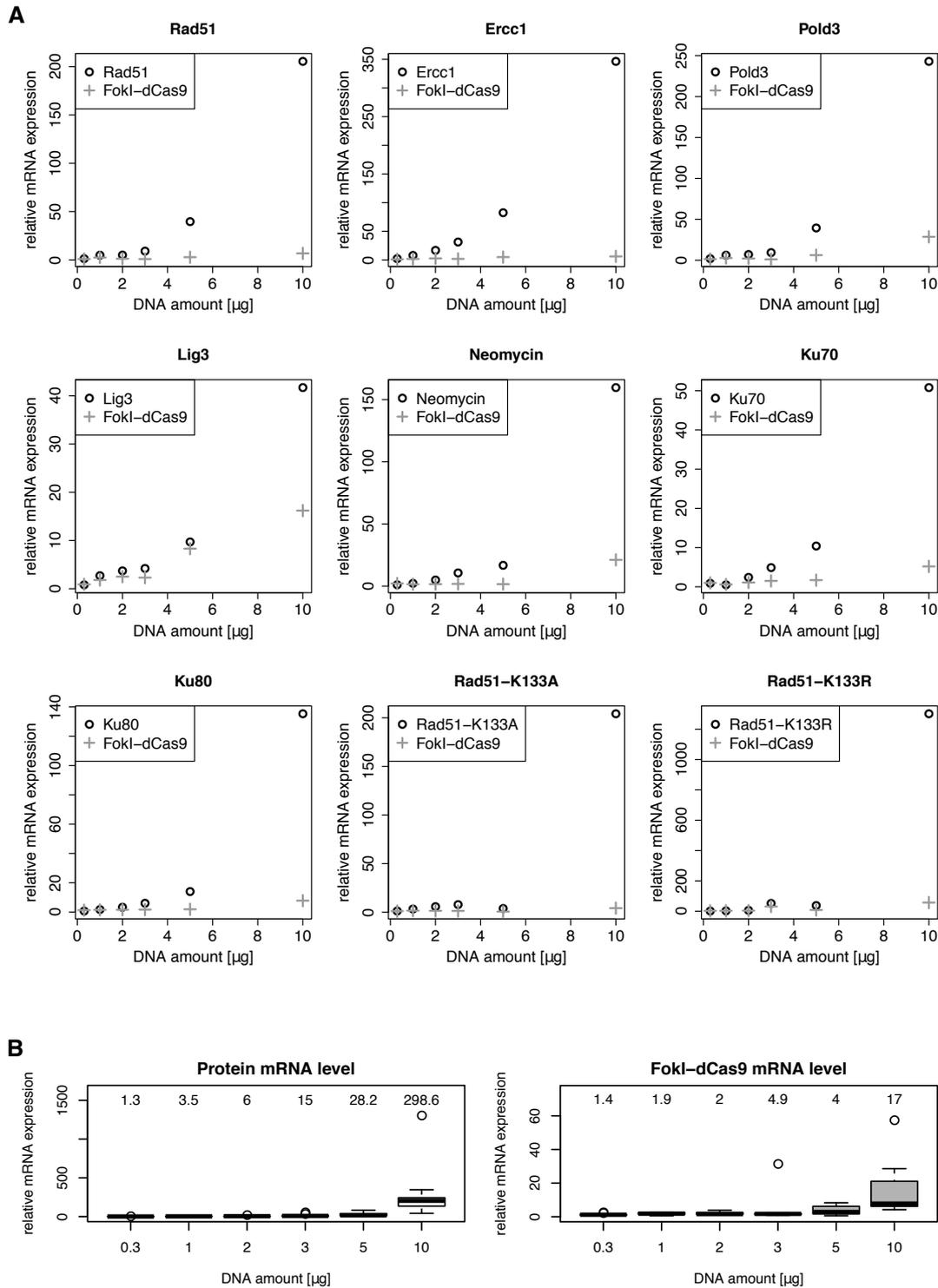
Figure S3.9: mRNA levels of FokI-dCas9 nuclease and overexpressed DNA repair pathway proteins. 0.3 μg, 1 μg, 2 μg, 3 μg, 5 μg or 10 μg DNA repair protein expression plasmids, as indicated, were co-transfected with the G3G7 sgRNA, FokI-dCas9 nuclease, donor, and dsRed expression plasmids. RNA was collected 24h after the transfection and mRNA levels were analyzed by RT-qPCR. (A) Graphs show relative mRNA expression of the DNA repair protein (closed circles) and FokI-dCas9 nuclease (grey crosses) separately per repair protein. (B) Protein (left) and FokI-dCas9 (right) mRNA expression data, as shown in A, were pooled together to calculate the average increase in mRNA expression for each DNA plasmid amount (written above each boxplot). Data are shown relative to the empty vector control (n=1).
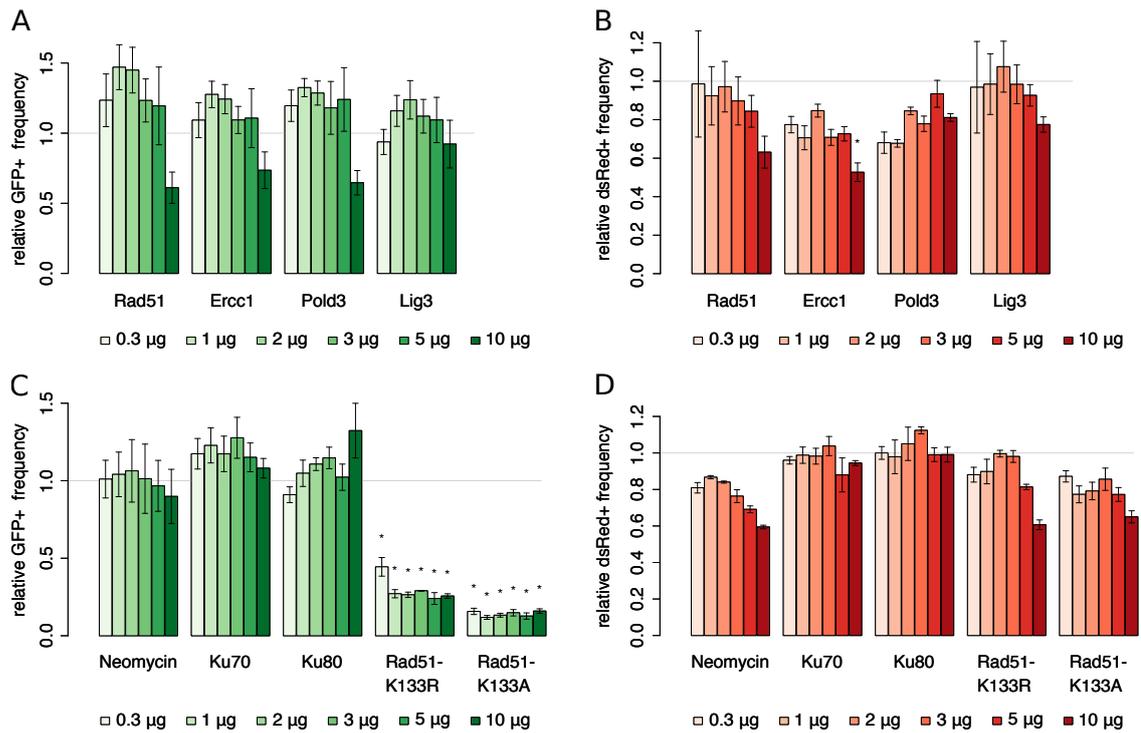
Figure S3.10: Overexpression of rate-limiting HR factors. GFP and dsRed expression was quantified by flow cytometry after the transfection of the indicated DNA repair protein and nuclease expression plasmids, as in Figure S3.9. Graphs show HR frequency as the proportion of GFP+ cells in response to the ectopic expression of HR rate-limiting factors (A) or control proteins (C), as well as the corresponding frequency of dsRed+ cells, as a proxy for the transfection efficiency and cell survival (B and D). The graphs show the frequency of GFP+ cells among the transfected dsRed+ cells (A and C) or frequency of dsRed+ cells (B and D) expressed relative to cells transfected with an empty vector control (n $\geq$ 3, error bars represent s.e.m.). P values (* P $<$ 0.05, ** P $<$ 0.01) were calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction.
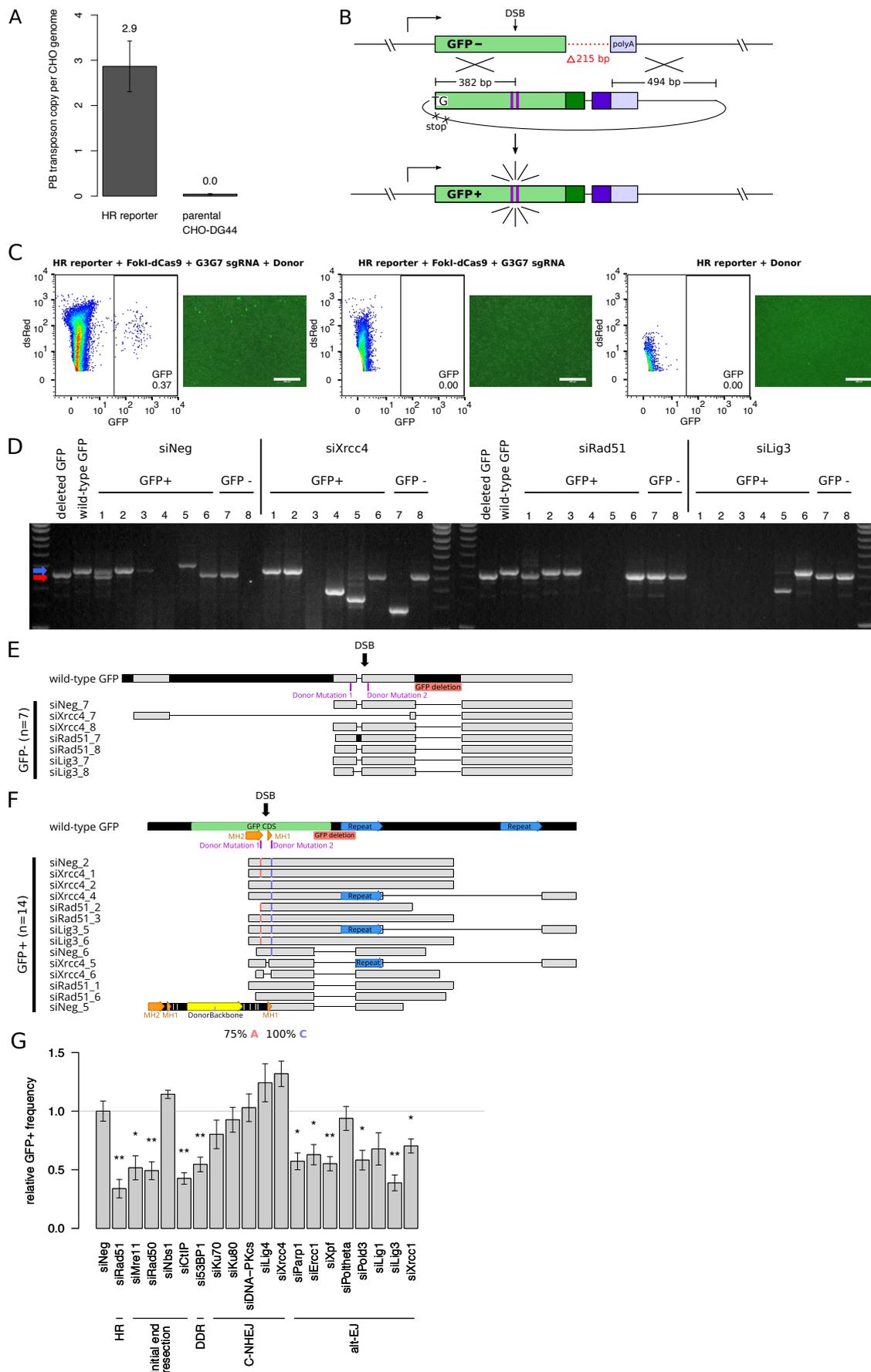
Figure S3.11: Caption on next page.

Figure S3.11: Characterization of the CHO-DG44 HR reporter cell line and analysis of siRNA knockdown effect on GFP reconstitution. (A) Copy number qPCR analysis of the number of integrated PB transposons per diploid CHO-DG44 genome. (B) Scheme of the HR assay developed in the adherent CHO-DG44 cell line. This assay shows a characteristic 215 bp frame-shift deletion (in red) within the GFP coding sequence and shares homology to the GFP donor plasmid (horizontal bracket lines). Upon DSBs induced by CRISPR nuclease, HR-mediated repair can be used to insert the missing GFP sequence leading to the reconstitution of GFP expression. The CRISPR target sites and donor plasmid are identical to those depicted in Figure 3.1A. (C) The CHO-DG44 HR reporter cells were co-transfected with G3G7 sgRNA, FokI-dCas9 nuclease, GFP donor and dsRed expression plasmids, or lacking one of these plasmids. Flow cytometry profiles show the GFP fluorescence of transfected dsRed+ cells only (Scale bar 300 μm). (D) GFP+ and GFP- CHO-DG44 cell clones were isolated after transfection with different siRNA pools and G3G7 sgRNA, FokI-dCas9 nuclease, donor and dsRed expression plasmids, and they were analyzed for their GFP sequences using PCR primers hybridizing to sequences located outside of the GFP donor homology arms (Fig. S3.2A). Expected amplicon sizes for repaired (2.7kb) and unrepaired (2.5kb) GFP sequences are marked with blue and red arrows, respectively. Clone siNeg_1 showed a mix of two sequences and was excluded from sequencing analysis. (E) Alignment of the GFP coding sequence of GFP- clones, with clone siXrcc4_7 showing a 1000 bp deletion. Detailed repair junction sequence analysis is shown in Table S3.7. (F) Sequence alignment of 14 analyzed repair products of GFP+ clones, as a comparison to the wild-type GFP sequence. The red and blue lines show upstream and downstream donor-derived PAM mutation, respectively (Donor Mutation 1 and 2). Four clones containing large rearrangements were detected: three GFP+ clones (siXrcc4_4, siXrcc4_5, siLig3_5) deleted the intervening sequence between two direct 215 bp repeats (blue arrows) downstream of the GFP CDS and one clone (siNeg_5) had a 500 bp insertion, which contained fragments derived from the donor plasmid backbone (yellow) and GFP sequence. SD-MMEJ MH1 and MH2 microhomologies bracketing the inserted sequences are represented as orange arrows. For detailed repair junction analysis, see Table S3.7. (G) Flow cytometric analysis of GFP and dsRed fluorescence of CHO-DG44 HR reporter cells treated with the indicated siRNA pools, G3G7 sgRNA, FokI-dCas9 nuclease, donor, and dsRed expression plasmids. All tested siRNA pools were validated previously to reduce target mRNA level by at least 50% in CHO-DG44 cells [141]. Values represent the frequency of GFP+ cells among the transfected dsRed+ cells, expressed relative to siNeg-treated control cells (n $\geq$ 3, error bars represent s.e.m.). P values (* P $<$ 0.05, ** P $<$ 0.01) were calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction.
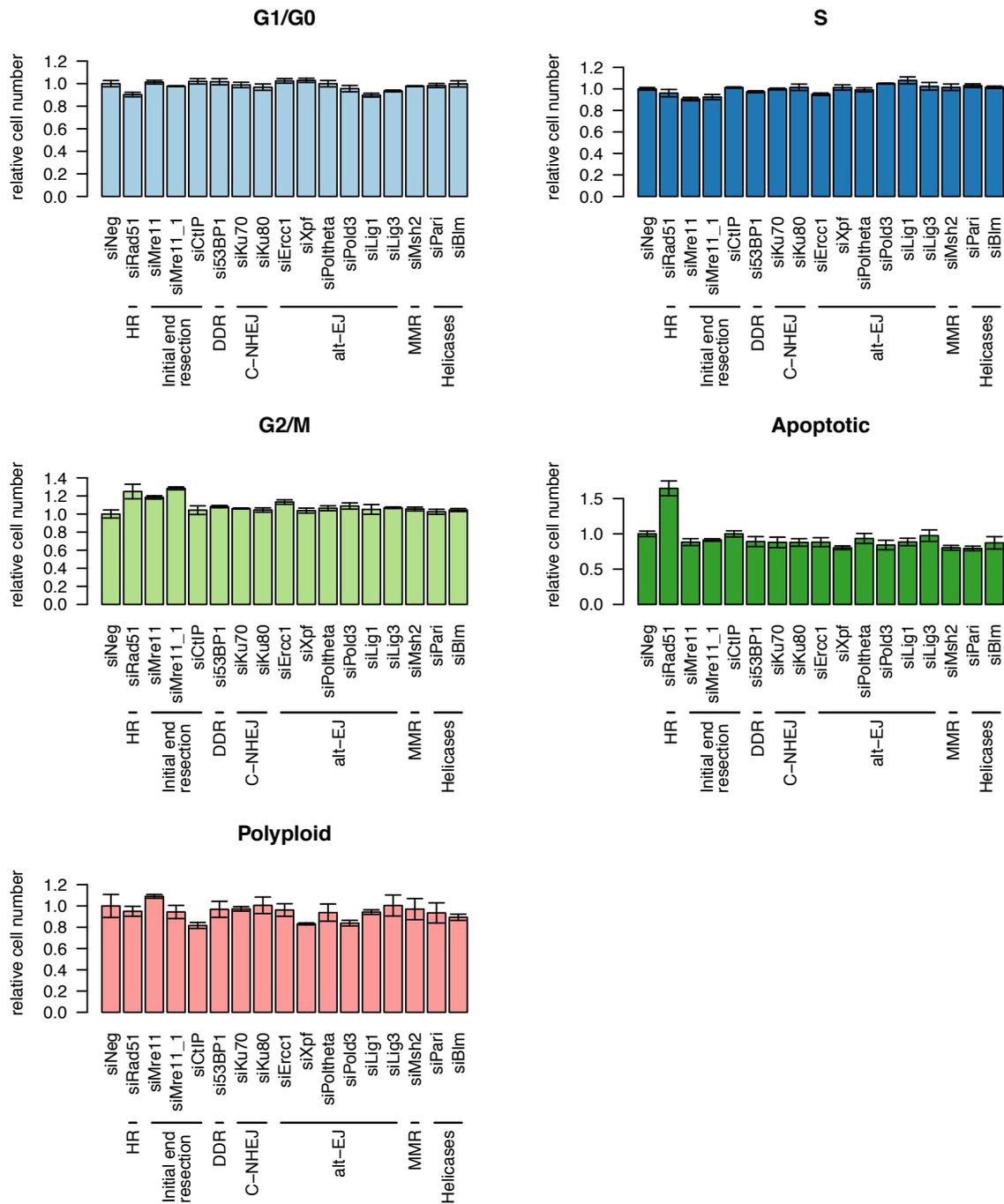
Figure S3.12: Effect of siRNA-mediated knockdowns on cell cycle distribution. Cycle distribution of siRNA-treated CHO-K1 reporter cells was assessed by flow cytometry using the Watson Pragmatic algorithm to determine G1/G0, S, G2/M, apoptotic and polyploid cell cycle phases. Data are normalized to cognate siNeg control values. siMre11 and siMre11_1 indicate transfections using the siRNA pool and the most efficient singular siRNA targeting Mre11, respectively. All other samples were treated with siRNA pools (n = 3, error bars represent s.e.m.).
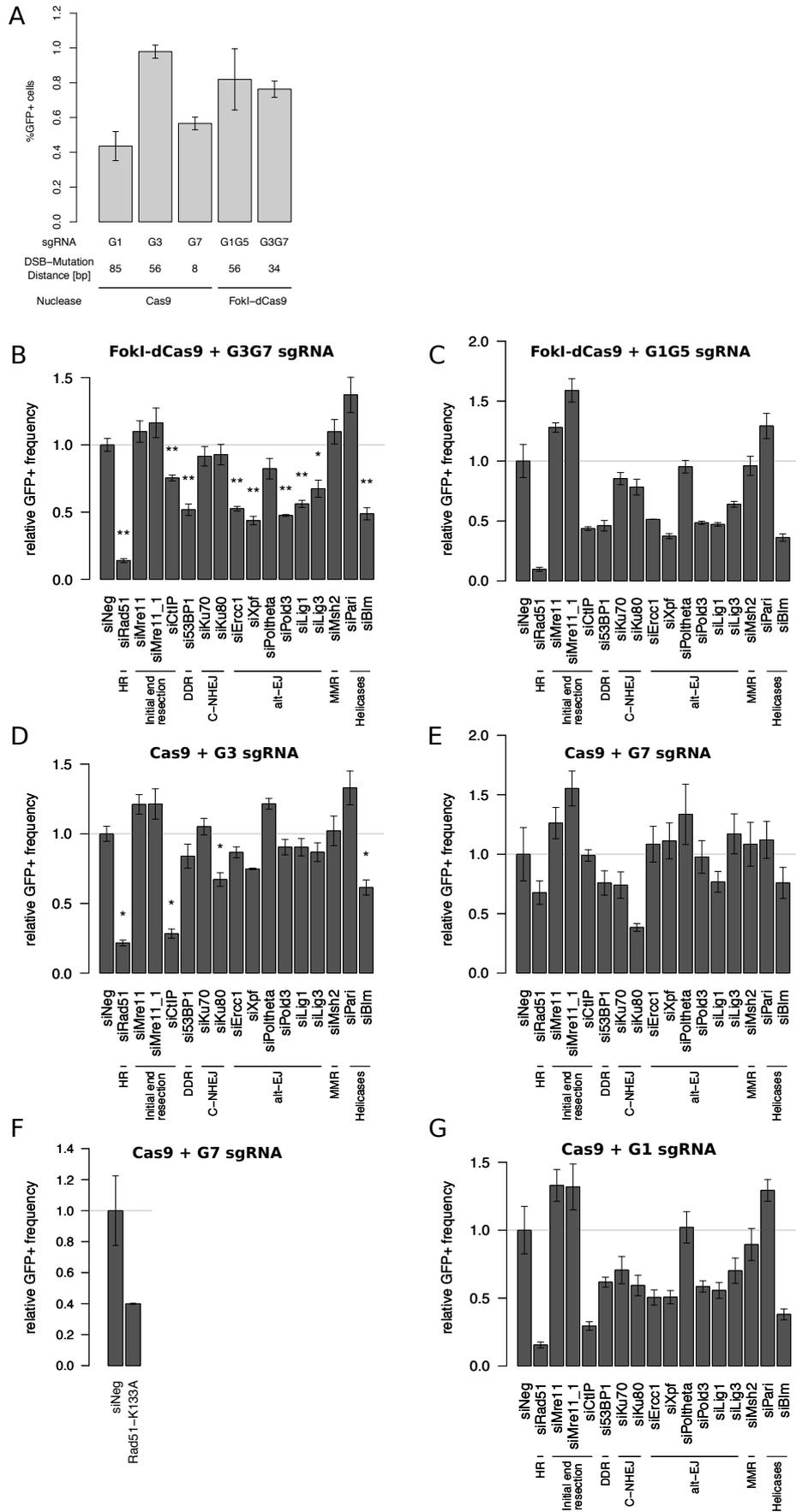
Figure S3.13: Caption on next page.

Figure S3.13: Comparison of different sgRNAs and nuclease systems DNA cleavage products relative to DNA repair knockdowns in the GFP reconstitution assay. (A) CHO-K1 reporter cells were transfected with distinct sgRNA (G1, G3, G7, G1G5, G3G7) and nuclease systems (Cas9, FokI-dCas9) and assessed for GFP reconstitution frequency using flow cytometry. Data show the frequency of GFP+ cells among the entire cell population ($n \geq 2$, error bars represent s.e.m.). (B-E, G) Effect of siRNA knockdown on GFP reconstitution for FokI-dCas9 nuclease using G3G7 sgRNA (B) and G1G5 sgRNA (C), or for wild-type Cas9 nuclease using G3 sgRNA (D), G7 sgRNA (E), and G1 sgRNA (G). (F) Effect of the dominant-negative Rad51-K133A mutant overexpression on GFP reconstitution using G7 sgRNA and wild-type Cas9. Each plot represents the frequency of GFP+ cells among the transfected dsRed+ cells, and values were normalized to the corresponding siNeg control. siMre11 and siMre11_1 samples were transfected using the siRNA pool and the most efficient singular siRNA targeting Mre11, respectively, while the other samples were treated with siRNA pools. Data shown in panel B-E and G were used for heat map and cluster analysis (see Figure 3.4) ($n \geq 2$, error bars represent s.e.m.). P values (* $P < 0.05$, ** $P < 0.01$) were calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction.

Table S3.1: siRNA target sequences.

Target gene[1]	siRNA	Sequence (5' - 3')
	siRif1_1	AGA GUC CAU UGC AUA UAA A
Rif1	siRif1_2	UGA UGA GCA UGC UGA AAC A
	siRif1_3	UGC UAC UAC UGG UGA UUU A
	siRev7_1	AAA CCA UGA CGU CGG AUA U
Rev7	siRev7_2	GAA CGA UGU GGA GAA GGU A
	siRev7_3	GAG CGA GCU CAU AAA CAC A
	siMsh2_1	GAU GCA GUC UAC AUU AAU A
Msh2	siMsh2_2	AAA GGA GAA CAG AUG AAU A
	siMsh2_3	GAA CAG AAU AGA AGA AAG A
	siMlh1_1	ACA CAC ACC CAU UCU UAU A
Mlh1	siMlh1_2	AAA UCC AAG UGA AGA GUA U
	siMlh1_3	GCA GCU GCC UCA AGU CAG A
	siFbxo18_1	AGU ACC AGC UGA AGA AGA A
Fbxo18	siFbxo18_2	GAA GAG GCA UAC CAA AUG A
	siFbxo18_3	CUU CUA CCC UGG UCA AGU A
	siRecql5_1	GAA AGU CUU UGG GUU UGA U
Recql5	siRecql5_2	GGC UGA UAA AGG GCU GUU A
	siRecql5_3	UUG UUG CGA CCA UCA GUU U
	siBlm_1	UAC AGA CUG UGA CGA CAA A
Blm	siBlm_2	AAG CUG ACU UCC UUU GAU A
	siBlm_3	ACA UGC UGC UUC AGU AAA U
	siWrn_1	ACA GAG UUG CAG UGA UUC A
Wrn	siWrn_2	GGC AGG UGU UGG AAU UGA A
	siWrn_3	UUA GCA AAU UUG GGU GAU A
	siRtel_1	GAA GCA CAC UCU AUU UAA A
Rtel	siRtel_2	GGG UUC UGA UGA CUU UGA A
	siRtel_3	AGU GGG AUG UAC AGU AGA A
	siKeap1_1	AGA CUA CCU GGU GCA GAU A
Keap1	siKeap1_2	UGA GUG GCG GAU GAU UAC A
	siKeap1_3	UGG AGG UGG UGU CCA UUG A
	siPari_1	UAA AGC UGG UGG CAA AUU A
Pari	siPari_3	UGU AGC AAC UUG ACU UCU A
	siPari_3	GGA AAC AGG UGG AUU UAG A

[1] siRNA sequences not listed here were published previously (Kostyrko et al., 2017)

Table S3.2: Primer sequences.

End-point PCR		
Experiment	Primer	Sequence (5' - 3')[1]
Cloning GFP gene correction vector	HRassay_F1	CTC ACT GTT CTC TCC CTC CG
	HRassay_R1	CCT TAA GAT CAC CCC AGG ATG TTG CCG T
	HRassay_F2	GGT GAT CTT AAG GAC GGC AGC GTG C
	HRassay_R2	TTC TCC GCT CCA TCG TTC AG
Cloning GFP donor plasmids	G1_F	TAC AAG ACa CGC GCC GAG
	G1_R	GCG CGt GTC TTG TAG TTG C
	G5_F	CTT CAA aGA GGA CGG CAA CA
	G5_R	GTC CTC tTT GAA GTC GAT GCC
	G3_F	GGC GAC ACa CTG GTG AAC
	G3_R	ACC AGt GTG TCG CCC TC
	G7_F	AAC ATC CTc GGG CAC AAG
	G7_R	CTT GTG CCC gAG GAT GTT G
	GFP_F3	CGC TCA CTG TTC TCT CCC TC
	GFP_R4h	TAC TCT GGC TAT GAA CCC TAC C
	GFP_Amut_KpnI_F	TAA TGG TAC CTG GTG AGC AAG GGC GAG
	GFP_Amut_R	TCT CCG CTC CAT CGT TCA GA
Analysis GFP repair junctions	GFP_Junction_F	CGT TCC CAA AGT CCT CCT GT
	GFP_Junction_R1	AAT CCG TCG CTG TGC ATT AG G
	GFP_Junction_R2	GTG AGG CGT GCT TGT CAA TG
Cloning overexpression plasmids	Rad51_NcoI_F	TAA GCA CCA TGG CTA TGC AGA TGC AGC TT
	Rad51_XbaI_R	TAC GGT TCT AGA GAA GAA GCC CAG AGA GCA GTC
	Ercc1_NcoI_F	TAT GCA CCA TGG ACC TTG GGA AAG ACG AG
	Ercc1_XbaI_R	TAC GCA TCT AGA GCA GCC TGG AAA ATG CTT TAT
	Pold3_BsaI_F	TAA GCA GGT CTC CCA TGG CGG AAC AGC TGT ATC
	Pold3_SpeI_R	TAA GCA ACT AGT GTC TTC TCT GTG ACC ACT CCA
	Lig3_BspHI_F	TAG GCA TCA TGA CTT TGG CTT TCA AGA TCC T
	Lig3_XbaI_R	TAA GCA TCT AGA GAA CGT GAG GAC AGC TCA AAG
	Rad51_K133A_F	CTG GGg ccA CAC AGA TCT GTC ATA CAT T
	Rad51_K133A_R	TGT GTg gcC CCA GTT CGG AAT TCT CC
	Rad51_K133R_F	CTG GGA gGA CAC AGA TCT GTC ATA CAT T
	Rad51_K133R_R	TGT GTC cTC CCA GTT CGG AAT TCT
	Neomycin_HindIII_F	TAGGTCAAGCTTACCGCCACCATGATTGAACAAGATGGA
	Neomycin_XbaI_R	TAAGCGTCTAGATCAGAAGAACTCGTCAA
	Ku70_PciI_F	AAA CCA ACA TGT CAG GGT GGG AAT C
	Ku70_SpeI_R	TTT ACT AGT CAG TTC TTC TGG AAG TG
	Ku80_BspMI_F	TAA GAC ACC TGC GCA ACA TGG CGT GGT CCG CTA
	Ku80_XbaI_R	TAC GGT TCT AGA TGG GGT GAT TTG CTC CTG AGC
qPCR and RT-qPCR		
Experiment	Primer	Sequence (5' - 3')
HR reporter copy number[2]	Puromycin_F	CCT CTG AGC TAT TCC AGA AGT
	Puromycin_R	GCT TGT ACT CGG TCA TGG G
	B2M_F	ACC ACT CTG AAG GAG CCC A
	B2M_R	TTG ACA CAG ATA GAG CTT CC

90

Donor plasmid integration[2]	GFP_Insert_F	AGA TCC GCC ACA ACA TCG AG
	GFP_Insert_R	GGG TGC TCA GGT AGT GGT TG
	Donor_Backbone_F	TCG GGG CTG GCT AAC TA TG
	Donor_Backbone_R	GCC TGA TGC GGT ATT TTC TCC
Unrepaired GFP deletion[2]	GFP_Deletion_F	GGT GAT CTT AAG GAC GGC AG
	GFP_Deletion_R	CGC TTC TCG TTG GGG TCT TT
siRNA knockdown and protein overexpression	B2M_RTqPCR_F	GAG TTT ACA CCC ACT GCG AC
	B2M_RTqPCR_R	CAT GTC TCG TTC CCA GGT GA
	53BP1_F	TCT CGA CAC CTT CCC AAT CT
	53BP1_R	GCT ATT ATT GTC TCC TTG CTC
	Rif1_F	GCC CCA CTT TAA CTG AAA ACC C
	Rif1_R	CAG TTT TCC ACG GCA CCT TG
	Rev7_R	GCT TCA CTG CGT CAA ACC AC
	Rev7_F	GAC AGC AGG GAG TCG AAA C
	Ku70_F	TCC AGC TTG TCT TCC TCC
	Ku70_R	TTC GCT TTA CAT ACA GGA GTG
	Ku80_F	CCC TGA TTG TGT GCA TGG AT
	Ku80_R	GTT GAT CTT GGC TGA ATG GG
	DNA-PKcs_F	CAT GAA TCA CAT TTC CTC CAG
	DNA-PKcs_R	GTC ATC AGC AGA TAA TCC CAA
	Lig4_F	GCA GAA AAG ATC AGG CAC TTT
	Lig4_R	TGC TGG GTA AAA AGA GTC TGT
	Xrcc4_F	TTC TTG GGA GGA AAC AGT GG
	Xrcc4_R	CCA TAG CCA TGT CAT CAG CT
	Rad51_F	CGG TTA GAG CAA TGT GGC A
	Rad51_R	ACG GTG TAA CGA GAT TGG C
	Mre11_F	TTC AGG CAT AGG GAG CAA A
	Mre11_R	CTC TGA GTG TCG TTC CTT CC
	Rad50_F	GGG ACA AGA AGA GTG GAC TGG
	Rad50_R	GCT CGT TCT TCA CAT TCC TCA
	Nbs1_F	ATC AGC AA TGT CCT CAT CC
	Nbs1_R	AGA TGG ATG CTC CTT GCT TT
	CtIP_F	TGA GCT TGC ATC AGT TCT TCA
	CtIP_R	GTC AAA GGG CAC ATC TTG G
	Parp1_F	AGA AGG GAA AGG ACA AGG AC
	Parp1_R	CAC TGG TGG AAC ATG CTT TC
	Ercc1_F	AGA CGG TGA AAA CGG GAG C
	Ercc1_R	GTC CCA GCA CAT AGT CAG G
	Xpf_F	ACA AGG TGG TAT TAT ATT CGC
	Xpf_R	ATG ATC CTG TGT GCT CTG TA
	Poltheta_F	ACT TCT ATG AAA AGG GTG GCA
	Poltheta_R	CCA TTT GTC GAT GCT GTC TG
	Pold3_F	ACC CAA GAC ACC AAC AAG GA
	Pold3_R	TCA CAT TCC CCT TTC CTG GT
	Lig1_F	GTG AAA CAA GAA GAG CCA GG
	Lig1_R	GGC ATC TTC AAT GGG GTG G
	Lig3_F	GCT ACA ATA CAA AGA CCC AGA

	Lig3_R	AAC ACT CTT AAC GAC TCC CG
	Xrcc1_F	GGG ACG AGA GGA GGA AGC
	Xrcc1_R	CCT GGG CTG TGA TAA CGA AT
	Rad52_F	GCC AGA AGG TGT GTT ATA TTG
	Rad52_R	ATT CTG CTG GGT GAT GGA GT
	Msh2_F	CAA AAC TGC TCA GGG ACA AAG
	Msh2_R	CCT CAA CAA AAG TCT CTA CTA
	Mlh1_F	AGG ACG TGG AGA TGT TGG AG
	Mlh1_R	CTT TCT GGG GTT GCT TGG AC
	Fbxo18_F	CAG CTA ACA CAT GAA CAA CAG C
	Fbxo18_R	AGA AAC CTG CTC TGA GAC CAC
	Recql5_F	CCT GAA GGC TCT TGG ACA G
	Recql5_R	CCT TTA GCC CTG CGT GGT AA
	Blm_F	GAA ATG CCT GCA AAG GAG CC
	Blm_R	CCA TCA CAC GAA TAA GCT GCT G
	Wrn_F	TGT GTA TGT CTG AGA GCA AAT GT
	Wrn_R	CAA AAT CCC GTA GAA GTT TCC AC
	Rtel_F	GGC TTC TCT TAC CTG CCT CTT
	Rtel_R	TCT TCA AAC TGC TGC TTG TGG
	Keap1_F	CGA TGG CAC TAA CCG ACT GA
	Keap1_R	GTA CAG CAC ACA GAC CCC AG
	Pari_F	TGA AAG TGT GAA AGT GGT TGA CC
	Pari_R	ACA GAA AAT CCC GTA GTT GGC T
	Neomycin_F	GAC CAC CAA GCG AAA CAT CG
	Neomycin_R	CCC TGA TGC TCT TCG TCC AG
	FokI-dCas9_F	AAT TCC GTT GGA TGG GCT GT
	FokI-dCas9_R	TCT GCC GTT TCG CCA CTA TC

Table S3.3: Plasmids used in this study.

Plasmid name	Description	Reference
wild-type Cas9 [Addgene #43861]	Expresses mammalian codon optimized Cas9 nuclease	(Fu et al., 2013)
FokI-dCas9 [Addgene #52970]	Expresses Fok1 nuclease domain fused to catalytically inactive Cas9 DNA-binding domain in mammalian cells	(Guilinger et al., 2014)
JG phU6-G1-sgRNA	single G1 sgRNA expression vector targeting GFP sequence	(Guilinger et al., 2014)
JG phU6-G3-sgRNA	single G3 sgRNA expression vector targeting GFP sequence	(Guilinger et al., 2014)
JG phU6-G7-sgRNA	single G7 sgRNA expression vector targeting GFP sequence	(Guilinger et al., 2014)
JG pUC19 double-hU6-sgRNA-G1&G5	double G1&G5 sgRNA expression vector targeting GFP sequence	(Guilinger et al., 2014)
JG pUC19 double-hU6-sgRNA-G3&G7	double G3&G7 sgRNA expression vector targeting GFP sequence	(Guilinger et al., 2014)
pCMV-DsRed-Express	dsRed expressing transfection control plasmid	Clonetech
pCS2+U5V5PBU3	*PiggyBac* transposase expression vector	(Ley et al., 2013)
pITR_SV40Puro-pGAPDHGFP_ITR	*PiggyBac* transposon containing a Puromycin and GFP gene cassettes	(Ley et al., 2013)
pHR-GFP-101del	HR GFP assay containing a GFP loss-of-function 101 bp deletion	This study
pUC19-G1G5-Amut	Donor plasmid containing two silent mutations in the PAM site for G1&G5 sgRNA	This study
pUC19-G3G7-Amut	Donor plasmid containing two silent mutations in the PAM site for G3&G7 sgRNA	This study
pGapdh-GFP	Overexpression cloning plasmid containing a strong GAPDH promoter upstream of the GFP sequence	(Le Fourn et al., 2014)
pGapdh-emptyVector	Overexpression control plasmid; devoid of GFP	This study
pGapdh-CgRad51, pGapdh-CgErcc1, pGapdh-CgPold3, pGapdh-CgLig3, pGapdh-Neomycin, pGapdh-CgKu70, pGapdh-CgKu80	Overexpression plasmids; GFP sequence of pGadph-GFP was replaced with Cricetulus griseus (Cg) Rad51, Ercc1, Pold3, Lig3, Neomycin, Ku70 or Ku80	This study
pGapdh-CgRad51-K133A	Overexpression plasmid expressing the dominant-negative CgRad51-K133A mutant deficient in ATP binding	This study
pGapdh-CgRad51-K133R	Overexpression plasmid expressing the dominant-negative CgRad51-K133R mutant deficient in ATP hydrolysis	This study

Table S3.4: Sequence analysis of mutated CHO-K1 HR reporter junction amplicons cloned into plasmid vectors.

Sample[1]		Sequence[2]	Features[3]	Repair pathway[4]	Second mutation	Maximal distance mutation to DSB[5]
Small mutations[6]						
siNeg_Red_6 GFP-	Genomic	5'-CGAGGGGCGACA**CCC**TGGTGAACCGCATCGAGCTGAAG**GGCCATC**G ACTTCAAG**GA**G**GA**CGGCAACATCC**TGG**GTGATCTTAAGGACGGCAG CGTGC-3'	17 bp deletion, 2 bp or 6 bp MH (26 bp downstream)	MMEJ or SD-MMEJ (loop-out)	no	<100
	Junction	5'-CGAGGGGCGACA**CCC**TGGTGAACCGCATCGAGCTGAAG**GG**----- ---------ACGGCAACATCC**TGG**GTGATCTTAAGGACGGCAG CGTGC-3'				
siNeg_Red_9 GFP-	Genomic	5'-AGTTCGAGGGCGACA**CCC**TGGTGAACCGCATCGAGCTGAAGGG**C** **ATC**GACTTCAAGGAGGACGGCAACATCC**TGG**GTGATCTTAAGGACG GCAGCGTGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGC GACGGCCCCGTGCTGCTGCCCGACAACCACTACCT-3'	42 bp deletion, 5 bp MH (87 bp downstream)	SD-MMEJ (snap-back)	no	<100
	Junction	5'-AGTTCGAGGGCGACA**C**------------------------------ ---------GGACGGCAACATCC**TGG**GTGATCTTAAGGACG GCAGCGTGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGC GACGGCCCCGTGCTGCTGCCCGACAACCACTACCT-3'				
siNeg_noDonor _Red_2 GFP-	Genomic	5'-GCGACA**CCC**TGGTGAACCGCATCGAGC**CT**GAAGGG**CATC**GAC**CT**TC AAGGAGGACGGCAAC-3'	12 bp deletion, 2 bp MH flanking DSB	unknown[7]	no	<100
	Junction	5'-GCGACA**CCC**TGGTGAACCGCATCGAGC**CT**------------**CT**TC AAGGAGGACGGCAAC-3'				
siNeg_noDonor _Red_3 GFP-	Genomic	5'-GCGACA**CCC**TGGTGAACCGCGA**T**CGAGCTGAAGGG**CATC**GAC**T**TC AAGGAGGACGGCAAC-3'	19 bp deletion, 1 bp MH flanking DSB	unknown[7]	no	<100
	Junction	5'-GCGACA**CCC**TGGTGAACCGCA**T**------------------**T**TC AAGGAGGACGGCAAC-3'				

Table S3.4: Continued

		Sequence	Deletion/MH	Mechanism		
siNeg_noDonor_Red_4	GFP-	Genomic	5'-GCGACACCCTGGTGAACCGCCATCGAGCTGAAGGGCATCGACTTC AAGGAGGACGGCAACATCCTGGGTGATCTTAAGGACGGCAGCGTGC AGCTCGCCGACCA-3'	14 bp deletion, 2 bp or 5 bp MH (51 bp downstream)	MMEJ or SD_MMEJ (snap-back)	no	<100
		Junction	5'-GCGACACCCTGGTGAACCGCCATCGAG-----------CTTC AAGGAGGACGGCAACATCCTGGGTGATCTTAAGGACGGCAGCGTGC AGCTCGCCGACCA-3'				
siNeg_GFP_10	GFP+	Genomic	5'-AAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAA CCGCCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCC TGGGGTGATCTTTA-3'	21 bp deletion, 2 bp or 6 bp MH (26 bp upstream)	MMEJ or SD-MMEJ (loop-out)	no	<100
		Junction	5'-AAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAA CCGCCATCG------------AGGAGGACGGCAACATCC TGGGGTGATCTTTA-3'				
siNeg_GFP_exp2_16	GFP+	Genomic	5'-AGTTCGAGGGCGACACCCTGGTGAACCGCCATCGAGCTGAAGGGC ATCGACTTCAAGGAGGAGGACGGCAACATCCTGGGTGATCTTAAGGTGA TCTTAAGGACGGCAGCGTGCAGCTCGCCGACCACTACC-3'	15 bp deletion, 5 bp MH (60 bp downstream)	SD-MMEJ (loop-out)	yes	<100
		Junction	5'-AGTTCGAGGGCGACACCCTGGTGAACCGCCATCGAGC----------------TCAAGGAGGAGGACGGCAACATCCTGGGTGATCTTAAGGTGA TCTTAAGGACGGCAGCGTGCAGCTCGCCGACCACTACC-3'				
Large mutations[8]							
siNeg_Red_4	GFP-	Genomic	5'-TGAAGGGCGGCAGGGCGGGGGCGGCCGCAGGCCGGATGTGTTCGC-580 bp-CATCGACTTCAAGGAGGACGGCAACATCCTGGGTGATCTTAA GGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAG-3'	594 bp deletion, 3 bp or 4 bp MH (15 bp downstream)	MMEJ or SD-MMEJ (snap-back)	yes	585
		Junction	5'-TGAAGGGCGGCAGGGCGGGGGCGGCCGCAGGCCGGATGTG-------580 bp----------TTCAAGGAGGACGGCAACATCCTGGGTGATCTTAA GGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAG-3'				

95

Table S3.4: Continued

				Mutation	Mechanism		
siNeg_Red_18	GFP-	Genomic	5'-GGGCGGACACCCTGGTGAACCGCCATCGAGCTGAAGGGCATCGACTTC-301bp-GGCATGGACGAGCTGTACAAGTAAAGCGGCCGC-3'	333 bp deletion, 7 bp MH	MMEJ	yes	326
		Junction	5'-GGGCGGACACCCTGGTGAACCGCCATCGAGCTG------------301bp-------------TACAAGTAAAGCGGCCGC-3'				
siNeg_GFP_exp2_16[9]	GFP+	Genomic	5'-CGCCGCCGGGATCACTCTCGGCCATGGACGAGAGCTGT-658bp-G CAATTGTTGTTGTTAACTTGTTTATTGCAGCTTATAATGGTTACAA ATAAAGCAATAGCATCACAAATTTCAC-3'	678 bp deletion, 2 bp MH or 4 bp MH (9 bp downstream)	MMEJ or SD-MMEJ (loop-out)	yes	1061
		Junction	5'-CGGCCGCCGCTTCGAGCAGACATG----------658bp-- -------TTGTTAACTTGTTTATTGCAGCTTATAATGGTTACAA ATAAAGCAATAGCATCACAAATTTCAC-3'				
siPari_GFP_20[9]	GFP+	Genomic	5'-AAAGTCCTCCTGTTTCATCCAAGC-347bp-CGCTCTCTGCTC CTCCTGTTCGACAGTCAGC-3'	367 bp deletion, 10 bp MH[10]	MMEJ	yes	1082
		Junction	5'---C CTCCTGTTCGACAGTCAGC-3'				

[1] The table shows Sanger sequencing data of repair junctions of CHO-K1 HR reporter cells treated with G3G7 sgRNA and FokI-dCas9 nuclease that contain small and large CRISPR-derived mutations. The sequences are derived from Sanger sequencing of amplicons cloned into plasmid vectors.
[2] Predicted 4 nt 5' overhanging DSB sites induced by the G3G7 sgRNA and FokI-dCas9 nuclease are highlighted in yellow, PAM site are shown in blue. Pre-existing microhomologies (MH) of the microhomology-mediated end-joining (MMEJ) repair mechanism are shown in green, while de novo MH of the synthesis-dependent microhomology-mediated end-joining (SD-MMEJ) mechanism are underlined with a double line. Deleted bases are depicted with a "-" sign.
[3] Size of mutation and MH length (in bp). The distance between priming site and the break site for de novo MH are shown in parenthesis.
[4] Most probable DSB repair mechanism based on manual junction analysis. Possible repair mechanisms include C-NHEJ, MMEJ, SD-MMEJ (snap-back), SD-MMEJ (loop-out), single strand annealing (SSA) and unknown. For snap-back SD-MMEJ mechanism, de novo priming sites are inverted repeats, while loop-out SD-MMEJ mechanisms uses priming sites with direct repeats (Khodaverdian et al., 2017). If junction sequence is compatible with more than one mechanism and both are equally likely, all potential pathways are listed.
[5] Depicted is the maximal distance (in bp) of the left or right mutation border to the FokI-dCas9 DSB site.
[6] The six junctions containing small mutations are also aligned in Supplementary Figure S6. Junctions were verified for homologies at break site and templated insertions (SD-MMEJ/TMEJ) using program described in (Schimmel et al., 2017).
[7] Unknown mechanism but similar junction pattern described in (Shin et al., 2017).

Table S3.4: Continued

[8] The four junctions containing large mutations are omitted in Supplementary Figure S6 for simplicity.
[9] Site of mutation is >50 bp apart from CRISPR/Cas9 target site. Mutation occurs outside of the GFP CDS and should not influence GFP expression.
[10] Junction sequence contains only 9 bp MH as start of the mutation corresponds to Sanger sequencing start site.

Table S3.5: Estimated coefficients of non-linear model testing the relationship between the amount of protein and the relative HR frequency for HR rate-limiting factors.

Estimated Coefficient[1]	Point Estimate	Standard Error	P-value (H0: $\beta = 0$)	Significance Level[2]
$\hat{\beta}_0$	1.085306	0.041569	< 2e-16	***
$\hat{\beta}_1$	0.091292	0.028334	0.00163	**
$\hat{\beta}_2$	-0.012825	0.002821	1.29e-05	***

[1] Estimated coefficients are derived from the non-linear model $relativeHR_i = \beta_0 + \beta_1 * \mu gProtein_i + \beta_2 * \mu gProtein_i^2 + \varepsilon_i$ using the Rad51, Ercc1, Lig3, and Pold3 overexpression data shown in Supplementary Figure S10A with a total of 126 observations. The F-Test for overall significance gives a p-value of 2.444e-07.
[2] *** p-value < 0.001; ** p-value < 0.01; * p-value < 0.05

Table S3.6: Estimated coefficients of non-linear model testing the relationship between the amount of protein and the relative dsRed+ frequency for rate-limiting HR factors.

Estimated Coefficient[1]	Point Estimate	Standard Error	P-value (H0: $\beta = 0$)	Significance Level[2]
$\hat{\beta}_0$	0.9429529	0.0389288	<2e-16	***
$\hat{\beta}_1$	-0.0174230	0.0265347	0.513	
$\hat{\beta}_2$	-0.0007418	0.0026416	0.779	

[1] Estimated coefficients are derived from the non-linear model $relativeTransfectionEfficiency_i = \beta_0 + \beta_1 * \mu gProtein_i + \beta_2 * \mu gProtein_i^2 + \varepsilon_i$ using the Rad51, Ercc1, Lig3, and Pold3 overexpression data shown in Supplementary Figure S10B with a total of 126 observations. The F-Test for overall significance gives a p-value of 0.009807.
[2] *** p-value < 0.001; ** p-value < 0.01; * p-value < 0.05

Table S3.7: Sequence analysis of mutated CHO-DG44 HR reporter junctions from isolated cell clones.

Sample[1]			Sequence[2]	Features[3]	Repair pathway[4]	Second mutation	Maximal distance mutation to DSB[5]
Small mutations[6]							
siRad51_7	GFP-	Genomic	5'-GAGGTGAAGTTCGAGGGCGACACCCCTGGTGAACCGCATCGAGCTGAAGG— -------------GCATCGACTTCAAGGAGGACGGC-3'	1 bp deletion, 25 bp templated insertion, 5 bp and 7 bp MH	SD-MMEJ (loop-out)	no	<100
		Junction	5'-GAGGTGAAGTTCGAGGGCGACACCCCTGGTGAACCGCATCGAGCTGAAGTT CGAGGGCGACACCCCTGGTGAACCGCATCGACTTCAAGGAGGACGGC-3'				
siLig3_8	GFP-	Genomic	5'-ACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGG CAACATCCTGGG-3'	15 bp deletion, 7 bp MH	MMEJ	no	<100
		Junction	5'-ACCCTGGTGAACCGCATCGA-------------CTTCAAGGAGGACGG CAACATCCTGGG-3'				
siXrcc4_5	GFP+	Genomic	5'-ACCCTGGTGAACCGCATCGAGCTGAGCTGAAGGGCATCGACTTCAAGGAGGACGG CAACATCCTGGG-3'	15 bp deletion, 4 bp MH	MMEJ	yes	<100
		Junction	5'-ACCCTGGTGAACCGCATCGAGCTG-------------AAGGAGGACGG CAACATCCTGGG-3'				
siXrcc4_6	GFP+	Genomic	5'-GAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGG CATCGACTTCAAGGAGGACGGCAACATCCTGGGCACAAGCTGGAGTACAAACT ACAACAGCCACAACGTCTATATCATG-3'	40 bp deletion	Unknown	no	<100
		Junction	5'-GAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCG------------- -------------ATCCTGGGCACAAGCTGGAGTACAAACT ACAACAGCCACAACGTCTATATCATG-3'				

		Mechanism		Size
siXrcc4_7	siNeg_7 / siXrcc4_7 / wild-type GFP	or 4 bp MH (18 bp downstream)	SD-MMEJ (loop-out)	no — 870
siNeg_5 [8]	siNeg_7 / siXrcc4_7 / siXrcc4_8 / siRad51_7 / siRad51_8 / siLig3_7 / siLig3_8	528 bp insertion, of which 525 bp G3G7 plasmid knock-in, 6 bp MH	SD-MMEJ (fold-back)	no — 528
siXrcc4_4 / siXrcc4_5 / siLig3_5		800 bp deletion, 215 bp direct repeat	SSA	yes — 1208

Genomic

Junction

[1] The table shows Sanger sequencing data of repair junctions of cells treated with G3G7 sgRNA and FokI-dCas9 nuclease that contain small and large CRISPR-deletions. Sanger sequencing of single cell clones.

[2] Predicted 4 nt 5' overhanging DSB sites induced by the FokI-dCas9 nuclease are highlighted in yellow, PAM site are shown in blue. Pre-existing microhomologies (MH) of the microhomology-mediated end-joining mechanism are shown in green, while de novo MH of the synthesis-dependent microhomology-mediated end-joining mechanism are coupled with a double line. Inserted bases are represented in red letters, deleted bases with a "-" sign. The distance between priming site and the break site for de novo MH are shown in parenthesis.

[3] Size of mutation and MH length (in bp).

[4] Most probable DSB repair mechanism based on manual junction analysis. Possible repair mechanisms include C-NHEJ, MMEJ, SD-MMEJ (snap-back), SD-MMEJ (loop-out), single strand annealing (SSA) and unknown. For snap-back SD-MMEJ mechanism, de novo priming sites are inverted repeats, while loop-out SD-MMEJ mechanisms uses priming sites with direct repeats (Khodaverdian et al., 2017). If junction sequence is compatible with more than one mechanism and both are equally likely, all potential pathways are listed.

[5] Depicted is the maximal distance (in bp) of the left or right mutation border to the FokI-dCas9 DSB site.

Table S3.7: Continued

[6] The four junctions containing small mutations are also aligned in Supplementary Figure S11E and S11F. Junctions were verified for homologies at break site and templated insertions (SD-MMEJ/TMEJ) using program described in (Schimmel et al., 2017).

[7] The five junctions containing large mutations are also aligned in Supplementary Figures S11E and S11F.

[8] Site of mutation is >50 bp apart from CRISPR/Cas9 target site. SSA-mediated mutations occur outside of the GFP CDS and should not influence GFP expression.

Table S3.8: Overview of sgRNAs and nucleases used in this study

Cell Line	CHO-K1					DG44
Mutation type	Insertion of 2 in-frame stop codons separated by 3 base pairs (9 bp), 101 bp deletion					Deletion
Mutation size [bp]	110 (101+9)					215
sgRNA[1]	G1	G3	G7	G1G5	G3G7	G3G7
Targeted DNA strand	non-template	non-template	template	both	both	both
Nuclease	wild-type Cas9	wild-type Cas9	wild-type Cas9	FokI-dCas9	FokI-dCas9	FokI-dCas9
Most frequent DSB structure	blunt[2]	blunt[2]	blunt[2]	4 nt 5' overhang	4 nt 5' overhang	4 nt 5' overhang
DSB – Mutation Distance [bp]	85	56	8	56	34	270
Protruding flap length [nt]	94[3]	65[3]	17[3]	65[3]	43[3]	270
Maximal homology length to donor plasmid [upstream|downstream; in bp]	331|1252	360|1252	408|1252	359|1252	382|1252	382|494
Nuclease cleavage activity [all cells][4]	46	60	22	5	9	9
Nuclease cleavage activity [transfected cells][4]	96	99	84	55	66	66
HR repair efficiency [%GFP][5]	0.44	0.98	0.57	0.82	0.76	0.29
HR dependency on alt-EJ activity	**Strong**	**Weak**	**Weak**	**Strong**	**Strong**	**Strong**

[1] Detailed DNA sequences and location are shown in Supplementary Figures S1A
[2] Wild-type Cas9 occasionally also leads to 1 nt 5' overhanging DSBs (Lemos et al., 2018)
[3] DSB - Mutation distance plus additional 9 base pairs coming from the insertion of two in-frame stop codons separated by 3 base pairs
[4] Assessed in CHO-DG44 cells (Supplementary Figure S1C and D)
[5] Data are shown in Supplementary Figures S13A

Chapter 4

Characterization and inactivation of endogenous retroviruses in Chinese hamster ovary cells

This chapter is based on a manuscript in preparation entitled "Characterization and inactivation of endogenous retroviruses in Chinese hamster ovary cells" by Duroy P.O.*, **Bosshard S.***, Schmid-Siegert, E., Neuenschwander, S, Arib G., Lemercier P., Masternak J., Roesch L., Buron F., Girod P.A., Xenarios, I., Mermod, N. * equal contribution

I contributed to the genome editing design, generated the CRISPR-Cas9-treated CHO-K1 cells, coordinated the development of the bioinformatic pipelines, prepared the samples for next-generation sequencing and analyzed the ERV mutations (Figs. 4.3, 4.4, 4.6, S4.1 and S4.3-S4.6 as well as Tables 4.2 and S4.1-S4.8). I also prepared the figures and tables and wrote the first version of the manuscript.

4.1 Abstract

Chinese hamster ovary (CHO) cells are the most commonly used protein production cell host for biopharmaceuticals. These cells are known to have budding type-C endogenous retroviruses (ERVs) embedded in their genome and to release retroviral-like particles in the culture supernatant. Although evidence for the infectivity of these particles is missing, their presence has raised safety concerns. As the genomic origin of these particles remains unclear, we systematically characterized the type-C ERV elements at the genome, transcriptome and viral particle level. We identified 173 type-C ERV sequences that clustered into two functionally conserved groups. Interestingly, transcripts from only one type-C ERV group were full-length with intact open reading frames, and corresponding viral RNA genomes were loaded into retroviral-like particles, suggesting that this group may produce functional viruses. Sequence analysis of the genomic RNA from viral particles indicated that they may result from few expressed ERV sequences. Using CRISPR-Cas9 genome editing, we disrupted the *gag* gene of the expressed ERV group. Comparison of CRISPR-derived mutations at the DNA and mRNA level led to the identification of a single ERV locus responsible for the release of viral RNA-loaded particles from CHO cells. Clones bearing a Gag loss-of-function mutation in this particular ERV locus showed a reduction of viral RNA-containing particles in the cell supernatant by over 250-fold. Notably, ERV mutagenesis did not compromise cell growth, cell size or recombinant protein production. Overall, our study highlights a new strategy to mitigate potential contaminations from CHO endogenous retroviruses during biopharmaceutical manufacturing.

4.2 Introduction

Contamination of biopharmaceutical products by adventitious agents such as viruses can interrupt drug supply and thereby imperil patient safety. Although viral contaminations of biopharmaceuticals are rare, they still occur [334], and mitigating the risk of viral contaminations in therapeutic protein preparations remains a top priority.

Chinese hamster ovary (CHO) cells are the most widely used mammalian expression system for biopharmaceutical products. Among others, CHO cells became a preferred production host due to their superior safety profile compared to other cell lines used for recombinant protein production. For instance, it was shown that CHO cells possess reduced susceptibility to certain viral infections [334]. This includes resistance to infections elicited by many human as well as murine retroviruses, with some of the latter being known to infect other mammalian cells [332, 333]. In addition, CHO cells, unlike other rodent cells, appeared to be unable to produce infective retroviruses that could replicate in mammalian cells, notably in human cells [333, 354, 386, 387]. However, viral-like particles (VLPs) have been detected both within CHO cells as well as budding off in the culture medium [388–392]. The presence of such VLPs raises safety and regulatory concerns, not only because there is a remaining risk of a possible hamster to human ERV transmission, but also because they interfere with and reduce the sensitivity of the detection of other adventitious agents.

These VLPs were detected independently by several laboratories, suggesting that they result from endogenous retroviruses (ERVs) that stably integrated into the CHO genome, rather than from an exogenous infection [352]. CHO cells possess two classes of ERVs: the intracisternal type-A ERVs (IAP), a defective ERV class forming immature particles in the cisternae of the endoplasmic reticulum [393], and the budding type-C ERVs [352, 354]. Although type-C ERV sequences remain incompletely characterized, a previous study estimated that approximately 100-300 type-C ERV sequences may be present in the CHO genome [354]. Some of them seemed to be full-length and actively transcribed proviruses, such as the ML2G retrovirus [352, 388]. However, the ML2G ERV sequences described by Lie et al., contain frameshift mutations in each of its gene (Gag, Pol and Env), indicating that the specific ERV sequence at this locus is not producing any VLP [352]. Nevertheless, this publication indicated that other members of this type of ERV sequence are transcribed and produce VLP. The ML2G transcript shares approximately

64% sequence identity to the murine leukemia virus (MLV) family.

As of today, CHO cells are believed to produce non-infective retroviral particles, as their infectivity could not be demonstrated. Nevertheless, the risk that at least one of the uncountable predicted type-C ERV proviruses in the CHO genome is or becomes infective cannot be excluded. This may happen if epigenetically silenced ERVs become expressed, as was observed upon some chemical treatments [355], if dysfunctional ERVs may acquire gain-of-function mutations or if ERVs recombine or trans complement each other. Such genetic changes are more likely to occur in immortalized cell lines, such as CHO cells, which have an overall increased genetic instability [368]. Notably, the close similarity of CHO type-C ERVs to the MLV family, a retrovirus family known to cross the species barrier and to infect even primate cells [394], further indicates that CHO particles may have the potential to become pathogenic for humans, as seen for other retroviruses [395]. Hence strategies to avoid viral contaminations originating from CHO endogenous sources are highly desirable.

The most promising strategy to efficiently prevent hamster ERV transmission is to inactivate retroviruses using CRISPR-Cas9-mediated mutagenesis. The programmable CRISPR-Cas9 RNA-guided nuclease system has already been employed to introduce DNA double strand breaks (DSBs) into proviral sequences in human and porcine cells [325, 396]. Imprecise DSB repair may lead to insertions and deletions within the viral sequences and inhibit viral activity. In a seminal paper, Yang et al. demonstrated that the CRISPR-Cas9 technology can be used to knock-out all 62 genomic porcine ERV sequences resulting in a more than 1000-fold reduction of ERV infectivity [325]. Although successful, viral inactivation remains technically challenging, due to high cytotoxicity, frequent genomic rearrangements and low editing efficiency [325, 326]. One explanation for the reduced editing efficiency of multi-loci sites compared to conventional editing of single genes might be the sheer number of ERV-like sequences that could serve as repair templates for precise, mutation-free repair, so antagonizing ERV mutagenesis and promoting chromosomal rearrangements. However, the incomplete characterization of type-C ERV sequences in CHO cells, as well as the absence of a clear link between the genomic type-C ERV sequences and viral particles, have hampered the establishment of a similar ERV inactivation strategy in CHO cells.

Here we sought to characterize in-depth the budding type-C ERV sequences at the genome, transcriptome and viral particle level using CHO-K1 cells. In contrast to previ-

ous studies, we identified transcribed type-C ERV group 1 sequences yielding full-length transcripts with open reading frames, suggesting that this ERV group results in potentially functional retroviruses. Using CRISPR-Cas9 genome editing, we mutated the expressed group 1 type-C ERV sequences, and showed that specific loss-of-function mutations within the *gag* gene of a single ERV suffice to decrease the release of functional viral RNA-loaded particles by more than 250-fold. This indicated that a single ERV locus is responsible for most type-C viral particles released from CHO cells. Altogether, our study provides a novel strategy to further improve the safety profile of CHO cells, paving the way to complete eradication of endogenous viral contaminations in cultures of CHO cells.

4.3 Material and Methods

Cell culture

Suspension-adapted Chinese hamster ovary (CHO-K1) derived cells were maintained in serum-free HyClone SFM4CHO medium supplemented with HyClone Cell boost 5 supplement (GE Healthcare), L-glutamine (Gibco), HT supplement (Gibco) and antibiotic-antimycotic solution (Gibco). CHO cell viability was assessed by Erythrosin B dye (Sigma-Aldrich) and viable cell density and cell size were quantified using the LUNA-FL Dual Fluorescence Cell Counter (Logos Biosystems). The cells were cultivated in 50ml TubeSpin bioreactor tubes (TPP, Switzerland) at 37 °C, 5% CO_2 in a humidified incubator with 180 rpm agitation rate and passed every 3-4 days.

Plasmid construction

The mammalian codon-optimized *Streptococcus pyogenes* Cas9 (SpCas9) nuclease expression plasmid JDS246 (Addgene plasmid #43861) [360] was used to introduce site-specific DSBs. The CRISPRseek R package [397] was applied to design single guide RNA (sgRNA) sequences that target the myristoylation (Myr) or PPYP motifs in the *gag* consensus sequence of group 1 type-C ERVs. Among all potential sgRNAs, three Myr (Myr2, Myr4, Myr8)- and five PPYP (PPYP5, PPYP6, PPYP7, PPYP13, PPYP20)-specific sgRNA sequences were selected as they mediate DSB cleavage no more than

25 bp apart from the target motif, and as they are predicted to have high sgRNA efficiency using various scoring tools (CRISPRseek [397]; Sequence Scan for CRISPR [398]; sgRNA scorer 1.0 [399]) (Table S4.1). Genome-wide off-target cleavage analysis for these sgRNA sequences was performed using the CRISPRseek R package using the CHO-K1 cell genome as reference sequence. SgRNA oligonucleotides were designed using the Zinc Finger Targeter software support tool [400, 401], and annealed sgRNA oligonucleotides were subsequently cloned into the mammalian sgRNA expression vector MLM3636 (Addgene plasmid #43860) as previously described [360]. For sgRNA sequences lacking a guanine (G) nucleotide at the 5' end, an additional, non-pairing G was appended to improve transcription from the sgRNA expression plasmid [370]. All primers used in this study were purchased from Microsynth AG (Balgach, Switzerland) and are listed in Tables S4.2-S4.4.

RNA-seq analysis of the cellular CHO mRNA

To complement the genomic CHO ERV characterization, the total cellular mRNA of wild-type CHO cells was sequenced using Illumina paired-end technology, resulting in a library of 83 mio reads of size 2x100 bp. This library was mapped with BWA [402] to the entire CHO-K1 transcriptome (version 2014) extended with six ERV sequences. The expression level per transcript was computed as the RPKM, normalizing the number of reads mapping to the specific transcript by the transcript length and the total number of reads mapping to the annotation to account for different library sizes.

VLP RNA extraction

To characterize viral sequences released within VLPs in the CHO supernatant, total VLP RNA was extracted from CHO culture supernatants using the Invitrogen PureLink Viral RNA/DNA mini kit (Thermo Fisher Scientific) according to the manufacturer's protocol with some modifications. The supernatants were used freshly prepared, or after only one freezing and thawing cycle. 500 μl of supernatant were loaded on a Corning Costar Spin-X column centrifuge tube with 0.22 μm membrane filter and centrifuged at 16000 g for 1 min. Approximately 12.5 units of RNase-free DNase (Macherey-Nagel) were added to 500 μl of CHO cell culture supernatants, which were incubated for 15 min at 37 °C to digest the residual DNA possibly present. The resulting extracts were then treated as

described in the PureLink Viral RNA/DNA mini kit protocol. The RNA recovered from the spin columns was resuspended in 30 μl of RNase-free water, followed by another DNase treatment using 10 units of RNase-free DNase (Macherey-Nagel) for 30 min at 37 °C. After the addition of EDTA at a 5 mM final concentration, a DNase denaturation step was made by incubating the extracts at 70 °C for 15 min. The samples were afterwards placed on a Microdialysis MF-Millipore Membrane Filter with 0.025 μm pore size (Merck-Millipore) for 15 min in order to remove salts such as EDTA remaining in the samples.

RNA-seq analysis of CHO VLPs

VLP RNA sequences from wild-type CHO cells and from clones D12 (treated with Myr2 sgRNA) and E10 (treated with PPYP6 sgRNA) were sequenced using Illumina paired-end technology, resulting in libraries of 187 mio, 306 mio and 283 mio of size 2x75 bp. Reads were first checked for sequence quality (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc`) before mapping them to 261 ERV sequences using BWA [402]. Expression levels per ERV were computed as the RPKM, normalizing the number of reads mapping to the specific transcript by the transcript length and the total number of reads mapping to the annotation to account for different library sizes.

Inactivation of ERV sequences, fluorescent cells enrichment and single cell isolation

CHO-K1 cells were seeded at 300,000 cells/ml one day prior to transfection. On the day of transfection, 700,000 cells were electroporated with 3700 ng of CRISPR-Cas9 and 1110 ng of Myr- or PPYP-specific sgRNA expression plasmids using the Neon transfection system (Thermo Fisher Scientific), according to the manufacturer's instructions. CRISPR-Cas9 and sgRNA expression plasmids were used at equimolar ratio. 200 ng of pCMV-DsRed-Express plasmid (Clonetech) was added to each transfection condition as transfection control. For CRISPR control experiments, the Myr- or PPYP-specific sgRNA plasmids were substituted with the empty sgRNA expression vector (empty vector control).

To enrich for transfected and ERV-mutated CHO cells, at least 70,000 cells were bulk-

sorted for the highest 30-40% of transfected dsRed expressing cell population 48-72h after transfection using the MoFlo Astrios EQ or FACSAria II cell sorters (Beckman Coulter). Cells were then briefly centrifuged to exchange medium and expanded. To isolate single cell clones, CRISPR-treated cells were incubated at room temperature with DAPI viability dye (BD Biosciences) for 15 min. Viable cells were single cell sorted into 96well plates using the FACSAria Fusion cell sorter (Beckman Coulter). Cell clones were recovered in HyClone SFM4CHO medium supplemented with L-glutamine, HT supplement, antibiotic-antimycotic solution and ClonaCell-CHO ACF Supplement (Stemcell Technologies) to increase post-sort survival. Flow cytometry data were analyzed using FlowJo software v10.4.2. Cells were first gated using side scatter (SSC) versus forward scatter (FSC) to separate the intact cell population from debris, and then selected for single cells in the SSC-H/SSC-W and FSC-H/FSC-W plots. This single cell population was then gated for dsRed+ cells with non-fluorescent cells as gating control.

ERV mutation efficiency

To assess the cleavage efficiency of ERV-specific sgRNAs, the frequency of ERV mutations was determined among the transcribed ERV sequences. Total RNA from CRISPR-treated polyclonal cell populations was extracted using the NucleoSpin RNA kit (Macherey Nagel) and reverse transcribed into cDNA using oligo(dT)15 primers and the GoScript Reverse Transcription System (Promega). For CRISPR-treated single cell clones, total RNA was isolated using the SV 96 Total RNA Isolation System (Promega) and reverse transcribed using GoScript Reverse Transcription Mix, Oligo(dT) (Promega). PCR amplification of the CRISPR target regions was carried out using OneTaq DNA polymerase (New England BioLabs) with group 1 type-C ERV-specific primers (Table S4.4). PCR products were analyzed by Sanger sequencing and analyzed for mutations. The mutagenesis frequency in CRISPR-treated polyclonal populations was determined by decomposition of the mixed Sanger sequencing chromatograms and comparison to untreated (wild-type) cells using the TIDE software [403].

Deep amplicon sequencing of CRISPR-targeted genomic regions

To assess the number of CRISPR-induced ERV mutations at the genome level, DNA was extracted from clones bearing a mutation in the expressed group 1 type-C ERV sequence

as well as from empty sgRNA vector-treated and wild-type cells using the DNeasy Blood & Tissue Kit (Qiagen). This extracted genomic DNA was used to prepare sequencing libraries in a two-step PCR approach as described in the Illumina "16S Metagenomic Sequencing Library Preparation" protocol with some modifications. Briefly, degenerate primers were designed using the Primer Design-M tool [404] to amplify approximately 300 bp of the genomic region flanking the Myr2 and PPYP6 sgRNA target sites of all predicted type-C ERV sequences (290 bp amplicon for Myr, 314 bp amplicon for PPYP, Table S4.3). Degenerate primers contained various 0-3 bp heterogeneity spacers to increase template complexity [405] and Myr or PPYP primers were mixed at the predicted genomic frequency. In the first PCR round, 100 ng of isolated genomic DNA was used to PCR amplify the Myr and PPYP target loci using KAPA HiFi HotStart ReadyMix (2X) (Kapa Biosystems) for 23 and 20 cycles, respectively. PCR amplicons were purified with AMPure XP beads (Beckman Coulter) using a 1:1 bead ratio. Amplicon quality and size were verified on an Agilent 2100 Bioanalyzer and DNA was quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific). In the second PCR round, Illumina Nextera XT Index sequencing adapters were added to 15 ng of purified amplicons using 8 PCR cycles. The final libraries were purified with AMPure XP beads (Beckman Coulter) using a 1:1.12 bead ratio. Library quality and size were verified using Fragment Analyzer (Advanced Analytical) and quantified using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific). Libraries were pooled at equimolar ratio, spiked with 25% PhiX and sequenced using 2x 250 bp paired-end sequencing on an Illumina Miseq System at the Genomic Technologies Facility of the University of Lausanne (Switzerland).

Analysis of deep amplicon sequencing

All sequenced Illumina paired-end reads were trimmed using trimmomatic (v0.36; IL-LUMINACLIP:config/daf.adapt.fa:2:30:10 LEADING:20 TRAILING:20 MINLEN:50), merged with FLASH2 (v2.2.00; –max-overlap 350) and converted to FASTA format. Spacer and primer sequences were translated into sequence profiles (pftools v3.0), flagged on all reads using pfsearchV3 (pftools) and subsequently excised.

Weighted profiles for both Myr and PPYP CRISPR-targeted regions were created as following: merged paired-end reads from wild-type control samples were cleaned (reads lengths exceeding Myr: > 300 bp, PPYP: > 400 bp removed) and clustered using cd-hit (v4.6.8, -n 11 -c 0.97 -A 0.95). Clusters with $> 0.3\%$ of members were kept and a con-

sensus sequence for each cluster created based on 100 random sequences (mafft v7.310; –globalpair –maxiterate 1000 and cons from EMBOSS suite v6.6.0). Multiple sequence alignments were done for these consensus sequences (mafft; –globalpair –maxiterate 1000) and translated into weighted profiles with pfmake (pftools suite 2.3.5.d; -0 -G 3 and a search-like scoring matrix). Profiles were calibrated using scores from pfsearchV3 against scrambled sequences (60 bp window).

The calibrated profiles were used to search both, wild-type control reads and CRISPR-treated sample reads with pfsearchV3 generating a psa output. Modifications observed in comparison to the profile were determined and further analyzed in R (v3.4.2). Alignment differences observed in wild-type control samples compared to the profile were removed in CRISPR-treated sample reads and only CRISPR-unique events kept. Events which appeared in $< 0.1\%$ of total reads were removed and reads from identical events clustered using mafft (< 1000 members: –globalpair –maxiterate 1000; > 1000 members: –globalpair –retree 1 –maxiterate 0) to generate one consensus sequence per observed CRISPR mutation (using cons). R packages used are: data. table, RColorBrewer, colorspace, seqinr, glue, knitr.

For all identified CRISPR-derived mutations, Illumina raw reads were clustered using the Jukes-Cantor genetic distance model under the UPGMA tree building method to test for ERV locus-specific genetic variations in the mutation flanking region.

Whole genome sequencing of ERV-mutated CHO clone

To identify mutated ERV loci in the whole CHO genome, high-molecular-weight DNA was extracted from the sgRNA PPYP6-treated E10 clone using the Blood & Cell culture DNA kit (Qiagen). DNA quality and quantity were verified using Fragment Analyzer (Advanced Analytical) and Quibit (Thermo Fisher Scientific), respectively. Five SMRT cells were sequenced on a PacBio Sequel system (Pacific Biosciences) at the Genomic Technologies Facility of the University of Lausanne (Switzerland). Each SMRT resulted in 5.9 Gbp – 7.2 Gbp in subreads and a N50 of 18.5 kbp. This yielded in total 34.15 Gbp of sequenced DNA in subreads which equals $\sim$14x theoretical coverage over the CHO genome. The consensus sequence describing the CRISPR-derived mutation at the PPYP site in clone E10 and the PPYP wild-type control cluster consensus sequences were used as reference for this analysis. WGS PacBio reads from the E10 clone were aligned

against these reference sequences (minimap2 v.2.8; -x map-pb –secondary=no; samtools v.1.8) and 2 subreads with the E10 PPYP mutation were identified (lengths of subreads: 4.3 kbp and 15 kbp). The latter contained sufficient genomic non-ERV sequence and was successfully mapped onto the NCBI CHO-K1 reference genome using minimap2.

Analysis of therapeutic protein expression

To assess the therapeutic protein production capacity of ERV-modified cells, polyclonal cell populations and cell clones previously treated with ERV-specific or empty sgRNA expression plasmids were electroporated with a trastuzumab immunoglobulin G1 (IgG1) heavy and light chain expression vector bearing a puromycin resistance gene [363]. As control, wild-type CHO-K1 cells were transfected with the same expression vector in parallel. Two days after transfection, cells were transferred to culture medium containing 5 μg/ml puromycin and selected for three weeks. Immunoglobulin titers from cultures of stable trastuzumab expressing cell populations were quantified during ten-days fed-batch cultures as previously described [363]. Briefly, cells were seeded at $0.3*10^6$ cells/ml in 5 ml initial culture volume without puromycin selection. Cell culture was fed with HyClone Cell boost 5 supplement (GE Healthcare) at 16% of the initial culture volume on days zero, two, three and six to eight of the cell cultivation. Cell density and viability was assessed at days three, six, eight and ten and immunoglobulin secretion in the cell culture supernatant was measured on days six, eight and ten by sandwich ELISA.

4.4 Results

4.4.1 Characterization of ERV elements in CHO-K1 cells

To search for ERVs present in CHO cells, the CHO-K1 genome was assembled *de novo* using PacBio long-read sequencing, and the previously reported IAP and ML2G murine retroviral sequences were searched in this assembly [352, 393]. Furthermore, we used as well profiles to complement and validate the ERV elements identified by sequence similarity. We found approximately 160 copies of IAP-like proviral elements within the CHO genome. In addition, we identified 173 gammaretrovirus type-C proviruses that shared at least 80% sequence identity to the ML2G sequence in CHO cells [352] (Table

4.1). Although the identified number of type-C proviruses is in line with previous estimations [354], we noticed that some ERV copies could not be successfully placed in the assembly suggesting that 173 copies are likely an underestimation of the total reservoir of type-C ERV elements in CHO cells. Among the identified 173 type-C ERV sequences, only 112 contained the *gag*, *pol* and *env* genes, as required to produce a functional ERV. Phylogenetic analysis of these full-length hamster type-C ERV sequences revealed their close similarity to other mammalian retroviral elements, such as the Feline leukemia virus (FeLV) and the Murine leukemia virus (MLV) (data not shown). Among these type-C ERV sequences, we identified two distinct groups, termed group 1 and group 2, which were composed of 101 and 36 members, respectively (Fig. 4.1A). Group 1 and group 2 type-C ERVs formed the predominant and functionally most conserved sequence clusters, with complete 5' LTR-*gag-pol-env*-3' LTR proviral structures, and they also shared most similarity to MLV elements, which are known to produce viral particles infecting primate cell lines [394]. This implied the ERVs of group 1 and 2 as the most likely candidates for viral particle formation.

Table 4.1: Number and frequency of distinct type-C ERV sequences detected in the genome, transcriptome and viral particles of CHO-K1 cells.

Detection level	Type-C ERV sequence number	ERV group relative frequency
Genomic DNA	173	group 1 ~ group 2
Cellular mRNA	3-32	group 1 > group2
Viral particles	1-5	group 1 only

Further sequence analysis highlighted that the *gag* and *pol* genes were highly conserved among group 1 and group 2 ERV sequences but that ERVs belonging to group 1 showed overall less diversity than ERVs from group 2 (Figs. 4.1B-4.1D). On average group 1 ERV sequences shared 99% sequence identity and likely form three subgroups (marked in blue, turquoise and green in Figs. 4.1B-4.1D). However, the overall high conservation of these ERV sequences and the frequency of residual PacBio derived errors in the assembled genome hampered the direct identification of which of these group 1 and group 2 ERVs may be functional and potentially active.

To complement the genomic CHO ERV characterization, the total cellular mRNA was
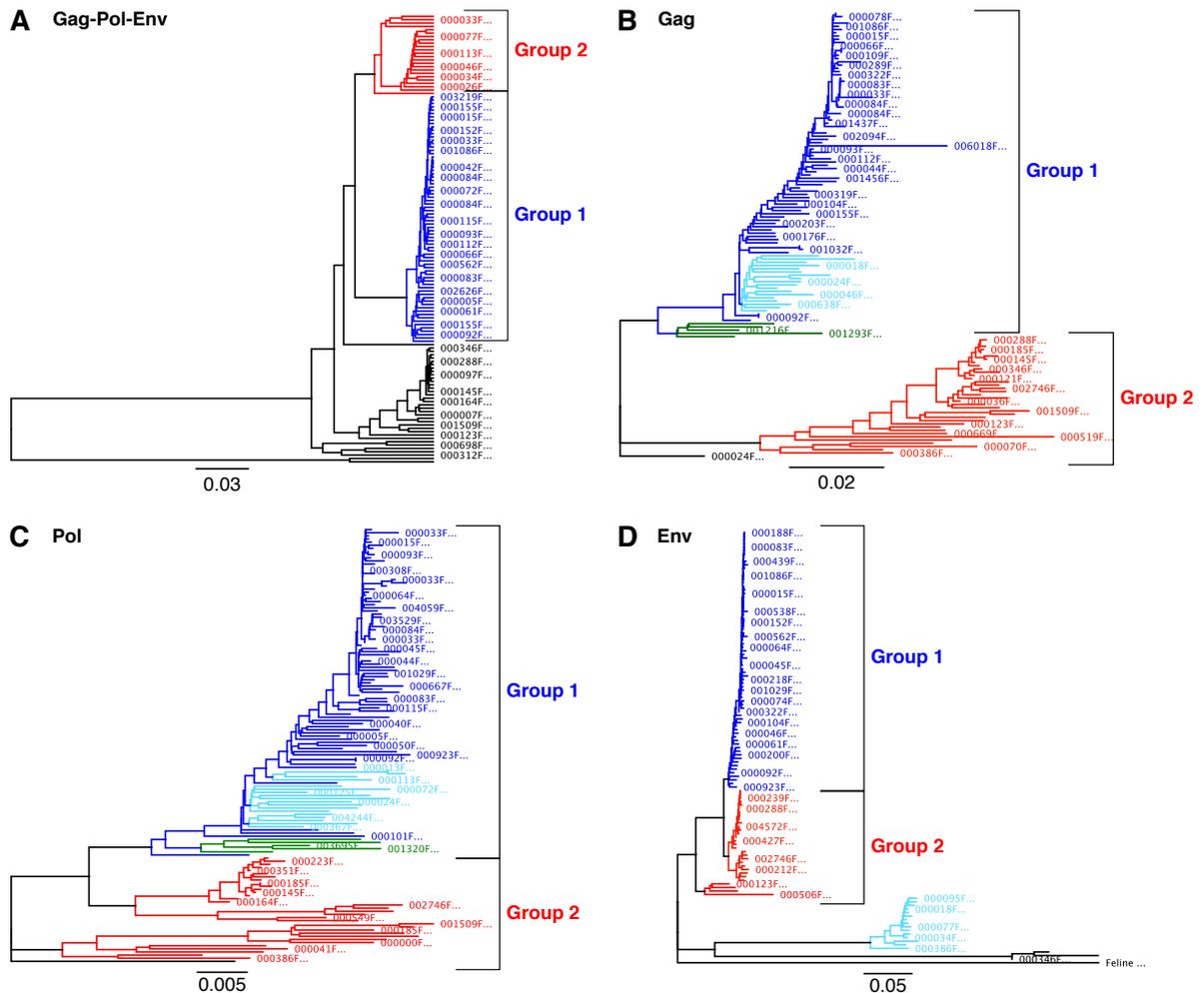
Figure 4.1: Phylogenetic analysis of full-length type-C ERV DNA sequences within the CHO genome. Phylogenetic trees were generated based on alignments of the complete *gag-pol-env* sequence (A) or the *gag* (B), *pol* (C) or *env* (D) sequence separately from 112 full-length type-C ERV. The scale under the tree shows the substitution rate per nucleotide. Group 1 with its three subclusters (in blue, turquoise and green) and group 2 (in red) are indicated by brackets. ERV sequences in black belong neither to group 1 nor group 2.

sequenced using Illumina short-read technology in order to refine the transcribed ERV sequences. Type-C ERV mRNAs were among the top 10 most abundant transcripts in CHO cells (data not shown). Mapping of these reads to type-C ERV representatives showed that 99.5% corresponded to group 1 and 2, indicating that these two groups contribute the vast majority of the transcribed ERVs of CHO cells. While the Illumina reads mapped mainly on two easily distinguishable group 2 ERV sequences, they mapped on approximately 30 group 1 ERV sequences (Fig. 4.2A). As group 1 ERVs are most highly conserved, this did not allow unambiguous attribution of these reads to one or few unique group 1 loci. Interestingly, both transcribed group 2 ERV sequences contained interrupted ORFs and/or missing coding sequences, one containing two deletions of a total of 2350 bp in the *pol*

gene and the second having one frameshift in the *gag* and *pol* genes, as well as three stop codon mutations in *pol*. These mutations were confirmed by Sanger sequencing. In contrast, the transcribed group 1 ERV sequences seemed to encode full-length *gag*, *pol* and *env* transcripts. Overall, this suggested that between 3 to 32 ERV loci are transcribed, corresponding to approximately 2-20% of the total ERV elements in CHO cells (Table 4.1). Such an ERV expression frequency agrees with previous reports indicating that the majority of ERVs are epigenetically silenced in cell lines and organisms [349]. Finally, among the total cellular mRNA, we also detected LTR-containing viral genomic RNA, indicating that CHO cells are capable of producing retroviral genomes that may be encapsulated and released as retroviral particles in the cell supernatant.

Moreover, we also characterized viral genomic RNA loaded into VLPs by Illumina sequencing. LTR-containing viral genomic RNA was enriched twenty-fold compared to the total cellular mRNA sequences (Fig. 4.2A). This indicated that CHO cells are able to shed retroviral particles containing genomic viral RNA into the cell supernatant. In-depth analysis of these viral RNA sequences indicated that group 1-derived reads were mostly present in the released viral particles (Fig. 4.2A). Moreover, these sequences mapped to between 1 and 5 different group 1 ERV loci, suggesting that only few group 1 ERV loci are responsible for the production of viral particles in CHO cells (Table 4.1).

To further characterize the functional group 1 type-C ERV sequences, we designed group 1-specific probes for Fluorescent in-situ hybridization (FISH) experiments. Using these probes, we detected approximately 50-100 group 1 ERV integration sites in the CHO-K1 genome (Fig. 4.2B). Retroviral integrations were dispersed throughout the CHO-K1 genome, with a possible integration hotspot in one of the smallest chromosomes. Additionally, when staining for group 1 nascent mRNAs, we observed a unique highly transcribed site, suggesting that only a single group 1 ERV locus might be transcriptionally active (Fig. 4.2C).

Altogether, systematic ERV characterization at the genome, transcriptome and viral particle level identified several group 1 type-C ERVs as strong candidates for the expression and release of functional retroviral particles from CHO-K1 cells. Although the high sequence identity among the type-C ERV sequences concealed the exact number of expressed ERV loci, these data suggested that mutating few transcribed group 1 ERV loci by genome editing might suffice to prevent ERV particle formation.
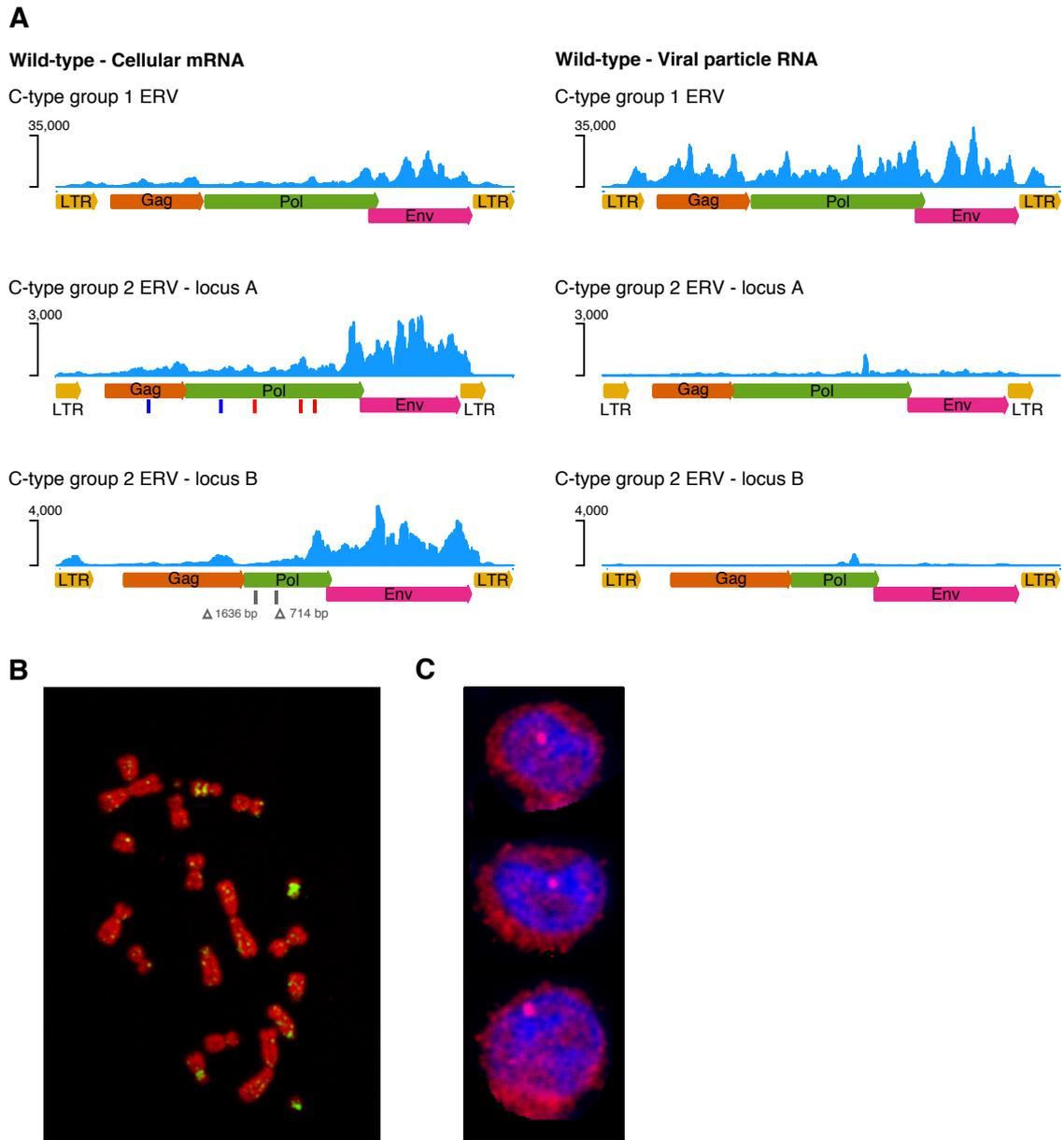
Figure 4.2: Characterization of expressed type-C ERV sequences in wild-type CHO cells. (A) Mapping of Illumina sequencing reads of total cellular RNA (left) or of viral particle RNA (right) obtained from CHO-K1 cells on group 1 and group 2 type-C ERV sequences. Reads were mapped to a consensus sequence for group 1 and on two distinct loci (locus A and locus B) for group 2. Signs present on the schematic representation of group 2 ERV on locus A and B show the mutation type occurring in these ERV sequences, blue for frameshift mutations, red for stop codon mutations and grey for deletions, with the deletion size indicated as the number of bases. (B) Representative metaphase spread of CHO-K1 chromosome FISH analysis using fluorescent probes specifically targeting group 1 type-C ERV. Chromosomal DNA is represented in red and the FISH signals of integrated retroviral sequences are shown as green dots (C) Three representative interphase CHO-K1 cells are shown, with mRNAs depicted in red and group 1 type-C ERV RNA illustrated in blue color. The bright purple dot represents the nascent group 1 mRNA at the transcription site.

4.4.2 Designing ERV-specific sgRNA sequences for CRISPR-Cas9 genome editing

To inhibit the release of potentially infective viral particles from CHO cells, we aimed to disrupt conserved ERV sequence motifs critical for viral particle release. The Gag protein plays a pivotal role during retrovirus budding, and, consistently, it was conserved among the type-C ERVs in CHO cells. However, in contrast to the *pol* gene, for instance, the *gag* sequences were sufficiently different to distinguish group 1 from group 2 type-C ERV sequences, allowing to specifically target group 1 ERV particles (Figs. 4.1B and 4.1C). We selected two conserved *gag* sequences involved in viral budding – the myristoylation (Myr) and the PPxY motifs – as targets for CRISPR-Cas9-mediated mutagenesis. The N-terminal Myr motif locates at a glycine residue at position 2 downstream of the ATG translation initiation codon (Fig. 4.3). Myristoylation of Gag is generally considered essential for targeting the protein to the host plasma membrane [406]. Mutations that directly interfere with Gag myristoylation, that block translation from the physiological start site or that create a loss-of-function *gag* transcript will perturb proper viral particle assembly at the plasma membrane, and hence block retroviral particle budding [406, 407]. In addition to Myr, the conserved proline-rich PPxY motif also contributes to retrovirus budding, likely by interacting with the ESCRT machinery [408], and its mutation strongly inhibits viral particle release [409]. The PPxY motif overlapped with a PPYP motif that is conserved in group 1 and group 2 CHO ERVs, which is termed PPYP hereafter to refer to this CHO-specific PPxY-related budding motif.

We designed eight sgRNAs against the group 1 *gag* consensus sequence: three constructs targeting the Myr motif (Myr2, Myr4, Myr8) and five constructs targeting the PPYP motif (PPYP5, PPYP6, PPYP7, PPYP13, PPYP20) (Fig. 4.3). The selected sgRNA sequences located close to the corresponding target motifs and were predicted to perfectly match between 33 and 117 target ERV sequences, but to target up to 283 sites when allowing a maximum of three mismatches and non-canonical protospacer adjacent motif (PAM) sites (Table S4.1). Importantly, all these potential cleavage sites map to ERV sequences, while other off-target sites in the CHO genome were not detected. Although these sgRNA sequences contain a multitude of predicted target sites, we hypothesized that expressed ERVs might be preferentially cleaved by the CRISPR-Cas9 nuclease, due its preference for open chromatin [410].

Provirus genome

~ 9.25 kb

Pol

5' LTR | Gag | Env | 3' LTR

Myr PPYP

Myr sgRNA target sites

Myr

Myr2

Gag

TTTGTCTTGTAGCTGTCATTTGTGCCCTCCTAAGCCTAGAAACTATGGGGCAAACTGTCACCACTCCTTTGTCCCTAACACTCTCCCACTGGAAAGATG
AAACAGAACATCGACAGTAAACACGGGAGGATTCGGATCTTTGATACCCCGTTTGACAGTGGTGAGGAAACAGGGATTGTGAGAGGGTGACCTTTCTAG

Myr4   Myr8

PPYP sgRNA target sites

PPYP   PPYP7

PPYP5   PPYP6

Gag

GGGAGAAGACCAGTTGGTTGATCTATTAACGGAGGAGCCCCCGCCATATCCGCCACTGCCGCCCCCACCAGAGGCAGAAGCGGACTCCGCCGCTGCCT
CCCTCTTCTGGTCAACCAACTAGATAATTGCCTCCTCGGGGGCGGTATAGGCGGTGACGGCGGGGGTGGTCTCCGTCTTCGCCTGAGGCGGCGACGGA

PPYP13   PPYP20

Figure 4.3: CRISPR-Cas9 target sites for ERV mutagenesis. The orientation and position of the eight sgRNA sequences designed to target the Myristoylation (Myr) and PPYP motifs of the gag group 1 type-C ERV consensus sequence are illustrated by grey arrows. The CRISPR-Cas9 DSB sites are shown by open triangles for sgRNAs targeting the forward strand (Myr2, PPYP5, PPYP6, PPYP7) and by filled triangles for sgRNAs targeting the reverse strand (Myr4, Myr8, PPYP13, PPYP20). The protospacer adjacent motif (PAM) sites are marked by blue bold letters.

To mutate the Gag budding motifs, CHO-K1 parental cells were transiently transfected with CRISPR-Cas9 and Myr or PPYP sgRNA expression plasmids together with a dsRed transfection control plasmid. For CRISPR control samples, the *gag*-specific sgRNA expression plasmids were replaced with a non-targeting empty vector sgRNA control plasmid (empty vector) or cells were left untreated (wild-type). Transfected dsRed positive (dsRed+) cells were bulk-sorted to enrich for cells containing mutations in the target motifs. Following treatments with ERV-specific sgRNAs, we noted an overall reduced frequency of transfected dsRed+ cells as well as a significant drop in dsRed fluorescence intensity in dsRed+ cells compared to control samples, suggesting that the most highly transfected cells may not survive because of a high frequency of genome cleavage (Figs. S4.1A and S4.1B). This effect was reduced for Myr4 sgRNA treated cells, which has the lowest number of predicted target sites. We also observed an elevated cell granularity following CRISPR treatment which inversely correlated with the frequency and expression intensity of dsRed+ cells (Figs. S4.1C and S4.1D). Highly granular cells were previously reported to consist of pro-apoptotic and/or dying cell populations [411]. Altogether, this provides evidence that CRISPR-mediated ERV cleavage impedes cell proliferation and

survival, especially in highly transfected cells, implying that ERV-specific sgRNAs efficiently introduce DSBs at multiple target sites in the CHO genome.

To estimate the CRISPR-mediated mutagenesis frequency within the expressed group 1 ERVs, the total cellular mRNA of bulk-sorted Myr- and PPYP-treated cells was reverse transcribed and PCR amplified. Polyclonal PCR products were then either directly sequenced or cloned into bacterial vectors prior to single colony sequence analysis. Based on these analyses, we estimated that the designed *gag*-specific sgRNAs introduced mutations in roughly 10 to 30% of the ERV mRNAs, and that the Myr2 or PPYP6 sgRNAs were most efficient (Fig. S4.2, Tables S4.5-S4.7). Interestingly, some of the recovered mutations were expected to block translation or introduce frameshifts, and thus should cause Gag loss-of-function phenotypes.

4.4.3 Isolation and characterization of ERV-mutated CHO-K1 clones

Given that roughly 10-15% of the expressed group 1 ERV sequences are predicted to be mutated, we hypothesized that a potential reduction in viral particle release would be difficult to detect within a polyclonal population. Thus, we isolated single CHO cell clones from bulk-sorted Myr2- or PPYP6-edited cell pools and screened for those having mutations in the expressed group 1 ERV sequences. 18 out of 95 screened Myr2 sgRNA-treated clones (18%) and 14 out of 181 screened PPYP6 sgRNA-treated clones (8%) contained group 1 ERV mutations at the mRNA level, in line with previous editing estimations (Table 4.2, Figure S4.2, Tables S4.5-S4.7). Among the Myr2-mutated clones, the majority possessed an identical 1 bp insertion upstream of the ATG start codon (Tables S4.6 and S4.8), which likely resulted from staggered CRISPR-Cas9 cleavage [262]. No clone treated with the PPYP6 sgRNA acquired a mutation disrupting the PPYP motif. Nonetheless, two Myr2- and eleven PPYP6-derived clones contained mutations either blocking translation or frameshifting the *gag* transcripts, hence making them promising candidates for reduced viral particle release. We also observed that the Sanger sequencing chromatogram of the repair junctions of all clones showed a clear mutated sequence and lacked background noise in the CRISPR flanking sequences. This supported our earlier hypothesis that only a single group 1 ERV locus might be prominently transcribed and lead to the production of viral particles in CHO cells.

Table 4.2: Detection of CRISPR-mediated mutations in the expressed type-C ERV sequences of edited CHO-K1 clones.

Sample	# screened clones	# mutated clones	Mutation frequency	Loss-of-function mutation frequency[1]
Myr2 sgRNA	95	18	19%	11%
PPYP6 sgRNA	181	14	8%	79%
Total	276	32	12%	45%

[1] Includes translation inhibition and frameshift mutations and is expressed relative to the number of mutated clones.

To further investigate the CRISPR-derived mutations at the genome level, we deep-sequenced the Myr and PPYP flanking regions of type-C ERVs in a subset of CHO. We selected two Myr2- and four PPYP6-edited clones with Gag loss-of-function mutations in the expressed group 1 type-C ERV sequences (clones CO2 and D12 for Myr2; A02, E10, K03 and K14 for PPYP6) as well as one Myr2-derived clone with a large mutation outside of the group 1 ERV coding (G09) and genotyped them along with wild-type and empty vector control samples (Table S4.8).

To detect CRISPR-derived mutations and distinguish them from sequence variations naturally occurring at each target site, we clustered the reads from wild-type CHO cells and used these cluster consensus sequences to create diversity profiles. When clustering by 97% sequence similarity, we identified 34 Myr and 28 PPYP clusters that represented the natural ERV sequence diversity present within the Myr and PPYP flanking regions (Figs. 4.4A and 4.4B, S4.3). Despite the observed sequence diversity, the Myr and PPYP motifs themselves were highly conserved, in agreement with their biological significance for viral budding. The identified clusters correlated well with the type-C ERV groups previously characterized from the CHO genome as well as with their predicted frequencies, corroborating our previous characterization of ERV sequences at the whole genome level (Figs. 4.1, 4.4A and 4.4B, Tables S4.9 and S4.10). For both targets, the largest cluster encompassed approximately 40% of all reads and its consensus sequence was identical to the group 1 type-C ERV sequence present in VLPs (yellow highlight, Figs. 4.4A and 4.4B). Among all clusters, 13 Myr and 8 PPYP clusters, including the largest clusters, could be targeted by the Myr2 and PPYP6 sgRNAs, accommodating for 61% and 72% of
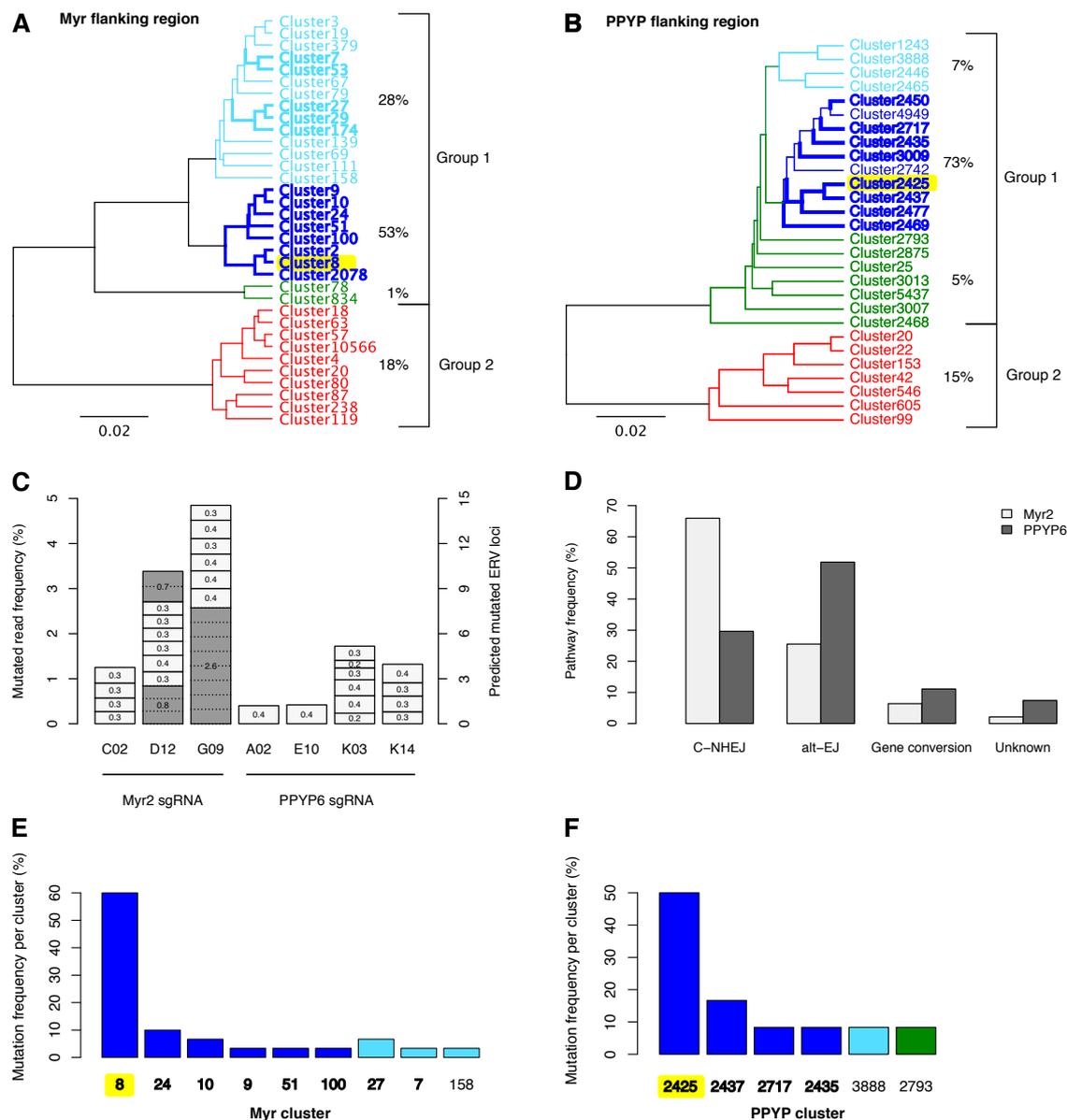
Figure 4.4: Assessment of sequence diversity in the Myr and PPYP motif flanking regions and analysis of CRISPR-derived mutations by deep DNA sequencing. Targeted amplification of approximately 300 bp surrounding the Myr and PPYP CRISPR target sites was performed using type-C ERV specific primers and amplicons were analyzed by Illumina deep sequencing. Clustering analysis was based on 97% similarity of wild-type CHO-K1 deep sequencing reads from the Myr (A) and PPYP (B) flanking sequences. Clusters are indicated in turquoise, blue, green or red lettering, according to the phylogenetic groups identified in Figure 4.1. Clusters containing the Myr2 sgRNA and PPYP6 sgRNA recognition sites and an adjacent PAM sequence are shown in bold, and the most abundant cluster per target site is highlighted in yellow. Values to the right represent the frequency of reads obtained for each subcluster relative to the total number of reads. (C) Number of distinct mutations and their corresponding read frequencies in seven clones (C02, D12, G09, A02, E10, K03, K14) isolated from Myr2 or PPYP6 sgRNA-treated polyclonal populations. All clones display mutations in the expressed group 1 type-C ERV locus. Grey shaded boxes represent mutations occurring at a higher than average read frequency ($> 0.4\%$, left-hand side axis) and the predicted number of ERV loci containing an identical mutation is indicated as dashed lines. Caption continued next page.

Figure 4.4: The estimated total number of mutated ERV loci of each clone is indicated by the right-hand side axis. (D) Frequency of Myr2 or PPYP6 sgRNA-induced repair junctions compatible with C-NHEJ, alt-EJ or HR-mediated gene conversion DSB repair mechanisms. Repair junctions incompatible with these three main DSB repair mechanisms are grouped as Unknown. A total of 74 DNA repair junctions (n_{Myr}=47, n_{PPYP}=27) obtained from both Sanger mRNA and Illumina deep DNA sequencing were analyzed. (E and F) Frequency of the wild-type CHO clusters representing best the mutation-flanking sequence of 30 Myr2- and 12 PPYP6-derived mutations of deep sequencing reads. Clusters containing the Myr2 or PPYP6 sgRNA recognition sites including an adjacent PAM site are shown in bold letters (on-targets), while clusters with sgRNA mismatches are shown in normal letters (off-targets). Off-target cluster possesses mismatches at position 13 or 15 in the sgRNA recognition site.

the captured read diversity, respectively (bold letters, Figs. 4.4A and 4.4B).

Using these wild-type CHO clusters and corresponding diversity profiles, we found between 1 and 7 distinct CRISPR-derived mutations per clone, including the mutations previously detected at the mRNA level (number of boxes, Fig. 4.4C; Table S4.6). The detected mutation range spanned from a 114 bp deletion up to a 78 bp insertion. As expected, CHO cells treated with the empty vector expression plasmid lacked additional mutations in the CRISPR target sites. Some mutations, for instance, a 1 bp insertion, occurred within all three genotyped Myr2-treated clones but were absent in the PPYP6 clones, as expected from sgRNA-specific repair outcomes [278].

Typically, a given mutation was detected at a read frequency of approximately 0.3%, which thus must represent a single ERV locus in the CHO genome (Fig. 4.4C). However, three Myr2-derived mutations were discovered at a read frequency well above 0.3%, with the same 1 bp insertion being present in 2.6% of all G09 clone reads. Consequently, this implies that the same mutation may occur more than once in the same clone. In support for this hypothesis, the reads of predicted single locus mutations (i.e. clones A02 or E10) were highly similar in the mutation flanking region, while the reads of abundant mutations (i.e. G09 1_1) contained variations in the mutation flanking regions, suggesting that the same mutations may have occurred repeatedly at distinct ERV loci (Fig. S4.4). In the case of G09 1_1, five ERV groups could be distinguished with one group having four-times more reads than the others, indicating that this mutation should have occurred at eight distinct ERV loci in the G09 clone. Therefore, we concluded that each clone acquired between 1 and 14 ERV mutations following transient CRISPR transfection (Fig. 4.4C). The identification of clones having only one mutated ERV at the DNA level, together

with the finding that this mutation was identical to the single mutation detected at the intracellular RNA level, further substantiated that a single group 1 type-C ERV locus is transcribed, and likely responsible for the release of type-C retroviral particles from CHO cells.

The repeated occurrence of identical mutations within one clone raised the question of whether they may result from gene conversion, a homologous recombination (HR)-related repair mechanism, in which a previous mutated ERV locus is used as template to repair other cleaved ERV sites. To find evidence for HR activity following Myr2- and PPYP6-mediated cleavage, we combined the previously obtained mRNA and DNA data and analyzed a total of 74 DNA repair junctions (n_{Myr}=47, n_{PPYP}=27). While Myr2 sgRNA-mediated cleavage led to an overall higher mutation frequency, with a preference for insertions, PPYP6 sgRNA mostly produced deletions (Fig. S4.5A). Notably, Gag loss-of-function mutations were observed in 70% of PPYP6 sgRNA-induced repaired junctions, but only in 30% of all Myr2 sgRNA-derived mutations (Fig. S4.5B). The majority of Myr2- and PPYP6-derived repair junctions were compatible with classical non-homologous end-joining (C-NHEJ) and alternative end-joining (alt-EJ) repair activities (Fig. 4.4D). C-NHEJ typically leads to small insertion and deletions, while alt-EJ utilizes microhomologies at the DSB site to anneal broken ends, which often results in larger and more complex mutations. Although alt-EJ repair is considered to be a backup pathway in most mammalian cells, we detected between 25-55% alt-EJ compatible junctions when targeting the *gag* gene, supporting our previous conclusions of intrinsically elevated alt-EJ activities in CHO cells [119, 141]. Among the alt-EJ repair junctions, some could be uniquely attributed to the microhomology-mediated end-joining (MMEJ) or the synthesis-dependent microhomology-mediated end-joining (SD-MMEJ) alt-EJ subpathways, while others were consistent with both MMEJ and SD-MMEJ repair [122, 123] (Fig. S4.5D). Interestingly, approximately 10% of all analyzed repair junctions contained either insertions templated from other ERV loci or from the same ERV locus but using a distant sequence, while others manifested apparent duplications devoid of microhomologies, as mediated by alt-EJ mechanisms. All of these latter junctions are consistent with homology-directed repair activities at Myr2- and PPYP6 target sites following CRISPR cleavage (Fig. 4.4D). Thus, HR-mediated gene conversion might indeed have caused the multiple occurrences of certain mutations.

Next, we assessed whether mutations occurred more frequently in some type-C ERV clusters, indicating a preferential cleavage of certain ERV loci. As expected, mutations

124

associated uniquely with clusters of group 1, but not of group 2, confirming sgRNA specificity for group 1 only (Figs. 4.4E and 4.4F). The majority of mutations located within the most abundant Myr or PPYP clusters, which presumably represent the actively transcribed and hence expressed ERVs. Additional mutations were observed in other clusters, which contained a Myr2 or PPYP6 sgRNA recognition sites adjacent to a PAM sequence (Figs. 4.4E and 4.4F, bold font). Surprisingly, we also witnessed CRISPR cleavage in Myr and PPYP clusters containing a one base pair mismatch to the sgRNA target site, supporting previous reports indicating that CRISPR-Cas9 tolerates small mismatches during target recognition [269] (Figs. 4.4E and 4.4F, normal font).

4.4.4 Identification of a unique viral particle-producing ERV locus in CHO-K1 cells

The Sanger chromatograms as well as the read frequencies of *gag* mutations observed during RNA and targeted DNA amplicon sequencing corroborated the assumption that a single group 1 type-C ERV locus is transcribed, and may therefore mediate viral particle production by CHO cells. To further substantiate this assumption, the genome of the E10 clone was fully sequenced using a PacBio approach, so as to obtain reads sufficiently long for the unambiguous determination of ERV integration sites. This clone was selected as it appeared to contain only a single mutated ERV, allowing to correlate its unique mutation at the RNA level with a potentially unique genomic locus (Fig. 4.4C). Analysis of the E10 clone genome sequence led to the identification of a single ERV locus bearing the mutation detected at the mRNA level (Figs. S4.6A and S4.6B). The predicted ERV integration site was then validated by PCR amplification and DNA Sanger sequencing using locus-specific primers located outside of the ERV sequence in the parental CHO cell line as well as the deep-sequenced clones. All deep-sequenced clones, which contain CRISPR-derived mutations at the mRNA level, possessed the identical mutation also at this ERV locus, further supporting that this genomic region harbors the expressed type-C ERV element (Fig. S4.6C). Interestingly, this particular ERV integration was found to be hemizygous, as the other allele was devoid of a corresponding ERV integration, and to have occurred into open chromatin between two moderately expressed CHO cell genes.

Next, we assessed whether Gag loss-of-function mutations in this expressed ERV locus may lead to the anticipated inhibition of viral particle budding. Besides the previously

characterized mutated clones, we analyzed in parallel their corresponding bulk-sorted polyclonal populations, as well as clones devoid of detectable mutations in the expressed group 1 ERV sequence (B01 for Myr2, B03 for PPYP6), as additional controls. First, viral particles were extracted from the supernatant of the CHO cell cultures and the amount of type-C viral genomes was quantified by RT-qPCR. Preliminary data suggested that viral particles produced by Gag loss-of-function mutants contain 80% less group 1 type-C genomic viral RNA than control samples, while the amount of group 2 genomic viral RNA remained close to detection limit (data not shown). To substantiate this finding, RNA extracted from the viral particles shed by the D12 (Myr2 sgRNA) and E10 (PPYP6 sgRNA) clones was Illumina sequenced. Remarkably, we observed a more than 250-fold reduction in reads mapping to the group 1 ERV sequence in both D12 and E10 when compared to wild-type CHO cells, while the trace amounts of reads mapping to group 2 remained close to the detection level (compare Figs. 4.5 and 4.2A). This indicated that mutations in the single expressed group 1 ERV sequence that block translation initiation (D12) or introduce a frameshift in the *gag* gene downstream of the PPYP motif (E10) are sufficient to severely reduce the budding of complete viral particles.
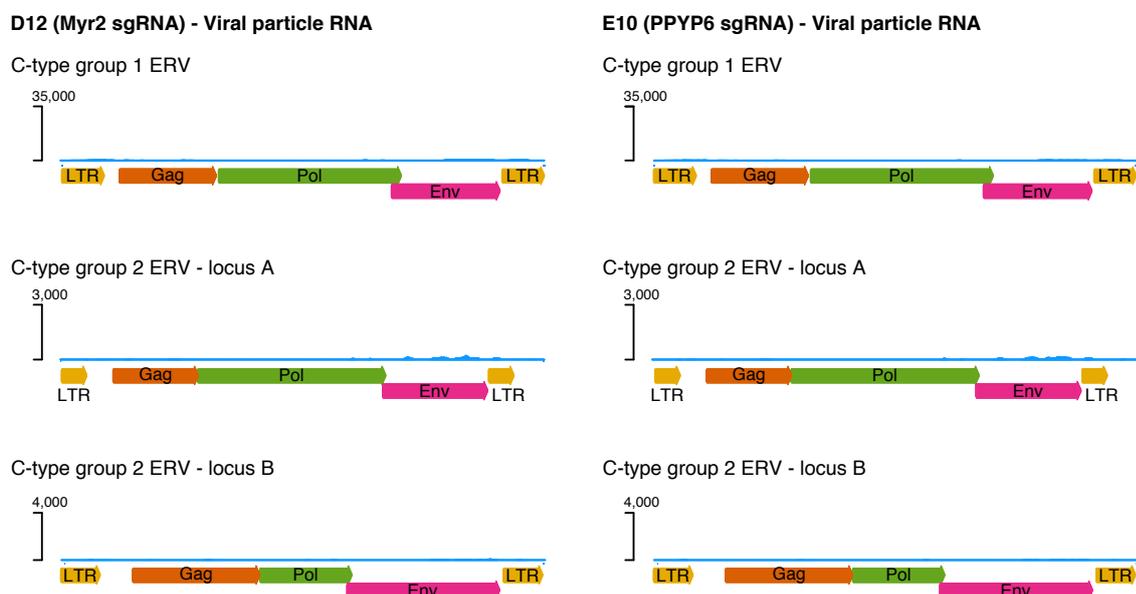
Figure 4.5: Viral particle RNA sequencing of CHO clones mutated in the expressed group 1 type-C ERV sequence. Mapping of viral particle RNA sequencing reads from a Myr2 sgRNA clone (D12, left panels) and a PPYP6 sgRNA clone (E10, right panels) on group 1 consensus sequence and group 2 locus A and locus B, as shown for the wild-type CHO viral particles (Fig. 4.2A). D12 and E10 mutants both contain Gag loss-of-function mutations in the functionally relevant group 1 type-C ERV locus.

126

4.4.5 Characterization of edited CHO cell lines displaying reduced viral budding

Having observed that CRISPR mutagenesis had efficiently inactivated viral particle release, we next tested whether ERV inactivation would affect other CHO cell properties, such as cell growth, cell size and therapeutic protein production. ERV-edited clones were found to proliferate at similar rates as polyclonal populations, wild-type and empty vector-treated cell controls, with a density reaching approximately $12.5*10^6$ cells/ml after five days in culture (Fig. 4.6A). Such a cell density concords with the expected CHO-K1 doubling time of roughly 20h [412]. Although two Myr2 sgRNA clones (C02, D12) and one PPYP6 sgRNA clones (K14) showed slightly modified cell cycle durations, the effect was not statistically significant. In addition, cell sizes tended to be elevated in ERV-edited cells, notably in the C02 clone, but they did not differ significantly when compared to the empty vector control cells (Fig. 4.6B).

Finally, we assessed the capacity of ERV-edited CHO cells to produce therapeutic proteins, a pivotal property of CHO cells for biotechnological use. We used the previously characterized ERV-mutated cells to generate polyclonal populations stably expressing a humanized therapeutic IgG immunoglobulin and quantified IgG secretion during ten-days fed-batch cultures. ERV-edited clones and polyclonal populations expressing the IgG protein demonstrated cell growth and cell viability properties similar to those of wild-type and empty vector control cells, as observed without therapeutic protein expression (Figs. 4.6C and 4.6D). IgG titer in the cell culture supernatants increased over the course of the fed-batch experiment, as expected from the accumulation of the secreted IgG protein, reaching around 300-400 mg/l at the end of the fed-batch for control cells and most ERV-edited cell clones (Fig. 4.6E). Thus, ERV mutagenesis did not globally affect the capability of CHO cells to produce IgG proteins. Interestingly, clone C02 (Myr2 sgRNA) secreted significantly fewer immunoglobulins, likely reflecting its reduced growth and increased cell size, while clones E10 and K03 (both PPYP6 sgRNA) produced 50% more IgG relative to the empty vector control. Overall, this indicated that CHO clones that were exposed to multi-locus ERV editing generally maintain normal CHO characteristics, while some clones, especially those with mutations in the PPYP region, appeared to have acquired a higher metabolic capacity to produce therapeutic proteins. However, this apparently augmented metabolism capacity could not be correlated to a specific ERV mutation type or to the total number of mutations, nor to cell growth or size, suggesting

clone-specific effects.
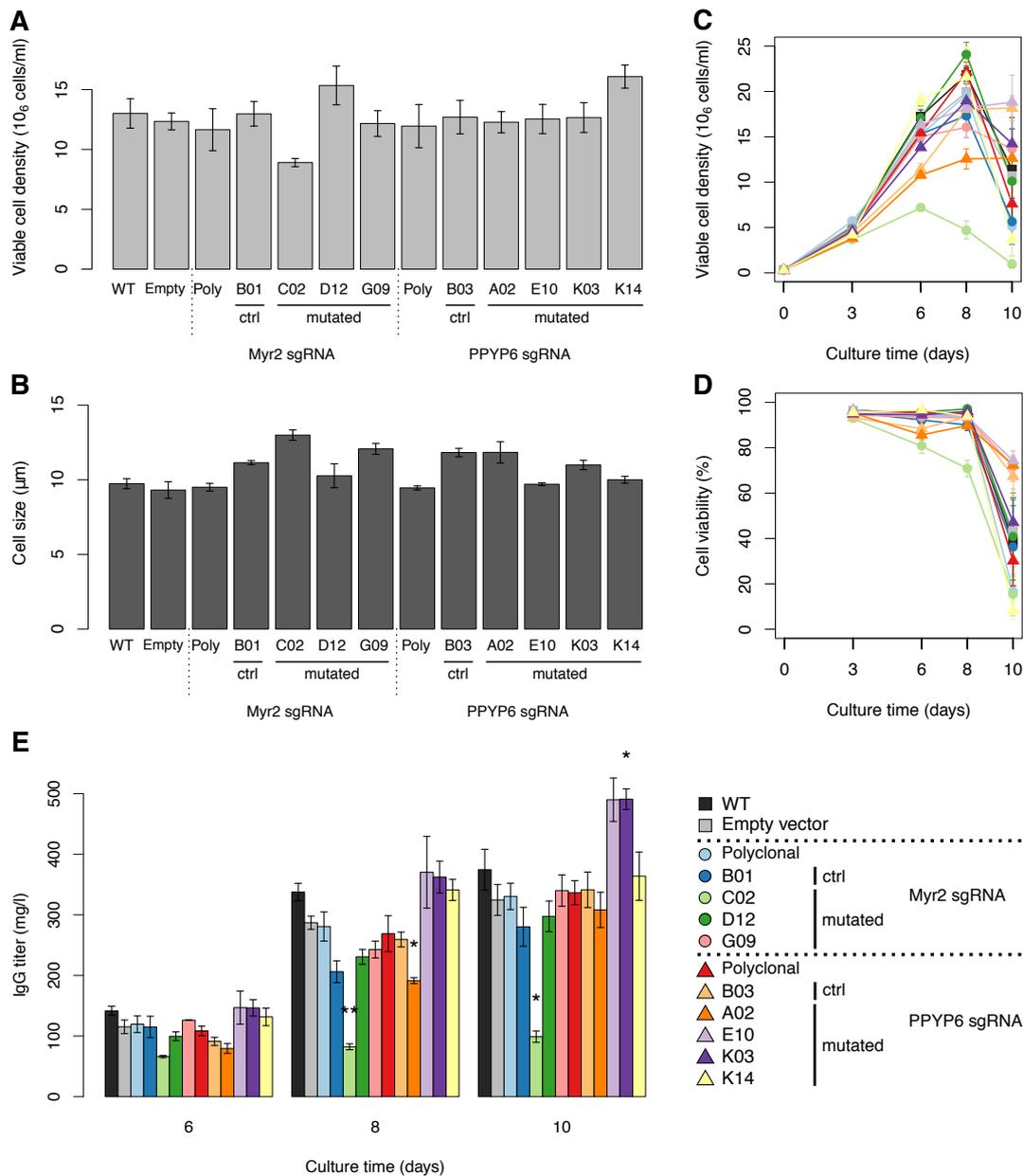
Figure 4.6: Assessment of cell growth, cell size and therapeutic IgG immunoglobulin production in ERV-mutated CHO cells. Viable cell density (A) and cell size (B) was measured in wild-type CHO cells (WT), empty sgRNA vector-treated cells (Empty), bulk-sorted polyclonal CRISPR-treated cells (Poly) as well as in clones containing mutations in the expressed ERV locus (C02, D12, G09, A02, E10, K03 and K14) or not (B01, B03) after five days of culture. The same samples were stably transfected to express an IgG immunoglobulin antibody and assessed for cell density (C), cell viability (D) and IgG production (E) during ten-days fed-batch cultures. Statistical significance relative to the empty vector control was calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction ($n \geq 2$, error bars represent s.e.m, $* P < 0.05$, $** P < 0.01$).

4.5 Discussion

CHO cells are the most widely used expression system for therapeutic proteins, but also a recognized source of adventitious VLPs for more than 40 years [388–391]. Although these particles were never shown to be infectious, their genomic origin and possible evolution remain mostly unknown. Thus, safety concerns have persisted, and ample precautions must be taken when purifying therapeutic proteins. In this study, we first addressed this issue by characterizing CHO endogenous retroviral elements at the genome, transcriptome and viral particle level, showing that CHO cells are able to release intact viral particles loaded with viral RNA genomes of group 1 type-C ERVs. The sequence encodes a full-length open reading frame, thus likely producing functional viral proteins. This finding challenges the only available study on CHO viral particle sequences, published in 1994, in which the authors detected only defective DNA sequences with numerous mutations in the ERV genes [352]. This apparent contradiction may be explained by improved current sequencing methods, allowing the in-depth characterization of expressed and released sequences. Using this updated viral particle RNA sequence, we curtailed the number of possible ERV loci responsible for the expression and release of CHO viral particles to a group of up to 30 well-conserved group 1 type-C ERV sequences in the CHO genome.

Next, we mutated the Myr and PPYP Gag budding motifs of the functionally relevant group 1 type-C ERV sequences using CRISPR-Cas9, so as to attempt to prevent ERV budding. After transient CRISPR-Cas9 expression, 10-15% of the isolated clones contained mutations in the expressed group 1 sequences, some of which causing Gag loss-of-function effects. Having introduced unique mutations into defined ERV sequences allowed us to pinpoint a single genomic ERV locus as the origin of viral type-C particle formation in CHO cells. Most interestingly, site-specific mutagenesis of this particular locus was sufficient to avoid release of viral particles carrying the viral genomic RNA. This indicated that the other ERVs present in the CHO genome may be unable to complement the Gag loss-of-function, nor became reactivated upon CRISPR-Cas9 mutagenesis, at least under the controlled cell culture conditions used in this study.

A common technical challenge for multi-locus genome editing is the presence of extensive DNA damage. This damage may be elicited by the multiple Cas9-induced DSBs, which usually activate p53 signaling and cause cell death [326–328, 413, 414]. The sgRNAs designed in this study were predicted to perfectly recognize roughly 60 distinct group

1 type-C ERV loci in the CHO genome, although only some of them should be transcribed and may thus be preferentially cleaved by Cas9. Indeed, CRISPR-Cas9 treated clones possessed between 1 to 14 different mutation sites following a single transient transfection, suggesting that CHO cells are able to handle the DNA damage response and repair of up to 14 separate DSBs. In comparison to primary cells where sometimes a single DSB break results in cell death [413], immortalized cell lines such as CHO cells typically encounter higher levels of endogenous DNA damage, and they are more likely to cope with and survive multi-loci genome editing, as seen here [415]. However, even in CHO cells, we observed a drop in cell proliferation and/or viability following a rather mild transient treatment with ERV-targeting sgRNAs, which correlated well with the predicted number of target sites. Elevated cytotoxicity might have prevented the isolation of even more highly mutated clones. This would explain why a recent study reporting the isolation of primary porcine cells containing mutations in up to 62 endogenous viral elements required anti-apoptotic treatments to suppress p53-mediated cell death [326].

Another challenge in multi-locus editing is the plurality of repetitive ERV sequences present in the CHO genome that could be used as template for HR repair, which may counteract efficient gene knock-out mediated by C-NHEJ and alt-EJ repair pathways. In CHO cells, HR activity is believed to be rather low compared to other cells [171, 291]. Typically, HR may precisely repair DSBs, but imprecise repair outcomes also occur [416]. Here we found that roughly 10% of the analyzed repair junctions at both sgRNA sites contained HR-compatible signatures, such as templated insertions from other ERV loci. Thus, we hypothesize that HR repair is active and possibly opposes efficient ERV mutagenesis. Although we succeeded in isolating clones with Gag loss-of-function mutations, reducing HR activity could be advantageous in some case, for instance when knock-out frequency needs to be maximized.

The genome editing strategy used in this study aimed to introduce Gag loss-of-function mutations that interfere with proper Gag protein synthesis and thereby prevent ERV budding. As expected, clones mutated in the expressed group 1 type-C ERV sequence showed unchanged mRNA expression levels of group 1 and group 2 ERVs, while being strongly impaired in releasing encapsulated viral RNA. In addition, ERV-mutated clones did not consistently differ in cell growth, cell size or therapeutic protein production compared to control samples. Hence, the differences between clones are likely clone-specific. Clonal variation is a common phenomenon when isolating clones from polyclonal populations and has even been noticed during clone subcloning [417, 418]. Clone-specific variabil-

ity may arise not only from genetic heterogeneity between the clones, for instance due to the acquisition of random and/or CRISPR-derived mutations, but also from stochastic fluctuations in protein expression and/or epigenetic effects [328, 418, 419].

The current study provides a proof-of-principle that the functionally active ERV locus can be selectively mutated using group 1 type-C specific sgRNAs. This offers novel avenues to improve the safety profile of CHO cells and thereby substantially reducing the number of virus inactivation and removal steps needed for viral clearance during biopharmaceutical production. Moreover, the study opens up prospective extensions for further improvements. For example, the discovery of a single ERV locus responsible for ERV expression and particle release in CHO cells would enable to excise the entire 10 kb long proviral genome using two site-specific sgRNAs, as done for HIV-infected human cells [420]. This approach for ERV mutagenesis might reduce the elicited DNA damage response and avoid the accumulation of defective ERV RNAs in the cytoplasm, which may both cause unclear side effects [421]. Furthermore, the current approach already greatly mitigates the risk of infective viral particles in the CHO supernatant by mutagenizing the expressed and particle-forming ERV, but it may not prevent the reactivation of silenced ERVs nor new adventitious infections. An interesting perspective is therefore to introduce mutations that can act preventively. For instance, work in FeLV and HIV showed that non-myristoylated Gag behaves in a dominant-negative manner leading to severe inhibition of viral particle release [422, 423]. Whether the similar Myr and possibly also the PPYP Gag mutations also manifest a dominant-negative phenotype in hamster ERVs, and whether this could prevent new infections in CHO cells, are interesting possibilities that will require further experimental validation.

4.6 Supplementary Data

Table S4.1: Predicted number of ERV target sites for Myr- and PPYP-specific sgRNAs in the CHO-K1 genome.

sgRNA name	sgRNA sequence (5'-3')	PAM sequence[1]	Number of mismatches allowed				Total
			0	1	2	3	
Myr2	TCCTAAGCCTAGAAACTATG	Canonical	59	29	16	26	147
		Non-canonical	-	-	1	16	
Myr4	CATAGTTTCTAGGCTTAGGA	Canonical	33	-	-	9	54
		Non-canonical	-		1	11	
Myr8	GAGTGTTAGGGACAAAGGAG	Canonical	117	30	-	36	218
		Non-canonical	-	-	2	33	
PPYP5	GTTGGTTGATCTATTAACGG	Canonical	61	30	12	5	114
		Non-canonical	-	-	-	6	
PPYP6	GCCACTGCCGCCCCCACCAG	Canonical	55	16	9	36	133
		Non-canonical	1	-	-	16	
PPYP7	GCCCCCACCAGAGGCAGAAG	Canonical	69	65	41	60	283
		Non-canonical	3	3	3	39	
PPYP13	GGCAGTGGCGGATATGGCGG	Canonical	58	16	14	42	142
		Non-canonical	1	2	1	8	
PPYP20	GCTTCTGCCTCTGGTGGGGG	Canonical	70	63	8	47	217
		Non-canonical	3	4	5	17	

---

[1] The canonical PAM sequence of SpCas9 is NGG
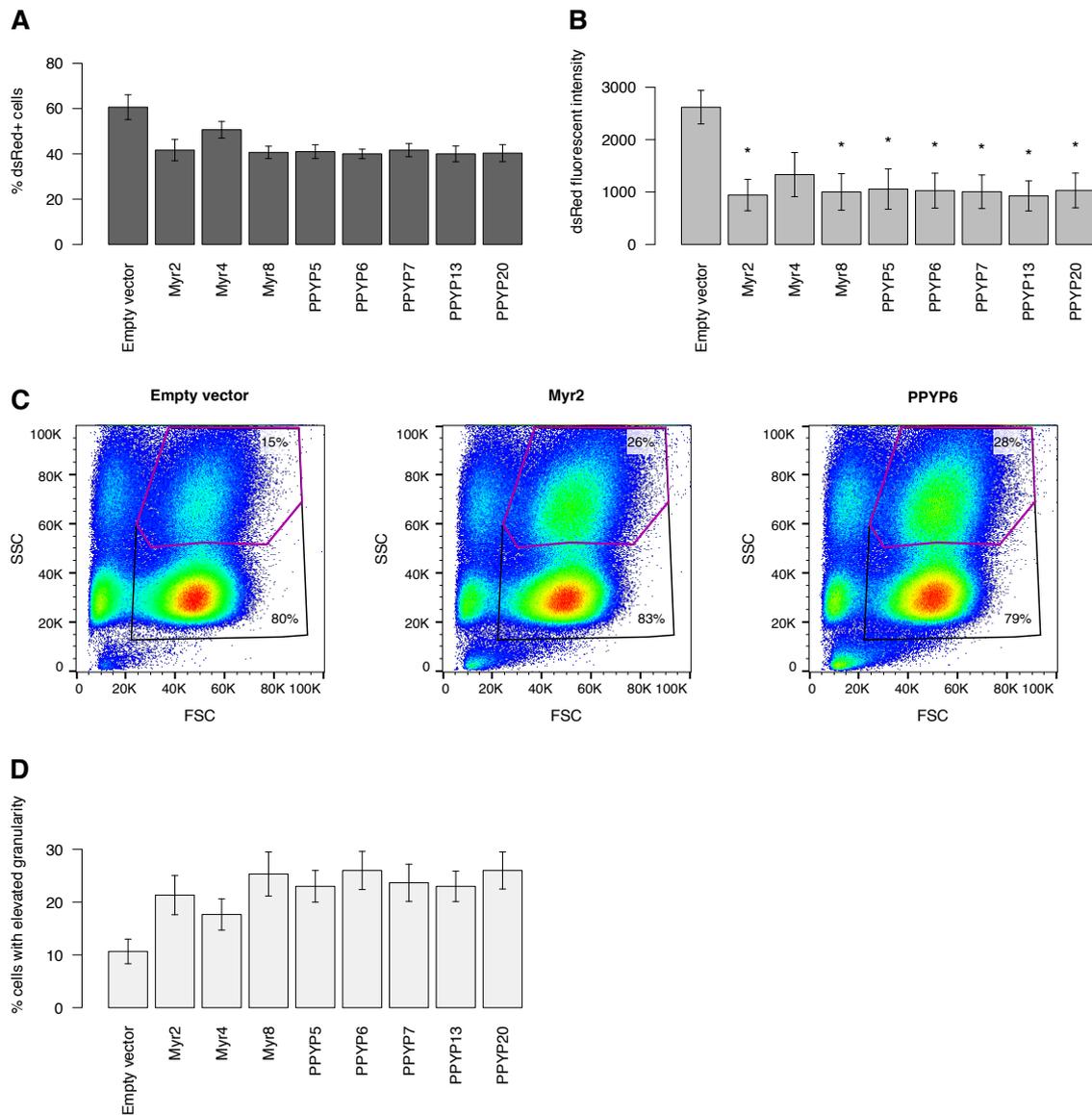
Figure S4.1: Assessment of *gag*-specific sgRNA-mediated CRISPR-Cas9 cleavage by flow cytometry. Analysis of the dsRed positive (dsRed+) cell frequency (A), the dsRed fluorescence intensity (B) and the frequency of high granularity cells (C and D) of CHO cells transfected with CRISPR-Cas9, Myr- or PPYP-specific sgRNAs (Myr2, Myr4, Myr8,PPYP5, PPYP6, PPYP7, PPYP13, PPYP20 sgRNAs) or a non-targeting empty vector control and dsRed transfection control expression plasmids. Panel C shows size (FSC) vs granularity (SSC) flow cytometry density plots of the empty vector-, Myr2 sgRNA- and PPYP6- treated cells. The larger black gate selects for intact non-debris cells while the smaller purple gate marks the CHO cell subpopulation with an elevated granularity level, as quantified in panel D. Statistical significance relative to the empty vector control was calculated using the two-tailed unpaired Student's t-test with Benjamini and Hochberg false discovery rate correction (n = 3, error bars represent s.e.m, * P < 0.05, ** P < 0.01).
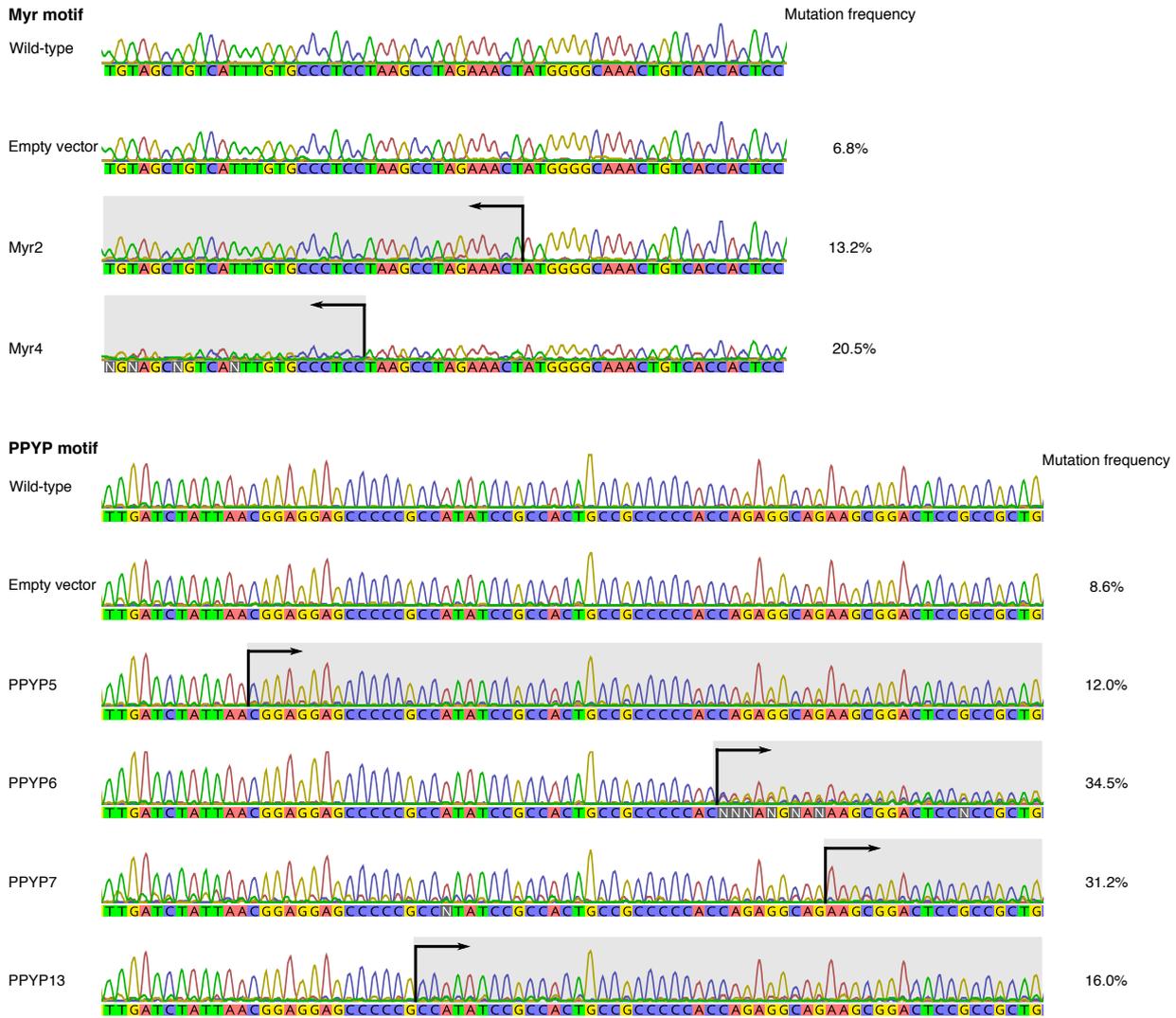
Figure S4.2: Estimation of *gag*-specific sgRNA-mediated CRISPR-Cas9 cleavage efficiency by targeted mRNA sequencing of polyclonal CHO populations. Indel mutation analysis of polyclonal PCR products obtained from reverse-transcribed cellular mRNA of bulk-sorted CRISPR-treated polyclonal populations using the indicated group 1 type-C specific primers. The mutation frequency was estimated by decomposition of the Sanger chromatogram [403]. The predicted mutation frequency relative to the untreated wild-type control sample is shown on the right of the chromatograms. The DSB site for each sgRNA is shown with a black line and the decomposition window, downstream of the DSB site relative to the sequencing direction indicated by an arrow, is shaded in grey.
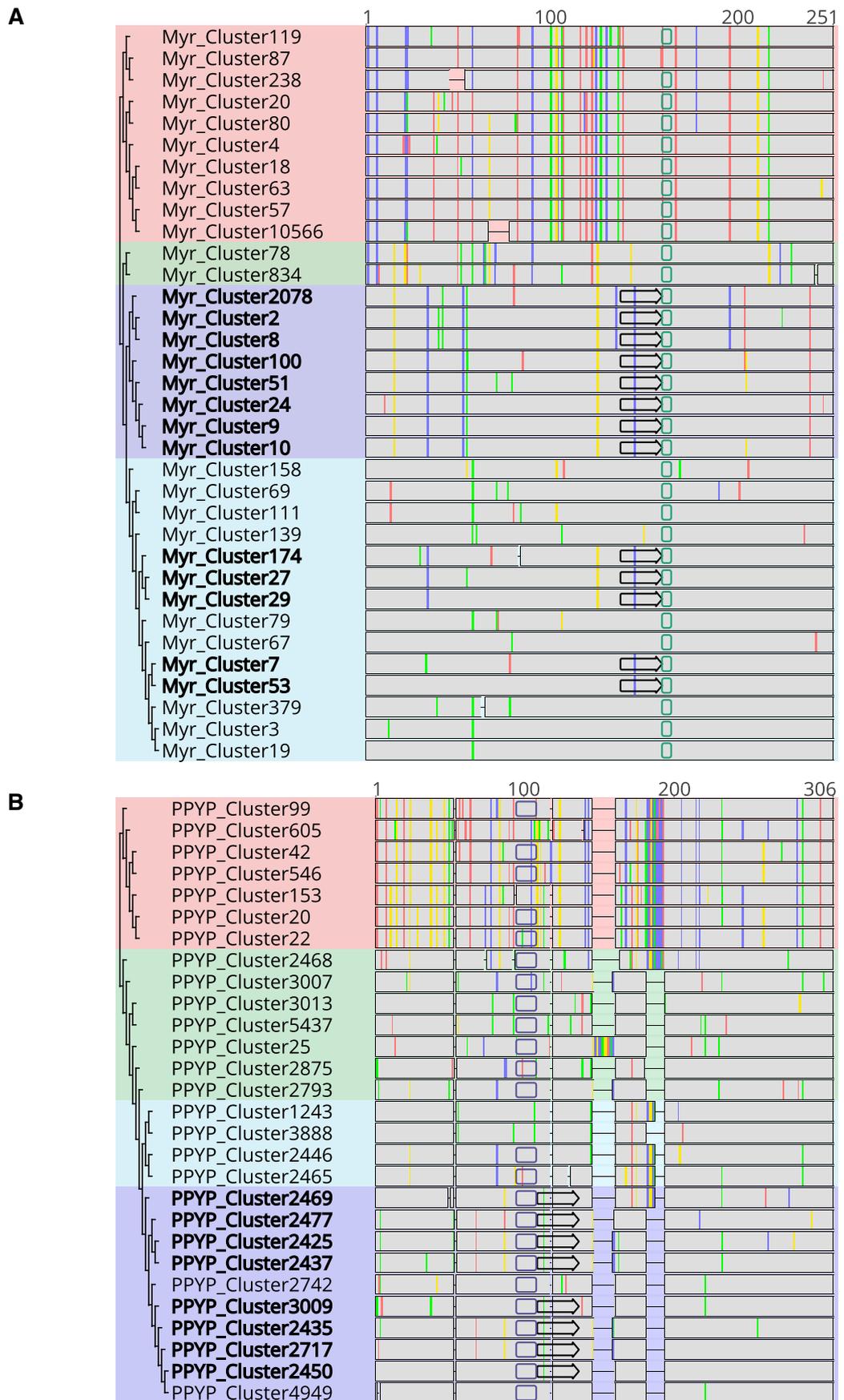
Figure S4.3: (Caption next page.)

Figure S4.3: Wild-type CHO consensus sequences of Myr and PPYP diversity clusters. Cluster sequences of Myr (A) and PPYP (B) flanking regions of deep-sequenced wild-type CHO cells. Colors correspond to the phylogenetic groups depicted in Figure 4.4A and 4.4B. Myr and PPYP clusters containing a sgRNA recognition site (black outlined arrow) with an adjacent PAM sequence are written in bold letters. Myr and PPYP motifs are indicated with turquoise and purple outlined boxes, respectively. The higher sequence complexity of the PPYP flanking region relative to the Myr flanking region is illustrated by missing sequences and colored lines depicting deletions or insertions and single nucleotide variants, respectively.
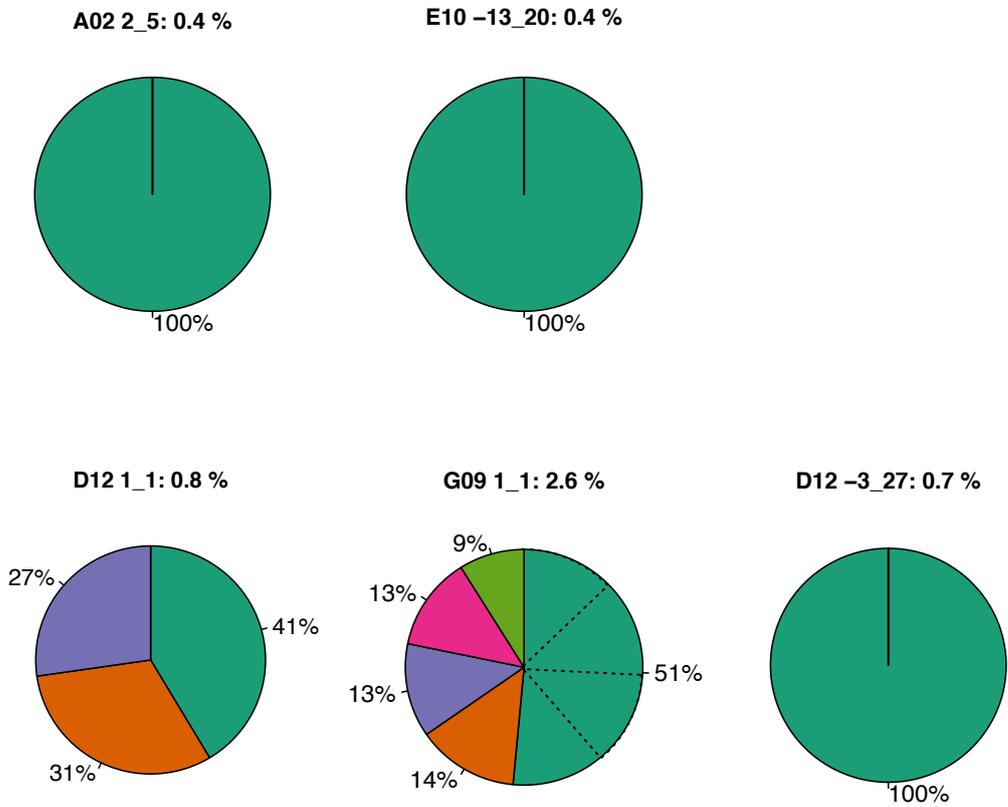
Figure S4.4: Characterization of ERV locus-specific mutations and their frequencies within clonal populations. Analysis of Illumina raw reads of mutations detected at normal (0.2-0.4%) or high (> 0.4%) read frequencies in different clones. Pie charts represent number and frequency of identified groups with identical CRISPR-derived mutation but distinct mutation flanking sequences (e.g. in D12_1_1 and G09_1_1). Dotted lines indicate the number of predicted ERV loci that could not be distinguished based on their flanking sequences.
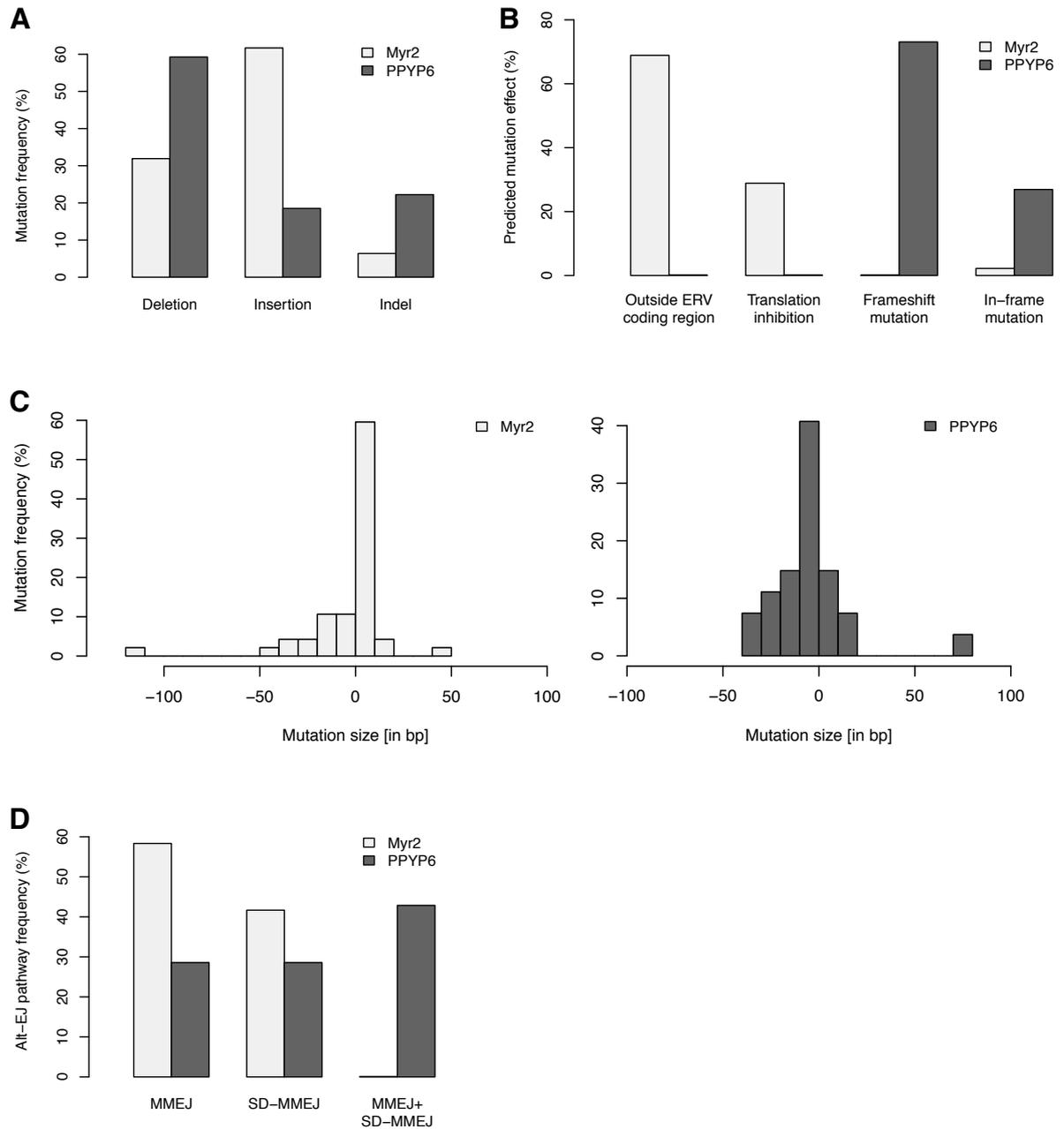
Figure S4.5: Characterization of Myr2 sgRNA- and PPYP6 sgRNA-mediated mutations and repair junctions. 47 Myr2- and 27 PPYP6-derived repair junctions were analyzed for sgRNA specific mutation signatures, including the elicited mutation type (Deletion, Insertion, Indel) (A), the mutation effect on Gag and ERV function (Outside ERV coding region, Translational inhibition, Frameshift mutation, In-frame mutation) (B), the mutation size distribution (C) and MMEJ and SD-MMEJ alt-EJ repair pathway activities. Indel mutations are defined in this figure as deletions coupled to insertions. Repair junctions compatible with both MMEJ and SD-MMEJ repair mechanisms are classified as "MMEJ + SD-MMEJ". Repair junctions were obtained from both Sanger mRNA and Illumina DNA deep sequencing.
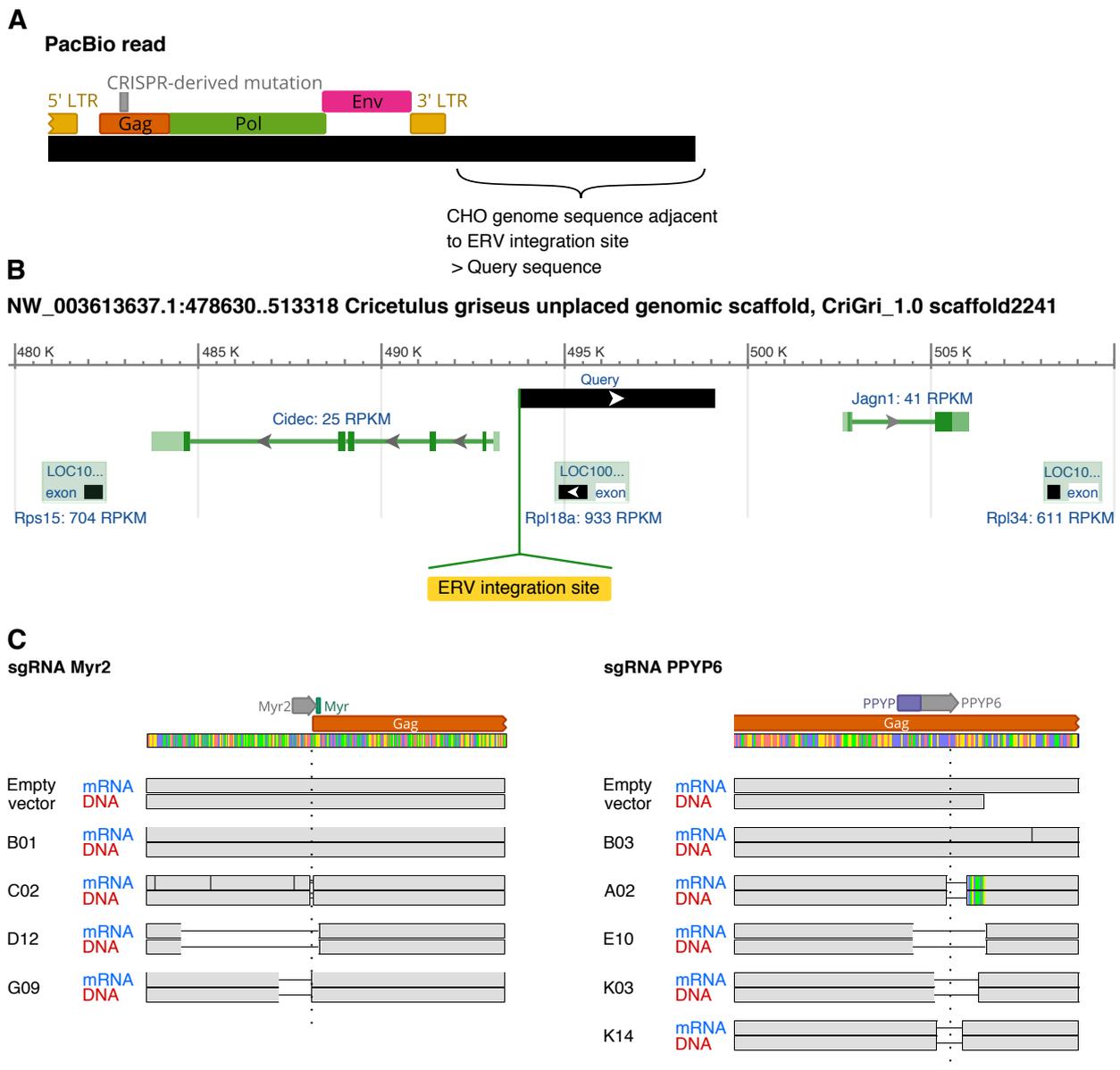
Figure S4.6: Identification of a unique functionally active group 1 type-C ERV locus. (A) Schematic representation of a 15 kb PacBio read obtained following whole genome sequencing of the E10 (PPYP6 sgRNA) clone. The read contains full-length *gag*, *pol*, *env* and 3' LTR sequences as well as the E10-specific CRISPR-mutation in the *gag* gene and extends into the CHO genome. (B) Alignment of the PacBio CHO genome-specific sequence against the publicly available NCBI CHO genome. The NCBI scaffold identifier is shown on top. The predicted group 1 type-C ERV integration site is highlighted in yellow. The genomic region surrounding the ERV integration site contains two protein-coding genes (Cidec, Jagn1) as well as three pseudogenes (Rps15, Rpl18a, Rpl34; shown with green backgrounds), as annotated by the NCBI. Cidec (cell death inducing DFFA like effector c) encodes for a lipid droplet protein involved in lipid metabolism [424], Jagn1 (jagunal homolog 1) encodes for an endoplasmatic reticulum protein involved in the early secretory pathway [425] and Rps15, Rpl18a, Rpl34 encode for ribosomal proteins. The predicted mRNA expression levels for each gene are estimated by RNA sequencing data and expressed as Reads Per Kilobase Million (RPKM). Continued next page.

Figure S4.6: (C) Sanger sequencing results of the Myr2 and PPYP6 sgRNA flanking regions. Sanger sequencing was performed on PCR amplicons obtained from total cellular mRNA using group 1 specific primers (in blue) or genomic DNA using primers specific to the expressed group 1 type-C ERV locus (in red). Clones C02, D12, G09, A02, E10, K03, K14 contain mutations in the functionally active group 1 type-C ERV locus, but clones (B01 and B03) as well as the empty vector controls do not. The predicted Myr2 and PPYP6 DSB sites are marked with a dotted line.

Table S4.2: Sequences of the sgRNAs and corresponding cloning primers used in this study. Output from Zifit software (http://zifit.partners.org/ZiFiT/ChoiceMenu.aspx) for cloning into MLM3636 sgRNA expression plasmid

sgRNA name	Orientation	5'-3 targetsite (without PAM)	Addition of G at 5' end for better U6 expression	Oligo 1	Oligo 2
Myr2	Forward strand	TCCTAAGCCTAGAAACTATG	GTCCTAAGCCTAGAAACTATG	ACACCGTCCTAAGCCTAGAAACTATGG	AAAAACCATAGTTTCTAGGCTTAGGACG
Myr4	Reverse strand	CATAGTTTCTAGGCTTAGGA	GCATAGTTTCTAGGCTTAGGA	ACACCGCATAGTTTCTAGGCTTAGGAG	AAAACTCCTAAGCCTAGAAACTATGCG
Myr8	Reverse strand	GAGTGTTAGGGACAAAGGAG	–	ACACCGAGTGTTAGGGACAAAGGAGG	AAAACCTCCTTTGTCCCTAACACTCG
PPYP5	Forward strand	GTTGGTTGATCTATTAACGG	–	ACACCGTTGGTTGATCTATTAACGGG	AAAACCCGTTAATAGATCAACCAACG
PPYP6	Forward strand	GCCACTGCCGCCCCACCAG	–	ACACCGCCACTGCCGCCCCACCAGG	AAAACCTGGTGGGGCGGCAGTGGCG
PPYP7	Forward strand	GCCCCCACCAGAGGCAGAAG	–	ACACCGCCCCCACCAGAGGCAGAAGG	AAAACCTTCTGCCTCTGGTGGGGGCG
PPYP13	Reverse strand	GGCAGTGGCGGATATGGCGG	–	ACACCGGCAGTGGCGGATATGGCGGG	AAAACCCGCCATATCCGCCACTGCCG
PPYP20	Reverse strand	GCTTCTGCCTCTGGTGGGGG	–	ACACCGCTTCTGCCTCTGGTGGGGGG	AAAACCCCCACCAGAGGCAGAAGCG

140

Table S4.3: Sequences of the PCR and Illumina sequencing primers used to characterize ERV genomic loci of edited CHO cells.

Primer Name	Full sequence (5'-3')	Illumina Adapter	Spacer	Gene-specific Primer	Primer ratio according to expected % of ERVs (total =1)
Myr Forward Primers					
Myr_Fa_3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCGCTTGAAGGATTTGCAATC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	ACC	GCTTGA**A**GGATTTG**C**AATC	0.15
Myr_Fb_0	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCTTGAGGGATTTGCAATC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG		GCTTGA**G**GGATTTG**C**AATC	0.2
Myr_Fb_1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCTTGAGGGATTTGCAATC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	T	GCTTGA**G**GGATTTG**C**AATC	0.2
Myr_Fb_2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGCTTGAGGGATTTGCAATC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	TT	GCTTGA**G**GGATTTG**C**AATC	0.2
Myr_Fb_3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACTGCTTGAGGGATTTGCAATC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	ACT	GCTTGA**G**GGATTTG**C**AATC	0.2
Myr_Fc_2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGCTTGAGGGATTTGTAATC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	GT	GCTTGA**G**GGATTTG**T**AATC	0.05
					1
Myr Reverse Primers					
Myr_R_0	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAAAGAGTAATCCATTTGCG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG		ACAAAGAGTAATCCATTTGCG	0.25
Myr_R_1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACAAAGAGTAATCCATTTGCG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	G	ACAAAGAGTAATCCATTTGCG	0.25
Myr_R_2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGACAAAGAGTAATCCATTTGCG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	CG	ACAAAGAGTAATCCATTTGCG	0.25
Myr_R_3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAGACAAAGAGTAATCCATTTGCG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	AAG	ACAAAGAGTAATCCATTTGCG	0.25
					1
PPYP Forward Primers					
PPYP_Fa_0	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACTCCAGCCTTTACCCTAC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG		ACTCCAGCCTTTACCCTAC	0.1
PPYP_Fa_1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAACTCCAGCCTTTACCCTAC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	A	A**C**TCCAGCCTTTACCCTAC	0.1
PPYP_Fb_0	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATTCCAACCTTTACCCTAC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG		A**AT**TCCA**AC**CTTTACCCTAC	0.2
PPYP_Fb_1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAATTCCAACCTTTACCCTAC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	G	A**AT**TCCA**AC**CTTTACCCTAC	0.2
PPYP_Fb_2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGAATTCCAACCTTTACCCTAC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	TG	A**AT**TCCA**AC**CTTTACCCTAC	0.2
PPYP_Fb_3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTATTCCAACCTTTACCCTAC	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	CTT	A**AT**TCCA**AC**CTTTACCCTAC	0.2
					1
PPYP Reverse Primers					
PPYP_Ra_1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGGTCTGATGCTGAGAATG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	G	GGTC**T**GATGCTGAGAATG	0.04
PPYP_Rb_0	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTCCGATGCTGAGAATG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG		GGTC**C**GATGCTGAGAATG	0.24
PPYP_Rb_1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGGTCCGATGCTGAGAATG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	T	GGTC**C**GATGCTGAGAATG	0.24
PPYP_Rb_2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGGTCCGATGCTGAGAATG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	GT	GGTC**C**GATGCTGAGAATG	0.24
PPYP_Rb_3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAGGGTCCGATGCTGAGAATG	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	AAG	GGTC**C**GATGCTGAGAATG	0.24
					1

Table S4.4: Sequences of the PCR and qPCR primers used in this study to characterize and validate ERV loci and expression

Amplification type	Primer 1	Primer 2
End-point PCR		
Locus group 1 type-C ERV validation1	CTCTGGTTCTTGCCTGCTGAGCT	TGGTCAATGTATATGAGGCGCT
Locus group 1 type-C ERV validation2	CTCTGGTTCTTGCCTGCTGAGCT	TAAGCCATTGGTGAAGGGTCA
Locus group 1 type-C ERV validation3	CTCTGGTTCTTGCCTGCTGAGCT	TGACGATATAAGCCACTTGA
Locus group 1 type-C ERV validation4	TTTTCTGGTGCCCTCTTGCCTGG	TAAGCCATTGGTGAAGGGTCA
Locus group 1 type-C ERV validation without ERV	CTCTGGTTCTTGCCTGCTGAGCT	TTGTGGAGCTGTGTGAGTGGTGG
qPCR		
Group 1 type-C ERV specific LTR	GGGAATTGAGTCTGCTGTACCA	ACAGAGTCTTTCAAATGAGGCG
Group 1 type-C ERV specific Gag	TGACGATATAAGCCACTTGA	ACCCCCAGACTATATTCCAGATA
Group 1 type-C ERV specific Env	CTATGTGCTGCCCTCAAGGA	GCCTCTCCCTAAGTTTGGCC
Group 2 type-C ERV specific	GAATAAAAGGTCAGGGCGTTGG	CTGACTTGGCTCTATCTTGGGT
Reference GAPDH	GCGACTTCAACAGTGACTCCCA	TGAGGTCCACCACTCTGTTGCT

Table S4.5: Detection of CRISPR-mediated mutations in expressed type-C ERV sequences cloned into plasmid vectors.

Sample	Analyzed sequence	Mutated sequences	Mutation frequency	Loss-of-function mutation frequency[1]
Myr2 sgRNA	12	2	17%	50%
PPYP6 sgRNA	56	4	7%	75%
PPYP13 sgRNA	12	1	8%	0%
Total	80	7	9%	42%

---

[1] Includes translation inhibition and frameshift mutations and is expressed relative to the number of mutated sequences.

Table S4.6: Sequence analysis of the junction sequences of edited ERVs.

Table S4.7: Sequence analysis of ERV-mutated CHO-K1 mRNA junction amplicons cloned into plasmid vectors.

Sample[1]		Sequence[2]	Features[3]	Mutation type[4]	Repair pathway[5]	Pattern score[6]	FORECasT[7]
Myr2 sgRNA (n=2)							
4_5	Genomic	5'-CTGTCATTTGTGCCCTCCTAAGCCTAGAAAACT–ATGGGGCAAA CTGTCACCACTCCTTTGTCCC–3'					
	Junction	5'-CTGTCATTTGTGCCCTCCTAAGCCTAGAAAACTTATGGGGCAAA CTGTCACCACTCCTTTGTCCC–3'	1 bp insertion, 1 bp MH	Outside ERV coding region	C-NHEJ[8]	-	17.8%
4_9	Genomic	5'-CTGTCATTTGTGCCCTCCTAAGCCTAGAAAACTATGGGGCAAAC TGTCACCACTCCTTTGTCCC–3'	12 bp deletion, 5 bp MH	Translation inhibition	MMEJ	329.4	13.4%
	Junction	5'-CTGTCATTTGTGCCCTCCTAAGCCTAG----------AAAC TGTCACCACTCCTTTGTCCC–3'					
PPYP6 sgRNA (n=4)							
1_6	Genomic	5'-CCTTTGATTCCTCCCAACCCCCTTCCCATTCCAACCTTTACC CTACCGTGATGAAAGACACTAAGGCTAAAGAAAGAAGACACCTAA GGTACTCCCTCCGGAGAAGACCAGTTGGTTGATCTATTAACGGAG GAGCCCCGGCCATATCCGCCATGCCGCCCCCACCAGAGGCAGAAGC GGACTCCGCGCTGCCTTGGCGGAAGCGGCCCCTGACCCTT–3'	1: 27 bp deletion, 7 bp MH (89 bp upstream) 2: Deletion + Insertion (Replacement of 29 bp with 2 bp; net = 27 bp deletion), 15 bp MH	In-frame mutation	1: SD-MMEJ (loop-out) 2: SD-MMEJ (loop-out)	-	-
	Junction	5'-CCTTTGATTCCTCCCAACCCCCTTCCCATTCCAACCTTTACC CTACCGTGATGAAAGACACTAAGGCTAAAGAAAGAAGACACCTAA GGTACTCCCTCCGGAGAAGACC----------------------- CCCCGGCCATATCCGCCATA----------------------- ----TCCGCGCTGCCTTGGCGGAAGCGGCCCCTGACCCTT–3'					
3_1	Genomic	5'-GCCCCCGGCCATATCCGCCACTGCCGCCGCCCCCAC–CAGAGGCAGA AGCGGACTCCGCCGCTGCCTTG–3'	1 bp insertion, 1 bp MH	Frameshift mutation	C-NHEJ[8]	-	3.1%
	Junction	5'-GCCCCCGGCCATATCCGCCACTGCCGCCCCCACCCAGAGGCAGA AGCGGACTCCGCCGCTGCCTTG–3'					

144

Table S4.7: Continued

3_10	Genomic	5′–TCCGGGAGAAGACCAGTTGGTTGATCTATTAACGGAGGAGCCCCGCCCGCTGCCTTGGCGGAAGCGGCCCC–3′			
	Junction	5′–TCCGGGAGAAGACCAGTTGGTTGATCTATTAACGGAGGAGCCCCGCCCGCTGCCTGCC–––––CCAGAGGCAGAAGCGGACTCCGCCGCTGCCTTGGCGGAAGCGGCCCC–3′	4 bp deletion, 2 bp MH or 5 bp MH (22 bp upstream)	Frameshift mutation	MMEJ or SD-MMEJ (loop-out)	NA	1.2%
8_6	Genomic	5′–ATGGATCCTGGACCACACGGCATCCCGATCAAGTGGCTTATATCGTCACTTGGGAGGCTTTGGTTCAGGACCCCTCCTTCCCCCTCTAACTCCTTTCTTACATCCCAAGGGCCCCTCTCTTCGGCCCCTACACCTCCCACTCCTTCGCTCCAACCGACCCATTCCTTCGGCCCCTTCCAATTCCAACCTTTACCCTACCGTTGATTCCTCCCAACCCCCCTTCCCATTCCAACCCCCCTTCCAACCTTTACCCTACCGTGATGAAAGACACTAAGGCTAAAGAAAAGAAAGACACCTAAGGTACTCCCTCCGGGAGAAGACCAGTTGGTTGATCTATTAACGGAGGAGCCCCCGCCATATCCGCCACTGCCGCCGCCCGCTGCCTGGCGG–3′				
	Junction	5′–ATGGATCCTGGACCACACGGCATCCCGATCAAGTGGCTTATATCGTCACTTGGGAGGCTTTGGTTCAGGACCCCTCCTTCCCCCTCTAACTCCTTTCTTACATCCCAAGGGCCCCTCTCTTCGGCCCCTACACCTCCCACTCCTTCGCTCCAACCGACCCATTCCTTCGGCCCCTTCCAATTCCAACCTTTACCCTACCGTGATGAAAGACACTAAGGCTAAAGAAAAGAAAGACACCTAAGGTACTCCCTCCGGGAGAAGACCAGTTGGTTGATCTATTAACGGAGGAGCCCCGCCATATCCGCCACTGCCGCCCCGCCGCTGCCTTGGCGG–3′	Deletion + Insertion; (Replacement of 9 bp with 26 bp; net= 17 bp insertion); templated inversed insertion possibly from same or other ERV allele[9]	Frameshift mutation	HR	-	-

PPYP13 sgRNA (n=1)

| PPYP13 (5_8) | Genomic | 5′–TTAACGGAGGAGCCCCGCCATATCCGCCACTGCCGCCCCCACCAGAGGCAGAAGCGGACTCCGCCGCTTGGCGGAAGCGGCCCC––––––––TGACCCTTCACCAATGGCTTA–3′ | 12 bp insertion[10], likely from another ERV allele | In-frame mutation | HR | - | - |
| | Junction | 5′–TTAACGGAGGAGCCCCGCCATATCCGCCACTGCCGCCCCCACCAGAGGCAGAAGCGGACTCCGCCGCTTGGCGGAAGCGGCCCCAGATCCACCACCTGACCCTTCACCAATGGCTTA–3′ | | | | |

145

Table S4.7: Continued

[1] The table shows mRNA Sanger sequencing data of the expressed ERV repair junctions of CHO-K1 cells treated with wild-type Cas9 nuclease and various sgRNAs (Myr2, PPYP6 and PPYP13). The sequences are derived from Sanger sequencing of cDNA PCR amplicons cloned into plasmid vectors.

[2] Predicted blunt-ended DSB sites induced by the various sgRNAs and the wild-type Cas9 nuclease are highlighted in yellow, PAM site are shown in bold blue, and the ATG translation initiation codon is depicted in brown letters. Myr and PPYP target motifs are highlighted in turquois and purple, respectively. Pre-existing microhomologies (MH) of the microhomology-mediated end-joining (MMEJ) repair mechanism are shown in bold green, while *de novo* MH of the synthesis-dependent microhomology-mediated end-joining (SD-MMEJ) mechanism are underlined with a double line or highlighted in cyan. Inserted bases are represented in bold red letters, deleted bases with a "-" sign, and replacements in bold black.

[3] Size of mutation and MH length (in bp). The distance between priming site and the break site for *de novo* MH are shown in parenthesis.

[4] ERV mutation types include in-frame mutations, out-of-frame mutations, translation inhibition (mutation of the ATG translation initiation codon) or mutations locating outside of the ERV coding region. Out-of-frame mutations and translation inhibition are likely, while in-frame mutations and mutations outside of the coding region are less likely to influence ERV expression and VLP formation.

[5] Most probable DSB repair mechanism based on manual junction analysis. Possible repair mechanisms include C-NHEJ, MMEJ, SD-MMEJ (snap-back), SD-MMEJ (loop-out), single strand annealing (SSA), homologous recombination (HR), and unknown. For snap-back SD-MMEJ mechanism, *de novo* priming sites are inverted repeats, while loop-out SD-MMEJ mechanisms uses priming sites with direct repeats (Khodaverdian et al., 2017). If the observed junction sequence is compatible with more than one mechanism and both appear equally likely, all potential pathways are listed. Junctions were verified for homologies at break site and templated insertions (SD-MMEJ) using program described in (Schimmel et al., 2017).

[6] Score of each repair pattern according to the MH size and the deletion length. Pattern score was calculated using the RGenome "Microhomology-Predictor" tool (http://www.rgenome.net/mich-calculator/) described in (Bae et al., 2014). The higher the score, the more likely the predicted mutation should be observed. The pattern score is only valid for repair junctions showing MHs at the break site (MMEJ-mediated repair).

[7] Predicted frequencies of CRISPR-Cas9 editing outcomes using the online tool FORECasT (Favoured Outcomes of Repair Events at Cas9 Targets; https://partslab.sanger.ac.uk/FORECasT) as described in (Allen et al., 2018). The higher the frequency, the more junctions are expected to contain the predicted mutation pattern. Only the frequencies of the predicted ten most frequent mutations are listed.

[8] Frequent 1 bp insertions consisting of a duplication of the 4th nucleotide were also observed previously (Lemos et al., 2018).

[9] DNA template sequence for insertion located 290 bp upstream.

[10] DNA template sequence for insertion located 71 bp downstream.

146

Table S4.8: Sequence analysis of the expressed mRNA ERV sequences of mutated CHO-K1 clones.

Clone[1]		Sequence[2]	Features[3]	Mutation type[4]	Repair pathway[5]	Pattern score[6]	FORECasT[7]
Myr2 sgRNA (n=18)							
C02	Genomic	5'-TGTCATTTGTGCCCTCCTAAGCCTAGAAACTATGGGGCAAAACTGTC ACCACTCCTTTGTCCCTAACACTCTCCCACTGGAA-3'	2 bp deletion, 1 bp MH	Translation inhibition	C-NHEJ	-	2.6%
	Junction	5'-TGTCATTTGTGCCCTCCTAAGCCTAGAAAC--TGGGGCAAAACTGTC ACCACTCCTTTGTCCCTAACACTCTCCCACTGGAA-3'					
D12	Genomic	5'-CGACTCTCTCTCAATTCCT-75bp-GAAAACTATGGGGCAAAACTGTC ACCACTCCTTTGT-3'	114 bp deletion, 3 bp MH	Translation inhibition	MMEJ	NA	-
	Junction	5'-CGACTCTCTCTCAA------95bp----------------ACTGTC ACCACTCCTTTGT-3'					
G09	Genomic	5'-TCTTTGTCTTGTGAGCTGTCATTTGTGCCCTCCTAAGCCTAGAAACT ATGGGGCAAAACTGTCACCACTCCTTTGTCCCTAACACTCTCCCACTGGA AAGATGTACAGGAATATGCTCATAACCAATCT-3'	27 bp deletion, 2 bp MH	Outside ERV coding region	MMEJ	51.8	-
	Junction	5'-TCTTTGTCTTGTGAGCTGTC------------------------- ATGGGGCAAAACTGTCACCACTCCTTTGTCCCTAACACTCTCCCACTGGA AAGATGTACAGGAATATGCTCATAACCAATCT-3'					
H02	Genomic	5'-AGCTGTCATTTGTGCCCTCCTAAGCCTAGAAACTATGGGGCAAAACT GTCACCACTCCTTTGTCCC-3'	3 bp deletion, no MH	Outside ERV coding region	C-NHEJ	-	-
	Junction	5'-AGCTGTCATTTGTGCCCTCCTAAGCCTAGA---TATGGGGCAAAACT GTCACCACTCCTTTGTCCC-3'					
A04 (n=14)	Genomic	5'-CTCCTAAGCCTAGAAACT-ATGGGGCAAAACTGTCACCACTCC-3'	1 bp insertion, 1 bp MH	Outside ERV coding region	C-NHEJ[8]	-	17.8%
	Junction	5'-CTCCTAAGCCTAGAAACTTATGGGGCAAAACTGTCACCACTCC-3'					

Table S4.8: Continued

PPYP6 sgRNA (n=14)						
A02	Genomic	5′-CCCC**CGGCCATATCCG**CCACTGCCGCGCCCCCC*CACC*AGA*GG*CAGAAGCGG* *ACT*CCGCCGCTGCCTTGGCGGAAGC-3′	Deletion + Insertion (Replacement of 20 bp with 10 bp; net=10 bp deletion); inverted templated insertion from three possible ERV alleles	Frameshift mutation	HR	-	-
	Junction	5′-CCCC**CGGCCATATCCG**CCACTGCCGCGCCCCC*ACTGCTTCTG*---------- --TCCGCCGCTGCCTTGGCGGAAGC-3′					
A07	Genomic	5′-CCCC**CGGCCATATCCG**CCACTGCCGCGCCCCC**CACC**AGA**GG****CA**GAGAAGCGG ACTCCGCCGCTGCCTTGGC-3′	7 bp deletion, 2 bp MH flanking DSB	Frameshift mutation	Unknown[9]	-	1.1%
	Junction	5′-CCCC**CGGCCATATCCG**CCACTGCCGCGCCCCC**CA**------**CA**GAGAAGCGG ACTCCGCCGCTGCCTTGGC-3′					
B11 (n=3)	Genomic	5′-CCCC**CGGCCATATCCG**CCACTGCCGCGCCCCCAC-C**A**GA**GG**CAGAGAAGCG GACTCCGCCGCTGCCTTGGC-3′	1 bp insertion, 1 bp MH	Frameshift mutation	C-NHEJ[8]	-	3.1%
	Junction	5′-CCCC**CGGCCATATCCG**CCACTGCCGCGCCCCCACC**A**GA**GG**CAGAGAAGCG GACTCCGCCGCTGCCTTGGC-3′					
D08	Genomic	5′-GGAGAAGACCAGTTGGTTGATCTATTAACGGAGGAGCCCCC**CGCCA TATCCG**CCACTGCCGCGCCCCC**CACC**AGA**GG****CA**GAGAAGCGGACTCCGCC-3′	9 bp deletion, 2 bp MH	In-frame mutation	MMEJ	191.4	2%
	Junction	5′-GGAGAAGACCAGTTGGTTGATCTATTAACGGAGGAGCCCCC**CGCCA TATCCG**CCACTGCCGCGCCCCC--------**CA**GAGAAGCGGACTCCGCC-3′					
E10	Genomic	5′-AGCCCC**CGGCCATATCCG**CCACTGCCGCGCCCCCACC**A**GA**GG**CAGAAGC GGAC**TCCGCC**GCTGCCTTG-3′		Frameshift mutation	MMEJ	172.7	-

Table S4.8: Continued

Sample	Type	Sequence	Mutation	Effect	Mechanism		%
	Junction	5'-AGCCCCCGGCCATATCCGCC---------------------------------GCTGCCTTG-3'	37 bp deletion, 6 bp MH				
G12 (n=2)	Genomic	5'-CCCCCGCCATATCCGCCCACTGCCGCGCCCCCACCAGAGGCAGAAGCGGACTCCGCGCCTGCCTTGGC-3'		In-frame mutation	MMEJ or SD-MMEJ (loop-out)	258.3	14.8%
	Junction	5'-CCCCCGCCATATCCGCCCACTGCCGCGCCCCCA---GAGGCAGAAGCGGACTCCGCGCCTGCCTTGGC-3'	3 bp deletion, 3 bp MH or 4 bp MH (4 bp downstream)				
K3	Genomic	5'-CCCCCGCCATATCCGCCCACTGCGCGCCCCCACCAGAGGCAGAAGCGGACTCCGCGCCTTGGCGGAAGCGG-3'		Frameshift mutation	MMEJ or SD-MMEJ (snap-back)	133.1	-
	Junction	5'-CCCCCGCCATATCCGCCCACTGC-------------------------------GGACTCCGCGCCTGCCTTGGCGGAAGCGG-3'	22 bp deletion, 2 bp MH or 5 bp MH (6 bp upstream)				
K9 (n=2)	Genomic	5'-CCCCCGCCATATCCGCCCACTGCCGCGCCCCCACCAGAGGCAGAAGCGGACTCCGCGCCTGCC-3'		Frameshift mutation	C-NHEJ	-	19.5%
	Junction	5'-CCCCCGCCATATCCGCCCACTGCCGCGCCCCCAC-AGAGGCAGAAGCGGACTCCGCGCCTGCC-3'	1 bp deletion, 1 bp MH				
K12	Genomic	5'-TTAACGGAGGAGCCCCGGCCATATCCGCCCACTGCCGCGCCCCCACCAGAGGCAGAAGCGGACTCCGC-3'		Frameshift mutation	Unknown	-	-
	Junction	5'-TTAACGGAGGAGCCCCGGCCATATCCGCCCACTGCCGCGCCCCCAC--AAGGCAGAAGCGGACTCCGC-3'	Deletion + Insertion (Replacement of 3 bp with 1 bp; net= 2 bp deletion)				
K14	Genomic	5'-AGGAGCCCCGGCCATATCCGCCCACTGCCGCGCCCCCACCAGAGGCAGAAGCGGACTCCCCGCCATATCCGCGGAAGCGGCCCCTGACCCTT-3'	13 bp deletion, 2 bp	Frameshift mutation		208.8	-

Table S4.8: Continued

Junction	5'-AGGAGCCCCCGCCATATCCGCCACTGCC--------------GCAGA AGCGGACTCCGCCGCTGCCTTGGCGGAAGCGGCCCCTGACCCTT-3'	MH or 5 bp MH (15 bp downstream)	MMEJ or SD-MMEJ (loop-out)

1 The table shows mRNA Sanger sequencing data of the expressed ERV repair junctions of CHO-K1 clones treated with wild-type Cas9 nuclease and the Myr2 or PPYP6 sgRNAs. The sequences are derived from Sanger sequencing of cDNA PCR amplicons. If the same repair junction was detected more than once, the number is indicated below each sample name as (n=).

2 Predicted blunt-ended DSB sites induced by the two sgRNAs and the wild-type Cas9 nuclease are highlighted in yellow, PAM site are shown in bold blue, and the ATG translation initiation codon is depicted in brown letters. Myr and PPYP target motifs are highlighted in turquoise and purple, respectively. Pre-existing microhomologies (MH) of the microhomology-mediated end-joining (MMEJ) repair mechanism are shown in bold green, while *de novo* MH of the synthesis-dependent microhomology-mediated end-joining (SD-MMEJ) mechanism are underlined with a double line. Inserted bases are represented in bold red letters, deleted bases with a "-" sign, and replacements in italic underlined with a single bold line.

3 Size of mutation and MH length (in bp). The distance between priming site and the break site for *de novo* MH are shown in parenthesis.

4 ERV mutation types include in-frame mutations, out-of-frame mutations, translation inhibition (mutation of the ATG translation initiation codon) or mutations locating outside of the ERV coding region. Out-of-frame mutations and translation inhibition are likely, while in-frame mutations and mutations outside of the coding region are less likely to influence ERV expression and VLP formation.

5 Most probable DSB repair mechanism based on manual junction analysis. Possible repair mechanisms include C-NHEJ, MMEJ, SD-MMEJ (snap-back), SD-MMEJ (loop-out), single strand annealing (SSA), homologous recombination (HR), and unknown. For snap-back SD-MMEJ mechanism, *de novo* priming sites are inverted repeats, while loop-out SD-MMEJ mechanisms uses priming sites with direct repeats (Khodaverdian et al., 2017). If the observed junction sequence is compatible with more than one mechanism and both appear equally likely, all potential pathways are listed. Junctions were verified for homologies at break site and templated insertions (SD-MMEJ) using program described in (Schimmel et al., 2017).

6 Score of each repair pattern according to the MH size and the deletion length. Pattern score was calculated using the RGenome "Microhomology-Predictor" tool (http://www.rgenome.net/mich-calculator/) described in (Bae et al., 2014). The higher the score, the more likely the predicted mutation should be observed. The pattern score is only valid for repair junctions having MHs at the break site (MMEJ-mediated repair). NA: not available.

7 Predicted frequencies of CRISPR-Cas9 editing outcomes using the online tool FORECasT (Favoured Outcomes of Repair Events at Cas9 Targets; https://partslab.sanger.ac.uk/FORECasT) as described in (Allen et al., 2018). The higher the frequency, the more junctions are expected to contain the predicted mutation pattern. Only the frequencies of the predicted ten most frequent mutations are listed.

8 Frequent 1 bp insertions consisting of a duplication of the 4th nucleotide were also observed previously (Lemos et al., 2018)

9 Unknown mechanism but similar junction pattern was described in (Shin et al., 2017)

150

Table S4.9: Cluster analysis of the Myr motif flanking sequences of the CHO type-C ERVs following Illumina deep sequencing. Clusters are compared to the CHO whole genome reference and number of sgRNA targets are calculated.

Myr			Illumina deep sequencing			Whole genome reference				Presence of Myr motif (glycine 2)
Cluster	abolute reads	% reads	ClusterSubgroup	ClusterFrequency	Cluster Frequency up to 100%	ERV number	% reads	Predicted sgRNA recognition	Predicted sgRNA recognition frequency	
Cluster10	16877	4.5	blue	48.6	53	9	8	yes	60.89556867	yes
Cluster100	1976	0.5	blue			1	1	yes		yes
Cluster2	2677	0.7	blue			1	1	yes		yes
Cluster2078	2348	0.6	blue			1	1	yes		yes
Cluster24	9703	2.6	blue			4	4	yes		yes
Cluster51	2465	0.7	blue			1	1	yes		yes
Cluster8	136722	36.3	blue			33	30	yes		yes
Cluster9	10236	2.7	blue			4	4	yes		yes
Cluster78	1603	0.4	green	0.8	1	3	3			yes
Cluster834	1231	0.3	green			0	0			yes
Cluster10566	1403	0.4	red	16.3	18	3	3			yes
Cluster119	2003	0.5	red			1	1			yes
Cluster18	1559	0.4	red			0	0			yes
Cluster20	28691	7.6	red			11	10			yes
Cluster238	3787	1.0	red			1	1			yes
Cluster4	1897	0.5	red			0	0			yes
Cluster57	16035	4.3	red			2	2			yes
Cluster63	1345	0.4	red			1	1			yes
Cluster80	1526	0.4	red			1	1			yes
Cluster87	3274	0.9	red			2	2			yes
Cluster111	6480	1.7	turquoise	26.0	28	1	1			yes
Cluster139	3869	1.0	turquoise			2	2			yes
Cluster158	2673	0.7	turquoise			1	1			yes
Cluster174	4503	1.2	turquoise			2	2	yes		yes
Cluster19	30056	8.0	turquoise			7	6			yes
Cluster27	6383	1.7	turquoise			3	3	yes		yes
Cluster29	6260	1.7	turquoise			3	3	yes		yes
Cluster3	16911	4.5	turquoise			3	3			yes
Cluster379	1674	0.4	turquoise			1	1			yes
Cluster53	6775	1.8	turquoise			3	3	yes		yes
Cluster67	3330	0.9	turquoise			1	1			yes
Cluster69	3578	1.0	turquoise			0	0			yes
Cluster7	3304	0.9	turquoise			2	2	yes		yes
Cluster79	2074	0.6	turquoise			1	1			yes
Total	376471	92		100	100	109	100	13		33

151

Table S4.10: Cluster analysis of the PPYP motif flanking sequences of the CHO type-C ERVs following Illumina deep sequencing. Clusters are compared to the CHO whole genome reference and number of sgRNA targets are calculated.

PPYP	Illumina deep sequencing					Whole genome reference				Presence of PPYP motif
Cluster	abolute reads	% reads	ClusterSubgroup	ClusterFrequency	Cluster Frequency up to 100%	ERV number	% reads	Predicted sgRNA recognition	Predicted sgRNA recognition frequency	Presence of PPYP motif
Cluster2425	169956	44.1	blue	62.0	73	48	39	yes	72.285897	yes
Cluster2435	3406	0.9	blue			2	2	yes		yes
Cluster2437	2688	0.7	blue			2	2	yes		yes
Cluster2450	41087	10.7	blue			10	8	yes		yes
Cluster2469	5038	1.3	blue			1	1	yes		yes
Cluster2477	5497	1.4	blue			3	2	yes		yes
Cluster2717	3387	0.9	blue			7	6	yes		yes
Cluster2742	1514	0.4	blue			1	1			yes
Cluster3009	4073	1.1	blue			2	2	yes		yes
Cluster4949	1857	0.5	blue			1	1			yes
Cluster25	2100	0.5	green	4.2	5	1	1			yes
Cluster2468	1367	0.4	green			1	1			yes
Cluster2793	1556	0.4	green			1	1			yes
Cluster2875	5251	1.4	green			1	1			yes
Cluster3007	1969	0.5	green			1	1			yes
Cluster3013	1797	0.5	green			1	1			yes
Cluster5437	2089	0.5	green			1	1			yes
Cluster20	3449	0.9	red	12.3	15	3	2			yes
Cluster22	27482	7.1	red			12	10			yes
Cluster42	2410	0.6	red			1	1			yes
Cluster99	2030	0.5	red			1	1			yes
Cluster153	6739	1.8	red			8	6			yes
Cluster546	1493	0.4	red			1	1			yes
Cluster605	3839	1.0	red			3	2			
Cluster1243	2466	0.6	turquoise	6.0	7	1	1			
Cluster3888	12156	3.2	turquoise			6	5			
Cluster2446	5412	1.4	turquoise			2	2			yes
Cluster2465	3173	0.8	turquoise			2	2			yes
Total	384960	84		100	100	124	100	8		24

Chapter 5

Conclusions and outlook

The advent of programmable nucleases, notably the CRISPR system, has sustainably impacted modern biology and is going to fundamentally shape it also in the future. CRISPR is currently omnipresent as the genome editing tool of choice in biological research. However, while CRISPR-mediated genome editing can be used to efficiently inactivate single genes, it remains relatively inefficient at inducing precise HR-mediated gene modifications as well as at simultaneously inactivating multiple genes by end-joining pathways. In this thesis, I aimed to address both of these limitations by better characterizing DSB repair mechanisms in the biotechnology-relevant CHO cells.

5.1   Identifying rate-limiting factors for HR repair

As a first part of this thesis, we designed a chromosomal HR gene correction assay in CHO cells in order to quantify how HR frequency alters upon knockdown and/or over-expression of DNA repair factors previously implicated in regulating HR repair. Our initial hypothesis was that the predominance of alt-EJ activity in CHO cells forms a major bottleneck for efficient HR-mediated genome editing. Surprisingly, we found that most alt-EJ factors positively contribute to HR in lieu of being anti-recombinogenic. By varying nuclease types and distance between the DSB and the gene correction site, we further demonstrated that alt-EJ contribution appears to be most relevant for the repair of short 5' overhanging DSB ends, as generated by the FokI-dCas9 system, or of blunt-ended DSB ends, as introduced by the wild-type Cas9 nuclease of *Streptococcus pyogenes*

(SpCas9), with long non-homologous protruding flaps. This indicates that alt-EJ factors may possess non-canonical activities to ensure bona fide HR repair, in addition to their well-established activity in end-joining pathways. The proposed interplay between alt-EJ and HR, thus, challenges the prevailing view that they are primarily competitors for the same DSB substrate.

Our current model assumes that in specific repair contexts alt-EJ factors non-canonically engage at an early as well as at a late stage of DSB repair. While alt-EJ factors may contribute to DSB recognition and processing at an early stage, they rather assist in processing HR intermediates at a late stage. In support for our model, growing evidence suggests that end-joining pathways may be involved in completing and/or rescuing HR intermediates [92, 141, 293, 373, 379]. For instance, the Ercc1-Xpf endonuclease complex appears not only to have key roles in NER and SSA pathways but also to remove non-homologous flap structures during HR in both yeast and mammalian cells, consistent with our data [373, 374]. Future work will be needed to understand how alt-EJ and HR activities are coordinated and whether these factors act as a complex through direct protein-protein interactions to stimulate HR-mediated gene conversion.

As a follow up to this study, it would be necessary to provide further support for an alt-EJ contribution to HR in other repair settings. For instance, it would be interesting to assess whether alt-EJ contributes to HR at DSBs induced by alternative CRISPR systems that may have a distinct globular protein structure, lead to different DSB ends and/or show other nuclease dynamics [248, 249]. Of particular interest to this study is the Cas9 ortholog of *Francisella novicida* (FnCas9) [257]. FnCas9 has the peculiarity that, on the one hand, it has the same target site requirements like the classical wild-type SpCas9, but on the other hand, it creates 4 nt 5' overhanging DSBs like the highly specific FokI-dCas9. This would allow validating and possibly extending our previous findings using another Cas9 ortholog. Moreover, it would be of great interest to test how the structure and origin of donor templates (e.g. whether they are single stranded or double stranded, or derived from non-viral or viral vectors) influence alt-EJ activity during HR. Ultimately, we could also assess whether alt-EJ factors contribute to HR-mediated large transgene insertions, as frequently used for biotechnological applications.

This work relied on CHO cells to advance our understanding of alt-EJ mechanisms and functions in mammalian cells. However, further investigations will be necessary to evaluate whether alt-EJ pathways fulfill similar roles in other cell types and/or organisms.

It is well-documented that cells can have intrinsically different DSB repair pathway activities, depending, among others, on their degree of differentiation and deficiencies in other DSB repair pathways. In the case of the immortalized CHO cells, alt-EJ pathways appear upregulated at the expense of C-NHEJ and HR activities, as we have also observed during this thesis [119, 141, 164, 171, 291]. This implies that DSB repair mechanisms and/or regulations may differ between CHO and other cells. In addition, different CHO cell lines are known to possess specific genetic diversity and are thus sometimes referred to as "quasispecies" [368]. In this study, most of the assessed repair factors – including many alt-EJ factors – showed consistent effects in both the adherent CHO-DG44 and the suspension-adapted CHO-K1 cell lines. This argues for a rather conserved alt-EJ role in CHO cells. Although alt-EJ pathways play only a minor role in primary and non-malignant cells, they are important in many cancer cells [116–118, 170]. Therefore, it would be tempting to corroborate our findings in cancer cells with similar DSB repair characteristics as CHO cells.

Despite the identification of numerous alt-EJ factors and subpathways, alt-EJ remains poorly defined. Several factors have complicated its in-depth characterization and separation from other repair mechanisms. First of all, alt-EJ is not a single repair pathway but a family of multiple subpathways. These alt-EJ pathways utilize partially overlapping but also distinct repair factors [121, 122, 141]. Alt-EJ pathways also appear to have co-opted many proteins from other repair pathways, notably from HR, SSA and excision repair pathways [128, 176, 177, 186, 426]. So far, Pol theta is the only alt-EJ repair protein with no function in other DNA repair pathways [132]. Additionally, alt-EJ protein requirements and/or mechanisms seem not to be conserved between lower and higher eukaryotes. For example, alt-EJ pathways are independent of Parp1, Pol theta and Lig3 in yeast, where these proteins are absent, and they require longer MH sequences for end annealing compared to mammalian cells [123]. Finally, alt-EJ pathways are not the only repair mechanisms that may rely on MHs as repair intermediate. Essentially MHs were implicated in C-NHEJ [24, 25], in a specific type of SDSA termed microhomology-mediated SDSA (MM-SDSA) [427], in microhomology-mediated BIR (MM-BIR) [428] as well as in a replication-based repair mechanism to regress stalled replication forks termed FoSTeS [429]. Single cell analysis and omics data will be valuable technologies to further improve our current DSB repair pathway understanding and, in particular, might help to better disentangle alt-EJ subpathways and define their corresponding protein requirements [430].

The inherently low HR frequency in most mammalian cells, including CHO cells, raises the intriguing question of which steps are rate-limiting in HR. Homology search is considered one of the major rate-limiting step of HR in eukaryotes [431]. Consistent with this assumption, we found that attempts to stimulate the formation of the Rad51 presynaptic complex, either by overexpressing the Rad51 recombinase or by knocking down the Pari anti-recombinase, increased HR. Therefore, it appears that the natural Rad51 levels in mammalian cells with large and highly complex genomes are kept in check to prevent unscheduled and/or non-allelic recombinations. Besides homology search, we provide evidence that end resection is another rate-limiting step of HR in CHO cells. More precisely, we found that knockdown of the Mre11 nuclease favors HR repair, while knockdown of the 53BP1 end protection protein reduces HR repair. This somewhat counterintuitive result suggests that end resection needs to occur "just to the right" extent to permit productive HR repair. In line with this hypothesis, several studies reported that excessive DSB end resection (known as hyper-resection) can lead to a switch from productive high-fidelity HR to highly mutagenic non-allelic recombination or SSA repair [366, 432, 433]. This implies that HR can be blocked not only by absence of DSB end resection but also by hyper-resection [366].

This work extends numerous previous studies attempting to improve HR and HR-mediated genome editing in mammalian cells. Most of these studies tested a single repair factor in a specific cellular context, which typically resulted in only moderate improvement rates and/or in conflicting result between cell types (e.g. [171, 309, 314]). This indicates tightly controlled HR activities and the absence of one single universal HR bottleneck sufficient to boost HR in distinct cell lines. Therefore, future improvement strategies will need to focus on modulating multiple repair factors and/or entire pathways in parallel taking into account cell line specific repair properties – as we attempted in this study with CHO cells. Ultimately, the aim would be to establish unique protein "cocktails" that allow stimulating HR, or any other DNA repair pathway, in a cell line-specific manner.

5.2   Inactivating endogenous viral elements

As a second part of this thesis, we performed multi-locus genome editing in order to inactivate group 1 type-C endogenous retroviruses (ERVs) presumably responsible for

the release of viral-like particles (VLPs) in CHO cells. Following transient CRISPR-Cas9 expression, 10-15% of the isolated clones contained group 1 ERV mutations at the mRNA level despite targeting more than 50 predicted genomic ERV loci. Moreover, we developed a bioinformatic pipeline to analyze CRISPR-derived mutations at the genome level by targeted DNA amplicon sequencing. This analysis highlighted that clones bearing mutations in the expressed group 1 type-C ERV sequence contained between 1 to 14 mutated ERV loci in total. Interestingly, disrupting the *gag* ERV gene of one particular ERV locus sufficed to decrease the amount of viral RNA-containing particles close to the detection limit, suggesting that a single ERV is responsible for VLP production in CHO cells. Clones with reduced functional VLP levels seemed not to have altered key CHO properties, including the capacity to produce therapeutic proteins. Therefore, the genome editing approach established here provides new avenues to increase the safety profile of CHO cells in biomanufacturing.

This study is a proof-of-principle that transient CRISPR-Cas9 expression followed by enrichment for transfected cells is suitable for inactivating multi-locus ERV sites in CHO cells. Previous studies attempting to mutate multiple ERV sequences in porcine cells failed to obtain acceptable editing frequencies using transient CRISPR-Cas9 expression and, instead, had to rely on stable integration of the sgRNA and Cas9 sequences into the to be edited genome [325, 326]. Although the overall editing frequencies were lower than those obtained from stable CRISPR-Cas9 expression, as expected, the transient CRISPR-Cas9 expression approach was sufficient to reduce the release of functional VLPs from CHO cells. This discrepancy may be explained by the different number of active ERV copies in porcine and hamster cells. Unlike porcine cells in which many ERVs remain active, we showed that presumably only a single group 1 type-C ERV locus is transcriptionally active and able to produce VLPs in CHO cells. Thus, the preference of CRISPR-Cas9 to cleave open chromatin seems to act as a natural selection mechanism to primarily mutate functionally relevant ERV loci [410]. However, if more or possible all ERV sequences were to be disrupted, for instance, to circumvent *trans* complementation of ERVs, additional strategies would be desired to further improve the efficiency of multi-locus genome editing by transient CRISPR-Cas9 expression.

In order to augment transiently expressed CRISPR-Cas9 multi-locus gene inactivations, one interesting strategy could be to prevent HR-mediated gene conversion. In fact, the large number of intact homologous ERV sequences may serve as donor template for HR repair. Thus, precise HR-mediated ERV reconstitution could oppose gene inactiva-

tion by inaccurate end-joining pathways. Supporting this hypothesis, we detected ERV repair junctions compatible with HR repair pathways, suggesting that HR contributes to the repair of CRISPR-targeted ERV sequences. Therefore, it would be interesting to assess ERV knock-out frequencies in HR-deficient backgrounds. Preliminary experiments combining ERV mutagenesis with knockdown of Rad51 in CHO cells were inconclusive. As HR inactivation appeared not to be necessary to disrupt the relevant ERV copy in CHO cells, we did not further pursue this strategy.

In addition to interfering with HR, gene inactivation at multiple target sites may be improved by channeling DSB repair into more mutagenic end-joining pathways. SSA belongs to the most mutagenic end-joining pathways. Unlike the C-NHEJ and SD-MMEJ pathways that can lead to apparent blunt end joints, SSA always leads to large deletions of sometimes several kb in length, almost exclusively causing loss-of-function mutations [101, 102]. Hence, one might try to block HR together with C-NHEJ to favor alt-EJ pathways, e.g. by simultaneously knocking down Rad51 and Ku. Furthermore, this approach could be extended in combination with a Pol theta knockdown to additionally block the MMEJ and SD-MMEJ pathways to avoid small in-frame mutations in favor of SSA. The bioinformatic pipeline developed in this study would allow to quantify and compare indel length and out-of-frame frequency for treated and control samples in the context of CHO ERV mutagenesis.

A major concern of CRISPR-Cas9 genome editing is the potential adverse side-effects, such as off-target mutagenesis and chromosomal aberrations. The transient CRISPR-Cas9 expression approach used in this study leads to timely restricted nuclease activity, potentially avoiding extensive DDR signaling and off-target mutagenesis, as typically observed following stable CRISPR-Cas9 expression [325, 326]. Despite these precautions, the established ERV-edited CHO cells would have to be further characterized prior to their use in biopharmaceutical production to ensure lack of adverse effects. Additional characterization steps could include, for instance, identifying potential CRISPR-induced off-target mutations in the ERV-edited CHO clones by whole-genome sequencing. Karyotype analysis would be another important characterization to check for undesired large deletions and/or complex rearrangements. Such mutations are difficult to detect by standard sequencing techniques, like PCR and Illumina paired-end sequencing, and have been systematically overlooked in the past although they could happen relatively frequently [299]. The risk of chromosomal aberrations is expected to be increased when introducing multiple DSBs simultaneously [326]. Hence, karyotyping ERV-edited CHO cells obtained

from transient CRISPR-Cas9 expression could be useful to demonstrate not only the absence of detectable chromosomal changes but also the superior genome integrity compared to stable expression of CRISPR-Cas9.

Another avenue of future research is to investigate whether ERV mutagenesis may affect the expression of viral and/or cellular genes. Preliminary RT-qPCR data suggest that clones mutated in the expressed group 1 type-C ERV sequences show unaffected mRNA levels of group 1 as well as the other expressed type-C ERV group. Moreover, we showed that ERV mutagenesis does not consistently alter CHO phenotypes (e.g. cell growth, cell size, therapeutic protein production) compared to control cells, implying the absence of critical changes in gene expression in the surviving clones. Nevertheless, it would be worthwhile complementing these experiments with an analysis of the transcriptome of edited CHO cells. For instance, such a transcriptome analysis could provide more global insight into how ERV editing influences the intracellular mRNA level of mutated as well as other non-mutated ERV sequences. In addition, it would allow us to detect potential changes in the expression of cellular genes. Such transcriptional changes may result from cellular adaptation to reduced VLP budding, off-target mutagenesis and cell responses to CRISPR genome editing. One particularly interesting protein to look at is p53. Altered or low p53 activities may have allowed the surviving clones to evade apoptosis induced by a strong DDR upon editing of numerous ERV sequences in parallel [434].

Besides creating loss-of-function mutations, other strategies could also prevent the release of ERV-derived particles from CHO cells. For instance, one possible strategy might be to introduce a dominant-negative mutation into the myristoylation motif in the *gag* gene. Studies using the feline leukemia virus (FeLV) and the human immunodeficiency virus (HIV) demonstrated that the substitution of the myristoylation glycine residue with an alanine severely inhibits particle release in a dominant-negative manner [422, 423]. To test whether the same mutation has a similar effect in CHO cells, this mutation could initially be expressed from a stably integrated plasmid vector and could later be introduced site-specifically by homology-directed repair. Unlike loss-of-function ERV mutations, this strategy could even prevent viral particle formation from reactivated ERVs as well as new viral infections. Another interesting strategy is to entirely excise the single ERV locus responsible for particle release in CHO cells. A similar approach was already deployed in HIV-infected human cells [420]. In contrast to multi-locus ERV mutagenesis, this approach requires two sgRNAs to introduce DSBs flanking the ERV sequence, which should greatly reduce the elicited DNA damage response and avoid the accumulation of

defective ERV RNAs in the cytoplasm.

An outstanding question is the potential function of ERVs in CHO cells. Although integration of viruses often negatively influences the fitness of the host organism (e.g. leading to diseases), their integration was as well associated with several positive effects (e.g increasing the genetic diversity by introducing new genes, providing new regulatory elements and protecting against other viral infections) [344, 345, 350]. The best-described examples for such positive effects are Syncytin proteins that are derived from the *env* gene of human ERVs and became co-opted for placenta development in mammalian embryogenesis [435]. Thus, one could advance the theory that those ERV sequences which remain full-length and/or expressed benefit the CHO cells' fitness. In that case, we would expect changes in the mRNA levels and/or higher susceptibility to viral infections in CHO cells in which the VLP-producing ERV locus has been excised by CRISPR-Cas9 editing. Alternatively, such intact ERV sequences could result from recent integrations, which have not yet accumulated deleterious mutations and/or became epigenetically silenced.

5.3   Perspectives in genome engineering

The current CRISPR-Cas9 applications are only the tip of the iceberg of what could be done with CRISPR nucleases. Besides genome editing, CRISPR holds great promises in labeling specific DNA sequences and activating as well as repressing gene expression [233]. A growing number of studies start focusing on the translation of CRISPR technologies into clinics, agriculture and biotechnology. For instance, future CRISPR applications might involve genetically modifying livestock and crop plants, controlling disease-causing mosquito populations via gene drives and germline editing [436]. As of today, 24 registered clinical trials are ongoing [437]. More will certainly arise in the future. However, limitations in CRISPR efficiency, as discussed in this thesis, as well as safety and ethical concerns still need to be thoroughly addressed before broad translational applications.

CRISPR-mediated genome editing is also becoming increasingly important for biotechnology applications, notably for genetically engineering CHO host cells. The main focus in CHO cell line engineering may be to improve productivity and cell line robustness [438]. In the past, CHO cells were genetically engineered to resist apoptosis by inactivating pro-apoptotic factors like Bax [439] or to produce less metabolic by-products

such as lactate, which negatively affect cell growth during large-scale cell culture [440]. Moreover, CHO cell lines lacking glutamine synthetase were established to facilitate the selection of high-producing clones using glutamine-depleted cell culture medium [441]. However, more recently, site-specific transgene integration into CHO cells has attracted growing interest [268, 442, 443]. Unlike random integration, targeted integration is expected to yield homogeneous and predictable transgene expression in a predefined genetic background [268]. This could drastically reduce the time required for selection and screening for high-producing clones. Interestingly, circumstantial evidence suggests that VLP levels are on average lower in clones derived from targeted rather than from random transgene integration (Conference presentation Merck company). This suggests that chromosomal rearrangements, as frequently observed upon random integration [216], may stimulate ERV expression and/or cause ERV reactivation.

At present, targeted integration has not replaced random integration as the standard approach for establishing CHO cell lines for biopharmaceutical production [444]. This has two main reasons. The first reason is that targeted transgene expression levels are typically inferior to those obtained from traditional random integrations. One explanation for this observation could be the considerably lower transgene copy number following targeted integration than random integration. While targeted integration typically results in a single transgene copy, random integration causes multicopy transgene arrays at one or few genomic sites. This raises the question of whether similar therapeutic protein levels can be achieved from targeted as from random transgene integrations or of whether optimal integration sites have just not yet been identified in the CHO genome. Optimal integration sites should support high and reproducible long-term transgene expression in well-characterized genomic loci without disrupting genes and/or regulatory elements. Such genomic loci are known as "safe harbors". Therefore, current attempts try to deploy transcriptomics to identify highly expressed genomic sites suitable as safe harbors [444]. It remains unclear whether a single best integration site exists or whether the optimal integration site is transgene dependent. The second reason is that HR-mediated precise integration of transgenes remains relatively inefficient in CHO cells, as extensively discussed during this thesis. Therefore, new strategies to boost HR in CHO cells are highly desired and will have broad implications for biomanufacturing.

References

[1] Ciccia A, Elledge SJ (2010) The DNA Damage Response: Making It Safe to Play with Knives. *Molecular Cell* 40(2):179–204.

[2] D'Adda Di Fagagna F (2008) Living on a break: Cellular senescence as a DNA-damage response. *Nature Reviews Cancer* 8(7):512–522.

[3] Jackson SP, Bartek J (2009) The DNA-damage response in human biology and disease. *Nature* 461(7267):1071–1078.

[4] Sulli G, Di Micco R, D'Adda Di Fagagna F (2012) Crosstalk between chromatin state and DNA damage response in cellular senescence and cancer. *Nature Reviews Cancer* 12(10):709–720.

[5] Goldstein M, Kastan MB (2015) The DNA Damage Response: Implications for Tumor Responses to Radiation and Chemotherapy. *Annual Review of Medicine* 66(1):129–143.

[6] Shiloh Y (2003) ATM and related protein kinases: safeguarding genome integrity. *Nature Reviews Cancer* 3(3):155–168.

[7] Maréchal A, Zou L (2013) DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harbor Perspectives in Biology* 5:1–17.

[8] Lamarche BJ, Orazio NI, Weitzman MD (2010) The MRN complex in double-strand break repair and telomere maintenance. *FEBS Letters* 584(17):3682–3695.

[9] Matsuoka S, et al. (2007) ATM and ATR Substrate Analysis Reveals Extensive Protein Networks Responsive to DNA Damage. *Science* 316(5828):1160–1166.

[10] Kantidze OL, Velichko AK, Luzhin AV, Petrova NV, Razin SV (2018) Synthetically Lethal Interactions of ATM, ATR, and DNA-PKcs. *Trends in Cancer* 4(11):755–768.

[11] Burma S, Chen BP, Murphy M, Kurimasa A, Chen DJ (2001) ATM Phosphorylates

Histone H2AX in Response to DNA Double-strand Breaks. *Journal of Biological Chemistry* 276(45):42462–42467.

[12] Ward IM, Chen J (2001) Histone H2AX Is Phosphorylated in an ATR-dependent Manner in Response to Replicational Stress. *Journal of Biological Chemistry* 276(51):47759–47762.

[13] Stewart GS, Wang B, Bignell CR, Taylor AMR, Elledge SJ (2003) MDC1 is a mediator of the mammalian DNA damage checkpoint. *Nature* 421:961.

[14] Riley T, Sontag E, Chen P, Levine A (2008) Transcriptional control of human p53-regulated genes. *Nature Reviews Molecular Cell Biology* 9(5):402–412.

[15] Vilenchik MM, Knudson AG (2003) Endogenous DNA double-strand breaks: Production, fidelity of repair, and induction of cancer. *Proceedings of the National Academy of Sciences* 100(22):12871–12876.

[16] Mehta A, Haber JE (2014) Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harbor Perspectives in Biology* 6(9):a016428.

[17] Garaycoechea JI, et al. (2018) Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* 553(7687):171–177.

[18] Handel MA, Schimenti JC (2010) Genetics of mammalian meiosis: Regulation, dynamics and impact on fertility. *Nature Reviews Genetics* 11(2):124–136.

[19] Soulas-Sprauel P, et al. (2007) V(D)J and immunoglobulin class switch recombinations: A paradigm to study the regulation of DNA end-joining. *Oncogene* 26(56):7780–7791.

[20] Papavasiliou FN, Schatz DG (2000) Cell-cycle-regulated DNA double-strand breaks in somatic hypermutation of immunoglobulin genes. *Nature* 408(6809):216–221.

[21] Beucher A, et al. (2009) ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2. *EMBO Journal* 28(21):3413–3427.

[22] Kerzendorfer C, O'Driscoll M (2009) Human DNA damage response and repair deficiency syndromes: Linking genomic instability and cell cycle checkpoint proficiency. *DNA Repair* 8(9):1139–1152.

[23] Mao Z, Bozzella M, Seluanov A, Gorbunova V (2008) Comparison of nonhomol-

ogous end joining and homologous recombination in human cells. *DNA Repair* 7(10):1765–1771.

[24] Pannunzio NR, Watanabe G, Lieber MR (2018) Nonhomologous DNA end-joining for repair of DNA double-strand breaks. *Journal of Biological Chemistry* 293(27):10512–10523.

[25] Chang HH, et al. (2016) Different DNA end configurations dictate which NHEJ components are most important for joining efficiency. *Journal of Biological Chemistry* 291(47):24377–24389.

[26] Bétermier M, Bertrand P, Lopez BS (2014) Is Non-Homologous End-Joining Really an Inherently Error-Prone Process? *PLoS Genetics* 10(1):e1004086.

[27] Odersky A, et al. (2002) Repair of sequence-specific 125I-induced double-strand breaks by nonhomologous DNA end joining in mammalian cell-free extracts. *Journal of Biological Chemistry* 277(14):11756–11764.

[28] Biehs R, et al. (2017) DNA Double-Strand Break Resection Occurs during Non-homologous End Joining in G1 but Is Distinct from Resection during Homologous Recombination. *Molecular Cell* 65(4):671 – 684.e5.

[29] Panier S, Boulton SJ (2014) Double-strand break repair: 53BP1 comes into focus. *Nature Reviews Molecular Cell Biology* 15(1):7–18.

[30] Chapman JR, et al. (2013) RIF1 Is Essential for 53BP1-Dependent Nonhomologous End Joining and Suppression of DNA Double-Strand Break Resection. *Molecular Cell* 49(5):858–871.

[31] Xu G, et al. (2015) REV7 counteracts DNA double-strand break resection and affects PARP inhibition. *Nature* 521(7553):541–4.

[32] Difilippantonio S, et al. (2008) 53BP1 facilitates long-range DNA end-joining during V(D)J recombination. *Nature* 456(7221):529–533.

[33] Gu Y, Jin S, Gao Y, Weaver DT, Alt FW (1997) Ku70-deficient embryonic stem cells have increased ionizing radiosensitivity, defective DNA end-binding activity, and inability to support V(D)J recombination. *Proceedings of the National Academy of Sciences* 94(15):8076–8081.

[34] Singleton BK, et al. (1997) Molecular and biochemical characterization of xrs mutants defective in Ku80. *Molecular and Cellular Biology* 17(3):1264–73.

[35] Mimori T, Hardin JA (1986) Mechanism of interaction between Ku protein and DNA. *Journal of Biological Chemistry* 261(22):10375–10379.

[36] Blier PR, Griffith AJ, Craft J, Hardin JA (1993) Binding of Ku Protein to DNA. *The Journal of Biological Chemistry* 268(10):7594–7601.

[37] Dynan WS, Yoo S (1998) Interaction of Ku protein and DNA-dependent protein kinase catalytic subunit with nucleic acids. *Nucleic Acids Research* 26(7):1551–1559.

[38] Mladenov E, Magin S, Soni A, Iliakis G (2013) DNA Double-Strand Break Repair as Determinant of Cellular Radiosensitivity to Killing and Target in Radiation Therapy. *Frontiers in Oncology* 3:113.

[39] Jeggo PA (1997) DNA-PK: At the cross-roads of biochemistry and genetics. *Mutation Research - DNA Repair* 384(1):1–14.

[40] Goodarzi AA, et al. (2006) DNA-PK autophosphorylation facilitates Artemis endonuclease activity. *EMBO Journal* 25(16):3880–3889.

[41] Riballo E, et al. (2004) A Pathway of Double-Strand Break Rejoining Dependent upon ATM, Artemis, and Proteins Locating to γ-H2AX Foci. *Molecular Cell* 16(5):715–724.

[42] Bertocci B, De Smet A, Weill JC, Reynaud CA (2006) Nonoverlapping Functions of DNA Polymerases Mu, Lambda, and Terminal Deoxynucleotidyltransferase during Immunoglobulin V(D)J Recombination In Vivo. *Immunity* 25(1):31–41.

[43] Andrade P, Martin MJ, Juarez R, Lopez de Saro F, Blanco L (2009) Limited terminal transferase in human DNA polymerase defines the required balance between accuracy and efficiency in NHEJ. *Proceedings of the National Academy of Sciences* 106(38):16203–16208.

[44] Lieber MR (2010) The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway. *Annual Review of Biochemistry* 79(1):181–211.

[45] Ahnesorg P, Smith P, Jackson SP (2006) XLF interacts with the XRCC4-DNA Ligase IV complex to promote DNA nonhomologous end-joining. *Cell* 124(2):301–313.

[46] Xing M, et al. (2015) Interactome analysis identifies a new paralogue of XRCC4 in non-homologous end joining DNA repair pathway. *Nature Communications* 6:6233.

[47] Petermann E, Helleday T (2010) Pathways of mammalian replication fork restart. *Nature Reviews Molecular Cell Biology* 11(10):683–687.

[48] Deans AJ, West SC (2011) DNA interstrand crosslink repair and cancer. *Nature Reviews Cancer* 11(7):467–480.

[49] Malkova A, Ira G (2013) Break-induced replication: functions and molecular mechanism. *Current Opinion in Genetics & Development* 23(3):271–279.

[50] Shibata A, et al. (2011) Factors determining DNA double-strand break repair pathway choice in G2 phase. *EMBO Journal* 30(6):1079–1092.

[51] Sonoda E, Hochegger H, Saberi A, Taniguchi Y, Takeda S (2006) Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair* 5(9-10):1021–1029.

[52] Aymard F, et al. (2014) Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nature Structural & Molecular Biology* 21(4):366–374.

[53] Shahar OD, et al. (2012) Live imaging of induced and controlled DNA double-strand break formation reveals extremely low repair by homologous recombination in human cells. *Oncogene* 31(30):3495–3504.

[54] Prakash R, Zhang Y, Feng W, Jasin M (2015) Homologous Recombination and Human Health: The Roles of BRCA1, BRCA2, and Associated Proteins. *Cold Spring Harbor Perspectives in Biology* 7(4):a016600.

[55] Waldman AS (2008) Ensuring the fidelity of recombination in mammalian chromosomes. *BioEssays* 30(11-12):1163–1171.

[56] Johnson RD (2000) Sister chromatid gene conversion is a prominent double-strand break repair pathway in mammalian cells. *The EMBO Journal* 19(13):3398–3407.

[57] Delacôte F, Lopez BS (2008) Importance of the cell cycle phase for the choice of the appropriate DSB repair pathway, for genome stability maintenance: The trans-S double-strand break repair model. *Cell Cycle* 7(1):33–38.

[58] Moynahan ME, Jasin M (2010) Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nature Reviews Molecular Cell Biology* 11(3):196–207.

[59] Rothkamm K, Krüger I, Thompson LH, Löbrich M (2003) Pathways of DNA double-strand break repair during the mammalian cell cycle. *Molecular and Cellular Biology* 23(16):5706–5715.

[60] Isono M, et al. (2017) BRCA1 Directs the Repair Pathway to Homologous Recombination by Promoting 53BP1 Dephosphorylation. *Cell Reports* 18(2):520–532.

[61] Cannavo E, Cejka P (2014) Sae2 promotes dsDNA endonuclease activity within Mre11-Rad50-Xrs2 to resect DNA breaks. *Nature* 514(7520):122–125.

[62] Sartori AA, et al. (2007) Human CtIP promotes DNA end resection. *Nature* 450(7169):509–514.

[63] Daley JM, Niu H, Miller AS, Sung P (2015) Biochemical mechanism of DSB end resection and its regulation. *DNA Repair* 32:66–74.

[64] Symington LS, Rothstein R, Lisby M (2014) Mechanisms and regulation of mitotic recombination in Saccharomyces cerevisiae. *Genetics* 198(3):795–835.

[65] Benson FE, Stasiak A, West SC (1994) Purification and characterization of the human Rad51 protein, an analogue of E. coli RecA. *The EMBO journal* 13(23):5764–5771.

[66] Baumann P, West SC (1998) Role of the human RAD51 protein in homologous recombination and double-stranded-break repair. *Trends in Biochemical Sciences* 23(7):247–251.

[67] Sy SMH, Huen MSY, Chen J (2009) PALB2 is an integral component of the BRCA complex required for homologous recombination repair. *Proceedings of the National Academy of Sciences* 106(17):7155–7160.

[68] Chun J, Buechelmaier ES, Powell SN (2013) Rad51 Paralog Complexes BCDX2 and CX3 Act at Different Stages in the BRCA1-BRCA2-Dependent Homologous Recombination Pathway. *Molecular and Cellular Biology* 33(2):387–395.

[69] Feng Z, et al. (2011) Rad52 inactivation is synthetically lethal with BRCA2 deficiency. *Proceedings of the National Academy of Sciences* 108(2):686–691.

[70] Orthwein A, et al. (2015) A mechanism for the suppression of homologous recombination in G1 cells. *Nature* 528(7582):422–426.

[71] Burkovics P, et al. (2016) The PCNA-associated protein PARI negatively regulates homologous recombination via the inhibition of DNA repair synthesis. *Nucleic acids research* 44(7):3176–3189.

[72] Simandlova J, et al. (2013) FBH1 helicase disrupts RAD51 filaments in vitro and modulates homologous recombination in mammalian cells. *Journal of Biological Chemistry* 288(47):34168–34180.

[73] Hu Y, et al. (2007) RECQL5/Recql5 helicase regulates homologous recombination and suppresses tumor formation via disruption of Rad51 presynaptic filaments. *Genes and Development* 21(23):3073–3084.

[74] Bugreev DV, Yu X, Egelman EH, Mazin AV (2007) Novel pro- and anti-recombination activities of the Bloom's syndrome helicase. *Genes & Development* 21(23):3085–3094.

[75] Forget AL, Kowalczykowski SC (2012) Single-molecule imaging of DNA pairing by RecA reveals a three-dimensional homology search. *Nature* 482(7385):423–427.

[76] Ragunathan K, Liu C, Ha T (2012) RecA filament sliding on DNA facilitates homology search. *eLife* 2012(1):e00067.

[77] Nasmyth K, Haering CH (2009) Cohesin: Its Roles and Mechanisms. *Annual Review of Genetics* 43(1):525–558.

[78] Ramani V, Shendure J, Duan Z (2016) Understanding Spatial Genome Organization: Methods and Insights. *Genomics, Proteomics and Bioinformatics* 14(1):7–20.

[79] Zhao W, et al. (2017) BRCA1–BARD1 promotes RAD51-mediated homologous DNA pairing. *Nature* 550(7676):360–365.

[80] Heyer WD, Li X, Rolfsmeier M, Zhang XP (2006) Rad54: The Swiss Army knife of homologous recombination? *Nucleic Acids Research* 34(15):4115–4125.

[81] Renkawitz J, Lademann CA, Jentsch S (2014) Mechanisms and principles of homology search during recombination. *Nature Reviews Molecular Cell Biology* 15(6):369–383.

[82] Spies M, Fishel R (2015) Mismatch Repair during Homologous and Homeologous Recombination. *Cold Spring Harbor Perspectives in Biology* 7(3):a022657.

[83] Li X, Heyer WD (2009) RAD54 controls access to the invading 3'-OH end after RAD51-mediated DNA strand invasion in homologous recombination in Saccharomyces cerevisiae. *Nucleic Acids Research* 37(2):638–646.

[84] McVey M, Khodaverdian VY, Meyer D, Cerqueira PG, Heyer WD (2016) Eukaryotic DNA Polymerases in Homologous Recombination. *Annual Review of Genetics* 50(1):393–421.

[85] Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW (1983) The double-strand-break repair model for recombination. *Cell* 33(1):25–35.

[86] Wyatt HDM, West SC (2014) Holliday Junction Resolvases. *Cold Spring Harbor Perspectives in Biology* 6(9):a023192.

[87] Zapotoczny G, Sekelsky J (2017) Human Cell Assays for Synthesis-Dependent

Strand Annealing and Crossing over During Double-Strand Break Repair. *G3: Genes, Genomes, Genetics* 7(4):1191–1199.

[88] Barber LJ, et al. (2008) RTEL1 Maintains Genomic Stability by Suppressing Homologous Recombination. *Cell* 135(2):261–271.

[89] Gari K, Décaillet C, Stasiak AZ, Stasiak A, Constantinou A (2008) The Fanconi Anemia Protein FANCM Can Promote Branch Migration of Holliday Junctions and Replication Forks. *Molecular Cell* 29(1):141–148.

[90] Prakash R, et al. (2009) Yeast Mphl helicase dissociates Rad51-made D-loops: Implications for crossover control in mitotic recombination. *Genes and Development* 23(1):67–79.

[91] van Brabant AJ, et al. (2000) Binding and Melting of D-Loops by the Bloom Syndrome Helicase. *Biochemistry* 39(47):14617–14625.

[92] Lydeard JR, Jain S, Yamaguchi M, Haber JE (2007) Break-induced replication and telomerase-independent telomere maintenance require Pol32. *Nature* 448(7155):820–823.

[93] Wilson MA, et al. (2013) Pif1 helicase and Polδ promote recombination-coupled DNA synthesis via bubble migration. *Nature* 502(7471):393–396.

[94] Smith CE, Llorente B, Symington LS (2007) Template switching during break-induced replication. *Nature* 447(7140):102–105.

[95] Anand RP, et al. (2014) Chromosome rearrangements via template switching between diverged repeated sequences. *Genes and Development* 28(21):2394–2406.

[96] Bzymek M, Thayer NH, Oh SD, Kleckner N, Hunter N (2010) Double holliday junctions are intermediates of DNA break repair. *Nature* 464(7290):937–941.

[97] Mitchel K, Zhang H, Welz-Voegele C, Jinks-Robertson S (2010) Molecular Structures of Crossover and Noncrossover Intermediates during Gap Repair in Yeast: Implications for Recombination. *Molecular Cell* 38(2):211–222.

[98] Cole F, Keeney S, Jasin M (2012) Preaching about the converted: how meiotic gene conversion influences genomic diversity. *Annals of the New York Academy of Sciences* 1267(1):95–102.

[99] Hum YF, Jinks-Robertson S (2017) Mitotic Gene Conversion Tracts Associated with Repair of a Defined Double-Strand Break in Saccharomyces cerevisiae. *Genetics* 207(1):115–128.

[100] Bhargava R, Onyango DO, Stark JM (2016) Regulation of Single-Strand Annealing and its Role in Genome Maintenance. *Trends in Genetics* 32(9):566–575.

[101] Puchta H (2005) The repair of double-strand breaks in plants: Mechanisms and consequences for genome evolution. *Journal of Experimental Botany* 56(409):1–14.

[102] Kolomietz E, Meyn MS, Pandita A, Squire JA (2002) The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes and Cancer* 35(2):97–112.

[103] Stark JM, Pierce AJ, Oh J, Pastink A, Jasin M (2004) Genetic Steps of Mammalian Homologous Repair with Distinct Mutagenic Consequences. *Molecular and Cellular Biology* 24(21):9305–9316.

[104] Sargent RG, et al. (2000) Role of the nucleotide excision repair gene ERCC1 in formation of recombination-dependent rearrangements in mammalian cells. *Nucleic acid research* 28(19):3771–3778.

[105] Sallmyr A, Tomkinson AE (2018) Repair of DNA double-strand breaks by mammalian alternative end-joining pathways. *Journal of Biological Chemistry* 293(27):10536–10549.

[106] Verma P, Greenberg RA (2016) Noncanonical views of homology-directed DNA repair. *Genes and Development* 30(10):1138–1154.

[107] Ma JL, Kim EM, Haber JE, Lee SE (2003) Yeast Mre11 and Rad1 proteins define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences. *Molecular and Cellular Biology* 23(23):8820–8828.

[108] Audebert M, Salles B, Weinfeld M, Calsou P (2006) Involvement of polynucleotide kinase in a poly(ADP-ribose) polymerase-1-dependent DNA double-strand breaks rejoining pathway. *Journal of Molecular Biology* 356(2):257–265.

[109] Liang L, et al. (2008) Human DNA ligases I and III, but not ligase IV, are required for microhomology-mediated end joining of DNA double-strand breaks. *Nucleic Acids Research* 36(10):3297–3310.

[110] Boboila C, et al. (2010) Alternative end-joining catalyzes robust IgH locus deletions and translocations in the combined absence of ligase 4 and Ku70. *Proceedings of the National Academy of Sciences* 107(7):3034–3039.

[111] Kent T, Chandramouly G, Mcdevitt SM, Ozdemir AY, Pomerantz RT (2015) Mech-

anism of microhomology-mediated end-joining promoted by human DNA polymerase θ. *Nature Structural and Molecular Biology* 22(3):230–237.

[112] Roth DB, Wilson JH (1986) Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. *Molecular and cellular biology* 6(12):4295–4304.

[113] Boulton SJ, Jackson SP (1996) Identification of a Saccharomyces cerevisiae Ku80 homologue: roles in DNA double strand break rejoining and in telomeric maintenance. *Nucleic Acids Research* 24(23):4639–4648.

[114] Corneo B, et al. (2007) Rag mutations reveal robust alternative end joining. *Nature* 449(7161):483–486.

[115] Truong LN, et al. (2013) Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proceedings of the National Academy of Sciences* 110(19):7720–5.

[116] Ceccaldi R, et al. (2015) Homologous-recombination-deficient tumours are dependent on Polθ-mediated repair. *Nature* 518(7538):258–262.

[117] Sallmyr A, Tomkinson AE, Rassool FV (2008) Up-regulation of WRN and DNA ligase III in chronic myeloid leukemia: consequences for the repair of DNA double-strand breaks. *DNA Sequence* 112(4):1413–1423.

[118] Bentley J, Diggle CP, Harnden P, Knowles MA, Kiltie AE (2004) DNA double strand break repair in human bladder cancer is error prone and involves microhomology-associated end-joining. *Nucleic Acids Research* 32(17):5249–5259.

[119] Kostyrko K, Mermod N (2015) Assays for DNA double-strand break repair by microhomology-based end-joining repair mechanisms. *Nucleic Acids Research* 44(6):e56.

[120] Higgins GS, Boulton SJ (2018) Beyond PARP-POLθ as an anticancer target. *Science* 359(6381):1217–1218.

[121] McVey M, Lee SE (2008) MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in Genetics* 24(11):529–38.

[122] Yu AM, McVey M (2010) Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Research* 38(17):5706–5717.

[123] Sfeir A, Symington LS (2015) Microhomology-Mediated End Joining: A Backup Survival Mechanism or Dedicated Pathway? *Trends in Biochemical Sciences* 40(11):701–714.

[124] Ceccaldi R, Rondinelli B, D'Andrea AD (2016) Repair Pathway Choices and Consequences at the Double-Strand Break. *Trends in Cell Biology* 26(1):52–64.

[125] Chen C, Umezu K, Kolodner RD (1998) Chromosomal rearrangements occur in S. cerevisiae rfa1 mutator mutants due to mutagenic lesions processed by double-strand-break repair. *Molecular Cell* 2(1):9–22.

[126] Wu W, Wang M, Mussfeldt T, Iliakis G (2008) Enhanced Use of Backup Pathways of NHEJ in G2 in Chinese Hamster Mutant Cells with Defects in the Classical Pathway of NHEJ. *Radiation Research* 170(4):512–520.

[127] Wang H, et al. (2003) Biochemical evidence for Ku-independent backup pathways of NHEJ. *Nucleic Acids Research* 31(18):5377–5388.

[128] Della-Maria J, et al. (2011) Human Mre11/human Rad50/Nbs1 and DNA ligase III/XRCC1 protein complexes act together in an alternative nonhomologous end joining pathway. *Journal of Biological Chemistry* 286(39):33845–33853.

[129] Wang M, et al. (2006) PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. *Nucleic Acids Research* 34(21):6170–6182.

[130] Haince JF, et al. (2008) PARP1-dependent kinetics of recruitment of MRE11 and NBS1 proteins to multiple DNA damage sites. *Journal of Biological Chemistry* 283(2):1197–1208.

[131] Audebert M, Salles B, Calsou P (2008) Effect of double-strand break DNA sequence on the PARP-1 NHEJ pathway. *Biochemical and Biophysical Research Communications* 369(3):982–988.

[132] Mateos-Gomez PA, et al. (2015) Mammalian polymerase θ promotes alternative NHEJ and suppresses recombination. *Nature* 518(7538):254–257.

[133] Lee K, Lee SE (2007) Saccharomyces cerevisiae Sae2- and Tel1-dependent single-strand DNA formation at DNA break promotes microhomology-mediated end joining. *Genetics* 176(4):2003–2014.

[134] Han J, et al. (2017) BRCA2 antagonizes classical and alternative nonhomologous end-joining to prevent gross genomic instability. *Nature Communications* 8(1):1470.

[135] Deng SK, Gibb B, De Almeida MJ, Greene EC, Symington LS (2014) RPA an-

tagonizes microhomology-mediated repair of DNA double-strand breaks. *Nature Structural and Molecular Biology* 21(4):405–412.

[136] Mateos-Gomez PA, et al. (2017) The helicase domain of Polθ counteracts RPA to promote alt-NHEJ. *Nature Structural and Molecular Biology* 24(12):1116–1123.

[137] Wyatt DW, et al. (2016) Essential Roles for Polymerase θ-Mediated End Joining in the Repair of Chromosome Breaks. *Molecular Cell* 63(4):662–673.

[138] Villarreal DD, et al. (2012) Microhomology directs diverse DNA break repair pathways and chromosomal translocations. *PLoS Genet* 8(11):e1003026.

[139] Audebert M, Salles B, Calsou P (2004) Involvement of poly(ADP-ribose) polymerase-1 and XRCC1/DNA ligase III in an alternative route for DNA double-strand breaks rejoining. *Journal of Biological Chemistry* 279(53):55117–55126.

[140] Khodaverdian VY, et al. (2017) Secondary structure forming sequences drive SD-MMEJ repair of DNA double-strand breaks. *Nucleic Acids Research* 45(22):12848–12861.

[141] Kostyrko K, et al. (2017) MAR-Mediated transgene integration into permissive chromatin and increased expression by recombination pathway engineering. *Biotechnology and Bioengineering* 114(2):384–396.

[142] Saito S, Maeda R, Adachi N (2017) Dual loss of human POLQ and LIG4 abolishes random integration. *Nature Communications* 8:16112.

[143] Zelensky AN, Schimmel J, Kool H, Kanaar R, Tijsterman M (2017) Inactivation of Pol θ and C-NHEJ eliminates off-target integration of exogenous DNA. *Nature Communications* 8(1):66.

[144] Chan SH, Yu AM, McVey M (2010) Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila. *PLoS Genetics* 6(7):e1001005.

[145] Yousefzadeh MJ, et al. (2014) Mechanism of Suppression of Chromosomal Instability by DNA Polymerase POLQ. *PLOS Genetics* 10(10):e1004654.

[146] He P, Yang W (2018) Template and primer requirements for DNA Pol θ-mediated end joining. *Proceedings of the National Academy of Sciences* 115(30):7747–7752.

[147] Kent T, Mateos-Gomez PA, Sfeir A, Pomerantz RT (2016) Polymerase θ is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *eLife* 5:e13740.

[148] Simsek D, et al. (2011) DNA ligase III promotes alternative nonhomolo-

gous end-joining during chromosomal translocation formation. *PLoS Genetics* 7(6):e1002080.

[149] Paul K, et al. (2013) DNA Ligases I and III Cooperate in Alternative Non-Homologous End-Joining in Vertebrates. *PLoS ONE* 8(3):e59505.

[150] Delacôte F, et al. (2002) An xrcc4 defect or Wortmannin stimulates homologous recombination specifically induced by double-strand breaks in mammalian cells. *Nucleic Acids Res* 30(15):3454–3463.

[151] Beumer KJ, et al. (2008) Efficient gene targeting in Drosophila by direct embryo injection with zinc-finger nucleases. *Proceedings of the National Academy of Sciences* 105(50):19821–19826.

[152] Bozas A, Beumer KJ, Trautman JK, Carroll D (2009) Genetic analysis of zinc-finger nuclease-induced gene targeting in Drosophila. *Genetics* 182(3):641–651.

[153] Germani A, et al. (2003) SIAH-1 interacts with CtIP and promotes its degradation by the proteasome pathway. *Oncogene* 22(55):8845–8851.

[154] Yun MH, Hiom K (2009) CtIP-BRCA1 modulates the choice of DNA double-strand-break repair pathway throughout the cell cycle. *Nature* 459(7245):460–463.

[155] Huertas P, Jackson SP (2009) Human CtIP mediates cell cycle control of DNA end resection and double strand break repair. *Journal of Biological Chemistry* 284(14):9558–9565.

[156] Lee GS, Neiditch MB, Salus SS, Roth DB (2004) RAG proteins shepherd double-strand breaks to a specific pathway, suppressing error-prone repair, but RAG nicking initiates homologous recombination. *Cell* 117(2):171–184.

[157] Cole F, Keeney S, Jasin M (2010) Evolutionary conservation of meiotic DSB proteins: More than just Spo11. *Genes and Development* 24(12):1201–1207.

[158] Bothmer A, et al. (2017) Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nature Communications* 8:13905.

[159] Ling AK, et al. (2018) Double-stranded DNA break polarity skews repair pathway choice during intrachromosomal and interchromosomal recombination. *Proceedings of the National Academy of Sciences* 115(11):2800–2805.

[160] Morimatsu K, Kowalczykowski SC (2014) RecQ helicase and RecJ nuclease provide complementary functions to resect DNA for homologous recombination. *Proceedings of the National Academy of Sciences* 111(48):E5133–E5142.

[161] Chakraborty A, et al. (2016) Classical non-homologous end-joining pathway utilizes nascent RNA for error-free double-strand break repair of transcribed genes. *Nature Communications* 7:13049.

[162] Kakarougkas A, et al. (2013) Opposing roles for 53BP1 during homologous recombination. *Nucleic Acids Research* 41(21):9719–9731.

[163] Lemaître C, et al. (2014) Nuclear position dictates DNA repair pathway choice. *Genes & Development* 28(22):2450–2463.

[164] Finnie NJ, Gottlieb TM, Bluntt T, Jeggot PA, Jackson SP (1995) DNA-dependent protein kinase activity is absent in xrs-6 cells: Implications for site-specific recombination and DNA double-strand break repair. *Proceedings of the National Academy of Sciences* 92(1):320–324.

[165] Jackson SP (1997) DNA-dependent protein kinase. *The International Journal of Biochemistry & Cell Biology* 29(7):935–938.

[166] Ghezraoui H, et al. (2014) Chromosomal Translocations in Human Cells Are Generated by Canonical Nonhomologous End-Joining. *Molecular Cell* 55(6):829–842.

[167] Orii KE, Lee Y, Kondo N, McKinnon PJ (2006) Selective utilization of nonhomologous end-joining and homologous recombination DNA repair pathways during nervous system development. *Proceedings of the National Academy of Sciences* 103(26):10017–10022.

[168] Mujoo K, et al. (2017) Differentiation of Human Induced Pluripotent or Embryonic Stem Cells Decreases the DNA Damage Repair by Homologous Recombination. *Stem Cell Reports* 9(5):1660–1674.

[169] Bauer M, et al. (2011) Human monocytes are severely impaired in base and DNA double-strand break repair that renders them vulnerable to oxidative stress. *Proceedings of the National Academy of Sciences* 108(52):21105–21110.

[170] Bauer M, Goldstein M, Heylmann D, Kaina B (2012) Human monocytes undergo excessive apoptosis following temozolomide activating the ATM/ATR pathway while dendritic cells and macrophages are resistant. *PLoS ONE* 7(6):e39956.

[171] Orlando SJ, et al. (2010) Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic Acids Research* 38(15):e152.

[172] Robertson AB, Klungland A, Rognes T, Leiros I (2009) Base excision repair: The long and short of it. *Cellular and Molecular Life Sciences* 66(6):981–993.

[173] Krokan HE, Bjørås M (2013) Base Excision Repair. *Cold Spring Harbor Perspectives in Biology* 5(4):a012583.

[174] Jacobs AL, Schär P (2012) DNA glycosylases: In DNA repair and beyond. *Chromosoma* 121(1):1–20.

[175] Tell G, Quadrifoglio F, Tiribelli C, Kelley MR (2009) The Many Functions of APE1/Ref-1: Not Only a DNA Repair Enzyme. *Antioxidants & Redox Signaling* 11(3):601–619.

[176] Parsons JL, Dianova II, Allinson SL, Dianov GL (2005) Poly(ADP-ribose) polymerase-1 protects excessive DNA strand breaks from deterioration during repair in human cell extracts. *FEBS Journal* 272(8):2012–2021.

[177] Woodhouse BC, Dianova II, Parsons JL, Dianov GL (2008) Poly(ADP-ribose) polymerase-1 modulates DNA repair capacity and prevents formation of DNA double strand breaks. *DNA Repair* 7(6):932–940.

[178] Sobol RW, et al. (1996) Requirement of mammalian DNA polymerase-β in base-excision repair. *Nature* 379(6561):183–186.

[179] Cannan WJ, Rashid I, Tomkinson AE, Wallace SS, Pederson DS (2017) The human ligase IIIα-XRCC1 protein complex performs DNA nick repair after transient unwrapping of nucleosomal DNA. *Journal of Biological Chemistry* 292(13):5227–5238.

[180] Pascucci B, Stucki M, Jonsson ZO, Dogliotti E, Hubsher U (1999) Long patch base excision repair with purified proteins. *J.Biol.Chem.* 274(47):33696–33702.

[181] Klungland A, Lindahl T (1997) Second pathway for completion of human DNA base excision-repair: Reconstitution with purified proteins and requirement for DNase IV (FEN1). *EMBO Journal* 16(11):3341–3348.

[182] Schärer OD (2013) Nucleotide excision repair in eukaryotes. *Cold Spring Harbor Perspectives in Biology* 5(10):a012609.

[183] de Laat WL, Jaspers NGJ, Hoeijmakers JHJ (1999) Molecular mechanism of nucleotide excision repair. *Genes & Development* 13(7):768–785.

[184] Gillet LCJ, Schärer OD (2006) Molecular Mechanisms of Mammalian Global Genome Nucleotide Excision Repair. *Chemical Reviews* 106(2):253–276.

[185] Hanawalt PC, Spivak G (2008) Transcription-coupled DNA repair: Two decades of progress and surprises. *Nature Reviews Molecular Cell Biology* 9(12):958–970.

[186] Matsunaga T, Mu D, Park CH, Reardon JT, Sancar A (1995) Human DNA Re-

pair Excision Nuclease: ANALYSIS OF THE ROLES OF THE SUBUNITS IN-VOLVED IN DUAL INCISIONS BY USING ANTI-XPG AND ANTI-ERCC1 ANTIBODIES . *Journal of Biological Chemistry* 270(35):20862–20869.

[187] Moser J, et al. (2007) Sealing of Chromosomal DNA Nicks during Nucleotide Excision Repair Requires XRCC1 and DNA Ligase IIIα in a Cell-Cycle-Specific Manner. *Molecular Cell* 27(2):311–323.

[188] Jiricny J (2006) The multifaceted mismatch-repair system. *Nature Reviews Molecular Cell Biology* 7(5):335–346.

[189] Bak ST, Sakellariou D, Pena-Diaz J (2014) The dual nature of mismatch repair as antimutator and mutator: For better or for worse. *Frontiers in Genetics* 5:287.

[190] Tham KC, Kanaar R, Lebbink JH (2016) Mismatch repair and homeologous recombination. *DNA Repair* 38:75–83.

[191] Kunkel TA, Erie DA (2005) DNA MISMATCH REPAIR. *Annual Review of Biochemistry* 74(1):681–710.

[192] Acharya S, et al. (1996) hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. *Proceedings of the National Academy of Sciences* 93(24):13629–12634.

[193] Dzantiev L, et al. (2004) A defined human system that supports bidirectional mismatch-provoked excision. *Molecular Cell* 15(1):31–41.

[194] Williams HL, Gottesman ME, Gautier J (2013) The differences between ICL repair during and outside of S phase. *Trends in Biochemical Sciences* 38(8):386–393.

[195] Sarkar S, Davies AA, Ulrich HD, McHugh PJ (2006) DNA interstrand crosslink repair during G1 involves nucleotide excision repair and DNA polymerase ζ. *EMBO Journal* 25(6):1285–1294.

[196] Moldovan GL, D'Andrea AD (2009) How the Fanconi Anemia Pathway Guards the Genome. *Annual Review of Genetics* 43(1):223–249.

[197] Niedernhofer LJ, et al. (2004) The Structure-Specific Endonuclease Ercc1-Xpf Is Required To Resolve DNA Interstrand Cross-Link-Induced Double-Strand Breaks. *Molecular and Cellular Biology* 24(13):5776–5787.

[198] Muniandy PA, Liu J, Majumdar A, Liu ST, Seidman MM (2010) DNA interstrand crosslink repair in mammalian cells: Step by step. *Critical Reviews in Biochemistry and Molecular Biology* 45(1):23–49.

[199] Long DT, Räschle M, Joukov V, Walter JC (2011) Mechanism of RAD51-Dependent DNA Interstrand Cross-Link Repair. *Science* 333(6038):84–87.

[200] Caldecott KW (2008) Single-strand break repair and genetic disease. *Nature Reviews Genetics* 9(8):619–631.

[201] Cannan WJ, Pederson DS (2016) Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin. *Journal of Cellular Physiology* 231(1):3–14.

[202] Swindall AF, Stanley JA, Yang ES (2013) PARP-1: Friend or foe of DNA damage and repair in tumorigenesis? *Cancers* 5(3):943–958.

[203] Fisher AEO, Hochegger H, Takeda S, Caldecott KW (2007) Poly(ADP-Ribose) Polymerase 1 Accelerates Single-Strand Break Repair in Concert with Poly(ADP-Ribose) Glycohydrolase. *Molecular and Cellular Biology* 27(15):5597–5605.

[204] Walther W, Stein U (2000) Viral vectors for gene transfer: A review of their use in the treatment of human diseases. *Drugs* 60(2):249–271.

[205] Ivics Z, et al. (2009) Transposon-mediated genome manipulation in vertebrates. *Nature Methods* 6(6):415–422.

[206] Yin H, et al. (2014) Non-viral vectors for gene-based therapy. *Nature Reviews Genetics* 15(8):541–555.

[207] Folger KR, Wong EA, Wahl G, Capecchi MR (1982) Patterns of integration of DNA microinjected into cultured mammalian cells: evidence for homologous recombination between injected plasmid DNA molecules. *Molecular and Cellular Biology* 2(11):1372–87.

[208] Kohli A, Leech M, Vain P, Laurie DA, Christou P (1998) Transgene organization in rice engineered through direct DNA transfer supports a two-phase integration mechanism mediated by the establishment of integration hot spots. *Proceedings of the National Academy of Sciences* 95(12):7203–7208.

[209] Grandjean M, et al. (2011) High-level transgene expression by homologous recombination-mediated gene transfer. *Nucleic Acids Research* 39(15):e104.

[210] Vasquez KM, Marburger K, Intody Z, Wilson JH (2001) Manipulating the mammalian genome by homologous recombination. *Proceedings of the National Academy of Sciences* 98(15):8403–8410.

[211] Wong EA, Capecchi MR (1987) Homologous recombination between coinjected DNA sequences peaks in early to mid-S phase. *Molecular and Cellular Biology* 7(6):2294–2295.

[212] Dai J, Cui X, Zhu Z, Hu W (2010) Non-homologous end joining plays a key role in transgene concatemer formation in transgenic zebrafish embryos. *International Journal of Biological Sciences* 6(7):756–768.

[213] Lundberg R, Mavinakere M, Campbell C (2001) Deficient DNA End Joining Activity in Extracts from Fanconi Anemia Fibroblasts. *Journal of Biological Chemistry* 276(12):9543–9549.

[214] Hamilton AA, Thacker J (1987) Gene recombination in X-ray-sensitive hamster cells. *Molecular and Cellular Biology* 7(4):1409–1414.

[215] Würtele H, Little KCE, Chartrand P (2003) Illegitimate DNA integration in mammalian cells. *Gene Therapy* 10(21):1791–1799.

[216] Merrihew RV, Marburger K, Pennington SL, Roth DB, Wilson JH (1996) High-frequency illegitimate integration of transfected DNA at preintegrated target sites in a mammalian genome. *Molecular and Cellular Biology* 16(1):10–18.

[217] Chen ZY, He CY, Ehrhardt A, Kay MA (2003) Minicircle DNA vectors devoid of bacterial DNA result in persistent and high-level transgene expression in vivo. *Molecular Therapy* 8(3):495–500.

[218] Wilson C, Bellen HJ, Gehring WJ (1990) Position effects on eukaryotic gene-expression. *Annual Review of Cell Biology* 6:679–714.

[219] Smith KR (2001) Theoretical mechanisms in targeted and random integration of transgene DNA. *Reproduction, Nutrition, Development* 41(6):465–85.

[220] Thomas KR, Deng C, Capecchi MR (1992) High-fidelity gene targeting in embryonic stem cells by using sequence replacement vectors. *Mol Cell Biol* 12(7):2919–2923.

[221] Tichy ED, et al. (2010) Mouse embryonic stem cells, but not somatic cells, predominantly use homologous recombination to repair double-strand DNA breaks. *Stem Cells Dev* 19(11):1699–1711.

[222] Capecchi MR (2005) Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. *Nature Reviews Genetics* 6(6):507–512.

[223] Buerstedde JM, Takeda S (1991) Increased ratio of targeted to random integration after transfection of chicken B cell lines. *Cell* 67(1):179–188.

[224] Adachi N, et al. (2006) The Human Pre-B Cell Line Nalm-6 Is Highly Proficient in Gene Targeting by Homologous Recombination. *DNA and Cell Biology* 25(1):19–24.

[225] Silva G, et al. (2011) Meganucleases and Other Tools for Targeted Genome Engineering: Perspectives and Challenges for Gene Therapy. *Current Gene Therapy* 11(1):11–27.

[226] Gaj T, Gersbach CA, Barbas CF (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology* 31(7):397–405.

[227] Donoho G, Jasin M, Berg P (1998) Analysis of gene targeting and intrachromosomal homologous recombination stimulated by genomic double-strand breaks in mouse embryonic stem cells. *Molecular and Cellular Biology* 18(7):4070–8.

[228] Rouet P, Smih F, Jasin M (1994) Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proceedings of the National Academy of Sciences* 91(13):6064–6068.

[229] Taghian DG, Nickoloff JA (1997) Chromosomal double-strand breaks induce gene conversion at high frequency in mammalian cells. *Molecular and Cellular Biology* 17(11):6386–6393.

[230] Stoddard BL (2011) Homing endonucleases: From microbial genetic invaders to reagents for targeted DNA modification. *Structure* 19(1):7–15.

[231] Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD (2010) Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics* 11(9):636–646.

[232] Mussolino C, Cathomen T (2012) TALE nucleases: tailored genome engineering made easy. *Current Opinion in Biotechnology* 23(5):644–650.

[233] Doudna JA, Charpentier E (2014) The new frontier of genome engineering with CRISPR-Cas9. *Science* 346(6213):1258096–1258096.

[234] Stoddard BL (2005) Homing endonuclease structure and function. *Quarterly Reviews of Biophysics* 38(1):49–95.

[235] Gallagher DN, Haber JE (2018) Repair of a Site-Specific DNA Cleavage: Old-School Lessons for Cas9-Mediated Gene Editing. *ACS Chemical Biology* 13(2):397–405.

[236] Haber JE (2012) Mating-type genes and MAT switching in Saccharomyces cerevisiae. *Genetics* 191(1):33–64.

[237] Colleaux L, D'Auriol L, Galibert F, Dujon B (1988) Recognition and cleavage site of the intron-encoded omega transposase. *Proceedings of the National Academy of Sciences* 85(16):6022–6026.

[238] Chen Z, Wen F, Sun N, Zhao H (2009) Directed evolution of homing endonuclease

I-SceI with altered sequence specificity. *Protein Engineering, Design and Selection* 22(4):249–256.

[239] Mussolino C, Cathomen T (2013) RNA guides genome engineering. *Nature Biotechnology* 31(3):208–209.

[240] Joung JK, Sander JD (2013) TALENs: A widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology* 14(1):49–55.

[241] Peng Y, et al. (2014) Making designer mutants in model organisms. *Development* 141(21):4042–4054.

[242] Charpentier E, Doudna JA (2013) Biotechnology: Rewriting a genome. *Nature* 495(7439):50–51.

[243] Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157(6):1262–1278.

[244] Hille F, et al. (2018) The Biology of CRISPR-Cas: Backward and Forward. *Cell* 172(6):1239–1259.

[245] Ronda C, et al. (2014) Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. *Biotechnology and Bioengineering* 111(8):1604–1616.

[246] Mali P, et al. (2013) RNA-Guided Human Genome Engineering via Cas9. *Science* 339(6121):823–826.

[247] Cong L, et al. (2013) Multiplex Genome Engineering Using CRISPR/Cas System. *Science* 339(6121):819–823.

[248] Makarova KS, et al. (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology* 13(11):722–736.

[249] Wu WY, Lebbink JH, Kanaar R, Geijsen N, Van Der Oost J (2018) Genome editing by natural and engineered CRISPR-associated nucleases. *Nature Chemical Biology* 14(7):642–651.

[250] Zetsche B, et al. (2015) Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* 163(3):759–771.

[251] Abudayyeh OO, et al. (2016) C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353(6299):aaf5573.

[252] Cox DBT, et al. (2017) RNA editing with CRISPR-Cas13. *Science* 358(6366):1019–1027.

[253] Deltcheva E, et al. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471(7340):602–607.

[254] Jinek M, et al. (2012) A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* 337(6096):816–822.

[255] Ran FA, et al. (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature* 520(7546):186–191.

[256] Esvelt KM, et al. (2013) Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature Methods* 10(11):1116–1123.

[257] Chen F, et al. (2017) Targeted activation of diverse CRISPR-Cas systems for mammalian genome editing via proximal CRISPR targeting. *Nature Communications* 8:14958.

[258] Hsu PD, et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology* 31(9):827–832.

[259] Zhang Y, et al. (2014) Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. *Scientific Reports* 4:5405.

[260] Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* 507(7490):62–67.

[261] Gong S, Yu HH, Johnson KA, Taylor DW (2018) DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell Reports* 22(2):359–371.

[262] Lemos BR, et al. (2018) CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proceedings of the National Academy of Sciences* 115(9):E2040–E2047.

[263] Taheri-Ghahfarokhi A, et al. (2018) Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Research* 46(16):8417–8434.

[264] Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE (2016) Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature Biotechnology* 34(3):339–344.

[265] Singh D, et al. (2018) Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). *Proceedings of the National Academy of Sciences* 115(21):5444–5449.

[266] Szczelkun MD, et al. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proceedings of the National Academy of Sciences* 111(27):9798–9803.

[267] Revyakin A, Liu C, Ebright RH, Strick TR (2006) Abortive Initiation and Productive Initiation by RNA Polymerase Involve DNA Scrunching. *Science* 314(5802):1139–1143.

[268] Lee JS, Kallehauge TB, Pedersen LE, Kildegaard HF (2015) Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Scientific Reports* 5:8572.

[269] Lin Y, et al. (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research* 42(11):7473–7485.

[270] Schaefer KA, et al. (2017) Unexpected mutations after CRISPR–Cas9 editing in vivo. *Nature Methods* 14(6):547–548.

[271] Slaymaker IM, et al. (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science* 351(6268):84–88.

[272] Kleinstiver BP, et al. (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 529(7587):490–495.

[273] Ran F, et al. (2013) Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* 154(6):1380–1389.

[274] Tsai SQ, et al. (2014) Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature Biotechnology* 32(6):569–576.

[275] Guilinger JP, Thompson DB, Liu DR (2014) Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nature Biotechnology* 32(6):577–582.

[276] Kim H, Kim JS (2014) A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics* 15(5):321–334.

[277] van Overbeek M, et al. (2016) DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell* 63(4):633–646.

[278] Bae S, Kweon J, Kim HS, Kim Js (2014) Microhomology-based choice of Cas9 nuclease target sites. *Nature Methods* 11(7):705–706.

[279] Ata H, et al. (2018) Robust Activation of Microhomology-mediated End Joining for Precision Gene Editing Applications. *PLoS Genetics* 14(9):e1007652.

[280] Maresca M, Lin VG, Guo N, Yang Y (2013) Obligate Ligation-Gated Recombination (ObLiGaRe): Custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome Research* 23(3):539–546.

[281] Suzuki K, et al. (2016) In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* 540(7631):144–149.

[282] Nakade S, et al. (2014) Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9. *Nature Communications* 5:5560.

[283] Aida T, et al. (2016) Gene cassette knock-in in mammalian cells and zygotes by enhanced MMEJ. *BMC Genomics* 17(1):1–18.

[284] Richardson CD, et al. (2018) CRISPR–Cas9 genome editing in human cells occurs via the Fanconi anemia pathway. *Nature Genetics* 50(8):1132–1139.

[285] Hoshijima K, Jurynec MJ, Grunwald DJ (2016) Precise Editing of the Zebrafish Genome Made Simple and Efficient. *Developmental Cell* 36(6):654–667.

[286] Moehle EA, et al. (2007) Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proceedings of the National Academy of Sciences* 104(9):3055–3060.

[287] Byrne SM, Ortiz L, Mali P, Aach J, Church GM (2015) Multi-kilobase homozygous targeted gene replacement in human induced pluripotent stem cells. *Nucleic Acids Research* 43(3):e21.

[288] Hollywood JA, Lee CM, Scallan MF, Harrison PT (2016) Analysis of gene repair tracts from Cas9/gRNA double-stranded breaks in the human CFTR gene. *Scientific Reports* 6:32230.

[289] Kan Y, Ruis B, Lin S, Hendrickson EA (2014) The Mechanism of Gene Targeting in Human Somatic Cells. *PLoS Genetics* 10(4):e1004251.

[290] Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533(7603):420–424.

[291] Cristea S, et al. (2013) In vivo cleavage of transgene donors promotes nuclease-mediated targeted integration. *Biotechnology and Bioengineering* 110(3):871–880.

[292] Chen F, et al. (2011) High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat Methods* 8(9):753–755.

[293] Paix A, et al. (2017) Precision genome editing using synthesis-dependent repair of Cas9-induced DNA breaks. *Proceedings of the National Academy of Sciences* 114(50):E10745–E10754.

[294] Storici F, Snipe JR, Chan GK, Gordenin DA, Resnick MA (2006) Conservative

Repair of a Chromosomal Double-Strand Break by Single-Strand DNA through Two Steps of Annealing. *Molecular and Cellular Biology* 26(20):7645–7657.

[295] Davis L, Maizels N (2016) Two Distinct Pathways Support Gene Correction by Single-Stranded Donors at DNA Nicks. *Cell Reports* 17(7):1872–1881.

[296] Kan Y, Ruis B, Takasugi T, Hendrickson EA (2017) Mechanisms of precise genome editing using oligonucleotide donors. *Genome Research* 27(7):1099–1111.

[297] Sakuma T, Yamamoto T (2017) Magic wands of CRISPR—lots of choices for gene knock-in. *Cell Biology and Toxicology* 33(6):501–505.

[298] Ma H, et al. (2017) Correction of a pathogenic gene mutation in human embryos. *Nature* 548(7668):413–419.

[299] Kosicki M, Tomberg K, Bradley A (2018) Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* 36(8):765–771.

[300] National Academy of Sciences and National Academy of Medicine and National Academies of Sciences Engineering and Medicine (2017) *Human Genome Editing: Science, Ethics, and Governance*. (The National Academies Press, Washington, DC).

[301] Holkers M, et al. (2014) Adenoviral vector DNA for accurate genome editing with engineered nucleases. *Nature Methods* 11(10):1051–1057.

[302] Deng C, Capecchi MR (1992) Reexamination of gene targeting frequency as a function of the extent of homology between the targeting vector and the target locus. *Molecular and Cellular Biology* 12(8):3365–71.

[303] Zhang JP, et al. (2017) Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome Biology* 18(1):1–18.

[304] Carlson-Stevermer J, et al. (2017) Assembly of CRISPR ribonucleoproteins with biotinylated oligonucleotides via an RNA aptamer for precise gene editing. *Nature Communications* 8(1):1711.

[305] Savic N, et al. (2018) Covalent linkage of the DNA repair template to the CRISPR-Cas9 nuclease enhances homology-directed repair. *eLife* 7:e33761.

[306] Mali P, et al. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnology* 31(9):833–838.

[307] Lin S, Staahl B, Alla RK, Doudna JA (2014) Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* 3:e04766.

[308] Gutschner T, Haemmerle M, Genovese G, Draetta GF, Chin L (2016) Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. *Cell Reports* 14(6):1555–1566.

[309] Lee JS, Grav LM, Pedersen LE, Lee GM, Kildegaard HF (2016) Accelerated homology-directed targeted integration of transgenes in Chinese hamster ovary cells via CRISPR/Cas9 and fluorescent enrichment. *Biotechnology and Bioengineering* 113(11):2518–2523.

[310] Vispé S, Cazaux C, Lesca C, Defais M (1998) Overexpression of Rad51 protein stimulates homologous recombination and increases resistance of mammalian cells to ionizing radiation. *Nucleic Acids Research* 26(12):2859–2864.

[311] Yáñez RJ, Porter AC (1999) Gene targeting is enhanced in human cells overexpressing hRAD51. *Gene Therapy* 6(7):1282–1290.

[312] Pierce AJ, Johnson RD, Thompson LH, Jasin M (1999) XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes and Development* 13(20):2633–2638.

[313] Kalvala A, et al. (2010) Enhancement of gene targeting in human cells by intranuclear permeation of the Saccharomyces cerevisiae Rad52 protein. *Nucleic Acids Research* 38(14):e149.

[314] Pinder J, Salsman J, Dellaire G (2015) Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Research* 43(19):9379–9392.

[315] Song J, et al. (2016) RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. *Nature Communications* 7:10548.

[316] Kim PM, Allen C, Wagener BM, Shen Z, Nickoloff JA (2001) Overexpression of human RAD51 and RAD52 reduces double-strand break-induced homologous recombination in mammalian cells. *Nucleic acids research* 29(21):4352–60.

[317] Fattah FJ, Lichter NF, Fattah KR, Oh S, Hendrickson EA (2008) Ku70, an essential gene, modulates the frequency of rAAV-mediated gene targeting in human somatic cells. *Proceedings of the National Academy of Sciences* 105(25):8703–8708.

[318] Liang F, et al. (1996) Chromosomal double-strand break repair in Ku80-deficient cells. *Proceedings of the National Academy of Sciences* 93(17):8929–8933.

[319] Bertolini LR, Bertolini M, Maga EA, Madden KR, Murray JD (2009) Increased gene targeting in Ku70 and Xrcc4 transiently deficient human somatic cells. *Molecular Biotechnology* 41(2):106–114.

[320] Chu VT, et al. (2015) Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature Biotechnology* 33(5):543–548.

[321] Riesenberg S, Maricic T (2018) Targeting repair pathways with small molecules increases precise genome editing in pluripotent stem cells. *Nature Communications* 9(1):2164.

[322] Bao Z, Cobb RE, Zhao H (2016) Accelerated genome engineering through multiplexing. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 8(1):5–21.

[323] Grav LM, et al. (2015) One-step generation of triple knockout CHO cell lines using CRISPR/Cas9 and fluorescent enrichment. *Biotechnology Journal* 10(9):1446–1456.

[324] Sakuma T, Nishikawa A, Kume S, Chayama K, Yamamoto T (2014) Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. *Scientific Reports* 4:5400.

[325] Yang L, et al. (2015) Genome-wide inactivation of porcine endogenous retroviruses (PERVs). *Science* 350(6264):1101 – 1104.

[326] Niu D, et al. (2017) Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science* 357(6357):1303 – 1307.

[327] Semaan M, Ivanusic D, Denner J (2015) Cytotoxic effects during knock out of multiple Porcine Endogenous Retrovirus (PERV) sequences in the pig genome by Zinc Finger Nucleases (ZFN). *PLoS ONE* 10(4):e0122059.

[328] Aguirre AJ, et al. (2016) Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discovery* 6(8):914–929.

[329] Dumont J, Euwart D, Mei B, Estes S, Kshirsagar R (2016) Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. *Critical Reviews in Biotechnology* 36(6):1110–1122.

[330] Walsh G (2018) Biopharmaceutical benchmarks 2018. *Nature Biotechnology* 36(12):1136–1145.

[331] Walsh G (2009) Post-Translational Modifications in the Context of Therapeutic Proteins: An Introductory Overview in *Post-translational Modification of Protein Biopharmaceuticals*, Wiley Online Books, ed. Walsh G. (John Wiley & Sons, Ltd), pp. 1–76.

[332] Hartley JW, Rowe WP (1976) Naturally occurring murine leukemia viruses in wild mice: characterization of a new "amphotropic" class. *Journal of Virology* 19(1):19–25.

[333] Shepherd AJ, Wilson NJ, Smith KT (2003) Characterisation of endogenous retrovirus in rodent cell lines used for production of biologicals. *Biologicals* 31(4):251–260.

[334] Berting A, Farcet MR, Kreil TR (2010) Virus susceptibility of Chinese hamster ovary (CHO) cells and detection of viral contaminations by adventitious agent testing. *Biotechnology and Bioengineering* 106(4):598–607.

[335] Xu L, et al. (2014) Role of Risk Assessments in Viral Safety: An FDA Perspective in *PDA Journal of Pharmaceutical Science and Technology*. Vol. 68, pp. 6–10.

[336] Chu L, Robinson DK (2001) Industrial choices for protein production by large-scale cell culture. *Current Opinion in Biotechnology* 12(2):180–187.

[337] Garnick RL (1996) Experience with viral contamination in cell culture. *Developments in Biological Standardization* 88:49–56.

[338] Dewannieux M, Ribet D, Heidmann T (2010) Risks linked to endogenous retroviruses for vaccine production: A general overview. *Biologicals* 38(3):366–370.

[339] Consortium IHGS, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

[340] Consortium MGS, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

[341] Stocking C, Kozak CA (2008) Endogenous retroviruses: Murine endogenous retroviruses. *Cellular and Molecular Life Sciences* 65(21):3383–3398.

[342] Aiewsakun P, Katzourakis A (2015) Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* 479-480:26–37.

[343] Ahlquist P (2006) Parallels among positive-strand RNA viruses, reverse-

transcribing viruses and double-stranded RNA viruses. *Nature Reviews Microbiology* 4(5):371–382.

[344] Stoye JP (2012) Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews Microbiology* 10(6):395–406.

[345] Jern P, Coffin JM (2008) Effects of Retroviruses on Host Genome Function. *Annual Review of Genetics* 42(1):709–732.

[346] Bénit L, Dessen P, Heidmann T (2001) Identification, Phylogeny, and Evolution of Retroviral Elements Based on Their Envelope Genes. *Journal of Virology* 75(23):11709–11719.

[347] Thompson PJ, Macfarlan TS, Lorincz MC (2016) Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. *Molecular Cell* 62(5):766–776.

[348] Swanstrom R, Wills JW (1997) Synthesis, Assembly, and Processing of Viral Proteins in *Retroviruses*, ed. Coffin JM, Hughes SH VH. (Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press), p. Available from: https://www.ncbi.nlm.nih.gov/books.

[349] Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics* 13:283.

[350] Kassiotis G (2014) Endogenous Retroviruses and the Development of Cancer. *The Journal of Immunology* 192(4):1343–1349.

[351] Akiyoshi DE, et al. (1998) Identification of a full-length cDNA for an endogenous retrovirus of miniature swine. *Journal of Virology* 72(5):4503–4507.

[352] Lie YS, et al. (1994) Chinese hamster ovary cells contain transcriptionally active full-length type C proviruses. *Journal of Virology* 68(12):7840–7849.

[353] Turner G, et al. (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current Biology* 11(19):1531–1535.

[354] Dinowitz M, et al. (1992) Recent studies on retrovirus-like particles in Chinese hamster ovary cells. *Developments in biological standardization* 76:201–207.

[355] Tihon C, Green M (1973) Cyclic AMP-amplified Replication of RNA Tumour Virus-like Particles in Chinese Hamster Ovary Cells. *Nature New Biology* 244:227.

[356] Brorson K, et al. (2002) Impact of cell culture process changes on endogenous retrovirus expression. *Biotechnology and Bioengineering* 80(3):257–267.

[357] Taruscio D, Mantovani A (2004) Factors regulating endogenous retroviral se-

quences in human and mouse. *Cytogenetic and Genome Research* 105(2-4):351–362.

[358] Contreras-Galindo R, López P, Vélez R, Yamamura Y (2007) HIV-1 Infection Increases the Expression of Human Endogenous Retroviruses Type K (HERV-K) in Vitro. *AIDS Research and Human Retroviruses* 23(1):116–122.

[359] Ogawa T, Watanabe Y, Meshi T, Okada Y (1991) Trans complementation of virus-encoded replicase components of tobacco mosaic virus. *Virology* 185(2):580–584.

[360] Fu Y, et al. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology* 31(9):822–826.

[361] Cormack B (1998) Directed Mutagenesis Using the Polymerase Chain Reaction. *Current Protocols in Neuroscience* 3(1):4.11.1–4.11.10.

[362] Ley D, et al. (2013) MAR Elements and Transposons for Improved Transgene Integration and Expression. *PLoS ONE* 8(4):e62784.

[363] Le Fourn V, Girod PA, Buceta M, Regamey A, Mermod N (2014) CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. *Metabolic Engineering* 21:91–102.

[364] Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research* 29(9):e45.

[365] Agudelo D, et al. (2017) Marker-free coselection for CRISPR-driven genome editing in human cells. *Nature Methods* 14(6):615–620.

[366] Ochs F, et al. (2016) 53BP1 fosters fidelity of homology-directed DNA repair. *Nature Structural and Molecular Biology* 23(8):714–721.

[367] Stark JM, et al. (2002) ATP hydrolysis by mammalian RAD51 has a key role during homology-directed DNA repair. *Journal of Biological Chemistry* 277(23):20185–20194.

[368] Wurm F (2013) CHO Quasispecies—Implications for Manufacturing Processes. *Processes* 1(3):296–311.

[369] Kostyrko K, Bosshard S, Urban Z, Mermod N (2015) A role for homologous recombination proteins in cell cycle regulation. *Cell Cycle* 14(17):2853–2861.

[370] Ran FA, et al. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protocols* 8(11):2281–2308.

[371] Paddock MN, Buelow BD, Takeda S, Scharenberg AM (2010) The BRCT Do-

main of PARP-1 Is Required for Immunoglobulin Gene Conversion. *PLOS Biology* 8(7):e1000428.

[372] Fishman-Lobell J, Haber JE (1992) Removal of nonhomologous DNA ends in double-strand break recombination: the role of the yeast ultraviolet repair gene RAD1. *Science* 258(5081):480 – 484.

[373] Ivanov EL, Haber JE (1995) RAD1 and RAD10, but not other excision repair genes, are required for double-strand break-induced recombination in Saccharomyces cerevisiae. *Molecular and Cellular Biology* 15(4):2245–2251.

[374] Al-Minawi AZ, Saleh-Gohari N, Helleday T (2008) The ERCC1/XPF endonuclease is required for efficient single-strand annealing and gene conversion in mammalian cells. *Nucleic Acids Research* 36(1):1–9.

[375] Tomkinson AE, Howes TRL, Wiest NE (2013) DNA ligases as therapeutic targets. *Translational Cancer Research* 2(3):203–214.

[376] Harrison JC, Haber JE (2006) Surviving the Breakup: The DNA Damage Checkpoint. *Annual Review of Genetics* 40(1):209–235.

[377] Moreno-Mateos MA, et al. (2015) CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nature Methods* 12(10):982–988.

[378] Misteli T, Soutoglou E (2009) The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nature Reviews Molecular Cell Biology* 10(4):243.

[379] Hartlerode AJ, Willis NA, Rajendran A, Manis JP, Scully R (2016) Complex Breakpoints and Template Switching Associated with Non-canonical Termination of Homologous Recombination in Mammalian Cells. *PLoS Genetics* 12(11):e1006410.

[380] Choi YJ, et al. (2014) Deletion of Individual Ku Subunits in Mice Causes an NHEJ-Independent Phenotype Potentially by Altering Apurinic/Apyrimidinic Site Repair. *PLOS ONE* 9(1):e86358.

[381] Fell VL, Schild-Poulter C (2012) Ku Regulates Signaling to DNA Damage Response Pathways through the Ku70 von Willebrand A Domain. *Molecular and Cellular Biology* 32(1):76–87.

[382] Yamaguchi-Iwai Y, et al. (1999) Mre11 is essential for the maintenance of chromosomal DNA in vertebrate cells. *The EMBO Journal* 18(23):6619–6629.

[383] Bressan DA, Baxter BK, Petrini JH (1999) The Mre11-Rad50-Xrs2 protein complex facilitates homologous recombination-based double-strand break repair in Saccharomyces cerevisiae. *Molecular and Cellular Biology* 19(11):7681–7687.

[384] Roques C, et al. (2009) MRE11–RAD50–NBS1 is a critical regulator of FANCD2 stability and function during DNA double-strand break repair. *The EMBO Journal* 28(16):2400–2413.

[385] Shy BR, Macdougall MS, Clarke R, Merrill BJ (2016) Co-incident insertion enables high efficiency genome engineering in mouse embryonic stem cells. *Nucleic Acids Research* 44(16):7997–8010.

[386] Emanoil-Ravier R, et al. (1991) Biological and molecular studies of endogenous retrovirus-like genes in Chinese hamster cell lines. *Developments in Biological Standardization* 75:113–122.

[387] Hojman F, Emanoil-Raivier R, Lesser J, Périès J (1989) Biological and molecular characterization of an endogenous retrovirus present in CHO/HBs-A Chinese hamster cell line. *Developments in Biological Standardization* 70:195–202.

[388] Anderson KP, Low MA, Lie YS, Keller GA, Dinowitz M (1991) Endogenous origin of defective retroviruslike particles from a recombinant Chinese hamster ovary cell line. *Virology* 181(1):305–311.

[389] Gould RR, Borisy GG (1977) The pericentriolar material in Chinese hamster ovary cells nucleates microtubule formation. *The Journal of Cell Biology* 73(3):601–615.

[390] Heine UI, Kramarsky B, Wendel E, Suskind RG (1979) Enhanced Proliferation of Endogenous Virus in Chinese Hamster Cells Associated with Microtubules and the Mitotic Apparatus of the Host Cell. *Journal of General Virology* 44(1):45–55.

[391] Manly KF, Givens JF, Taber RL, Zeigel RF (1978) Characterization of Virus-like Particles Released from the Hamster Cell Line CHO-K1 After Treatment with 5-Bromodeoxyuridine. *Journal of General Virology* 39(3):505–517.

[392] Reuss FU (1992) Expression of intracisternal A-particle-related retroviral element-encoded envelope proteins detected in cell lines. *Journal of Virology* 66(4):1915–1923.

[393] Anderson KP, et al. (1990) Presence and transcription of intracisternal A-particle-related sequences in CHO cells. *Journal of Virology* 64(5):2021 – 2032.

[394] Donahue RE, et al. (1992) Helper virus induced T cell lymphoma in nonhuman primates after retroviral mediated gene transfer. *The Journal of Experimental Medicine* 176:1125–1135.

[395] Urnovitz HB, Murphy WH (1996) Human endogenous retroviruses: nature, occur-

rence, and clinical implications in human disease. *Clinical Microbiology Reviews* 9(1):72–99.

[396] Kaminski R, et al. (2016) Elimination of HIV-1 Genomes from Human T-lymphoid Cells by CRISPR/Cas9 Gene Editing. *Scientific Reports* 6:22555.

[397] Zhu LJ, Holmes BR, Aronin N, Brodsky MH (2014) CRISPRseek: A Bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. *PLoS ONE* 9(9):e108424.

[398] Xu H, et al. (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Research* 25(8):1147–1157.

[399] Chari R, Mali P, Moosburner M, Church GM (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nature Methods* 12(9):823–826.

[400] Sander JD, Zaback P, Joung JK, Voytas DF, Dobbs D (2007) Zinc Finger Targeter (ZiFiT): An engineered zinc finger/target site design tool. *Nucleic Acids Research* 35:599–605.

[401] Sander JD, et al. (2010) ZiFiT (Zinc Finger Targeter): An updated zinc finger engineering tool. *Nucleic Acids Research* 38:462–468.

[402] Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.

[403] Brinkman EK, Chen T, Amendola M, Van Steensel B (2014) Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Research* 42(22):1–8.

[404] Yoon H, Leitner T (2015) PrimerDesign-M: A multiple-alignment based multiple-primer design tool for walking across variable genomes. *Bioinformatics* 31(9):1472–1474.

[405] Fadrosh DW, et al. (2014) An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* 2(1):6.

[406] Morikawa Y, et al. (1996) Complete Inhibition of Human Immunodeficiency Virus Gag Myristoylation Is Necessary for Inhibition of Particle Budding. *Journal of Biological Chemistry* 271(5):2868–2873.

[407] Wapling J, Srivastava S, Shehu-Xhilaga M, Tachedjian G (2007) Targeting Human Immunodeficiency Virus Type 1 Assembly, Maturation and Budding. *Drug Target Insights* 2:159–182.

[408] Henzy JE, Gifford RJ, Johnson WE, Coffin JM (2014) A Novel Recombinant Retrovirus in the Genomes of Modern Birds Combines Features of Avian and Mammalian Retroviruses. *Journal of Virology* 88(5):2398–2405.

[409] Segura-Morales C, et al. (2005) Tsg101 and Alix Interact with Murine Leukemia Virus Gag and Cooperate with Nedd4 Ubiquitin Ligases during Budding. *Journal of Biological Chemistry* 280(29):27004–27012.

[410] Daer RM, Cutts JP, Brafman DA, Haynes KA (2017) The Impact of Chromatin Dynamics on Cas9-Mediated Genome Editing in Human Cells. *ACS Synthetic Biology* 6(3):428–438.

[411] Gosselin K, et al. (2009) Senescent keratinocytes die by autophagic programmed cell death. *The American Journal of Pathology* 174(2):423–435.

[412] Byrne G, et al. (2018) CRISPR/Cas9 gene editing for the creation of an MGAT1-deficient CHO cell line to control HIV-1 vaccine glycosylation. *PLoS Biology* 16(8):e2005817.

[413] Ihry RJ, et al. (2018) P53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nature Medicine* 24(7):939–946.

[414] Wang T, et al. (2015) Identification and characterization of essential genes in the human genome. *Science* 350(6264):1096–1101.

[415] O'Connor MJ (2015) Targeting the DNA Damage Response in Cancer. *Molecular Cell* 60(4):547–560.

[416] Guirouilh-Barbat, Lambert S, Bertrand P, Lopez BS (2014) Is homologous recombination really an error-free process? *Frontiers in genetics* 5:175.

[417] Lee JS, et al. (2018) Revealing Key Determinants of Clonal Variation in Transgene Expression in Recombinant CHO Cells Using Targeted Genome Editing. *ACS Synthetic Biology* 7(12):2867–2878.

[418] Pilbrough W, Munro TP, Gray P (2009) Intraclonal protein expression heterogeneity in recombinant CHO cells. *PLoS ONE* 4(12):e8432.

[419] Sigal A, et al. (2006) Variability and memory of protein levels in human cells. *Nature* 444(7119):643–646.

[420] Hu W, et al. (2014) RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. *Proceedings of the National Academy of Sciences* 111(31):11461–11466.

[421] Schmieder V, et al. (2018) Enhanced Genome Editing Tools For Multi-Gene

Deletion Knock-Out Approaches Using Paired CRISPR sgRNAs in CHO Cells. *Biotechnology Journal* 13(3):1700211.

[422] Kawada S, Goto T, Haraguchi H, Ono A, Morikawa Y (2008) Dominant Negative Inhibition of Human Immunodeficiency Virus Particle Production by the Non-myristoylated Form of Gag. *Journal of Virology* 82(9):4384–4399.

[423] Manrique ML, Celma CC, González SA, Affranchino JL (2001) Mutational analysis of the feline immunodeficiency virus matrix protein. *Virus Research* 76(1):103–113.

[424] Puri V, et al. (2007) Fat-specific Protein 27, a Novel Lipid Droplet Protein That Enhances Triglyceride Storage. *Journal of Biological Chemistry* 282(47):34213–34218.

[425] Boztug K, et al. (2014) JAGN1 deficiency causes aberrant myeloid cell homeostasis and congenital neutropenia. *Nature Genetics* 46(9):1021–1027.

[426] Caldecott KW (2003) XRCC1 and DNA strand break repair. *DNA Repair* 2(9):955–969.

[427] Hicks WM, Kim M, Haber JE (2010) Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science* 329(5987):82–85.

[428] Hastings PJ, Ira G, Lupski JR (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics* 5(1):e1000327.

[429] Lee JA, Carvalho CM, Lupski JR (2007) A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell* 131(7):1235–1247.

[430] Derks KW, Hoeijmakers JH, Pothof J (2014) The DNA damage response: The omics era and its impact. *DNA Repair* 19:214–220.

[431] Wilson JH, Leung WY, Bosco G, Dieu D, Haber JE (1994) The frequency of gene targeting in yeast depends on the number of target copies. *Proceedings of the National Academy of Sciences* 91(1):177 – 181.

[432] Chung WH, Zhu Z, Papusha A, Malkova A, Ira G (2010) Defective Resection at DNA Double-Strand Breaks Leads to De Novo Telomere Formation and Enhances Gene Targeting. *PLOS Genetics* 6(5):e1000948.

[433] Canny MD, et al. (2017) Inhibition of 53BP1 favors homology-dependent DNA re-

pair and increases CRISPR–Cas9 genome-editing efficiency. *Nature Biotechnology* 36(1):95–102.

[434] Thompson DB, et al. (2018) The Future of Multiplexed Eukaryotic Genome Engineering. *ACS Chemical Biology* 13(2):313–325.

[435] Lokossou AG, Toudic C, Barbeau B (2014) Implication of human endogenous retrovirus envelope proteins in placental functions. *Viruses* 6(11):4609–4627.

[436] Carroll D (2017) Genome Editing : Past, Present, and Future. *Yale Journal of Biology and Medicine* 90(4):653–659.

[437] (2019) ClinicalTrials.gov.

[438] Datta P, Linhardt RJ, Sharfstein ST (2013) An 'omics approach towards CHO cell engineering. *Biotechnology and Bioengineering* 110(5):1255–1271.

[439] Cost GJ, et al. (2010) BAK and BAX deletion using zinc-finger nucleases yields apoptosis-resistant CHO cells. *Biotechnology and Bioengineering* 105(2):330–340.

[440] Zhou M, et al. (2011) Decreasing lactate level and increasing antibody production in Chinese Hamster Ovary cells (CHO) by reducing the expression of lactate dehydrogenase and pyruvate dehydrogenase kinases. *Journal of Biotechnology* 153(1):27–34.

[441] Fan L, et al. (2012) Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. *Biotechnology and Bioengineering* 109(4):1007–1015.

[442] Sakuma T, et al. (2015) Homologous recombination-independent large gene cassette knock-in in CHO cells using TALEN and MMEJ-directed donor plasmids. *International Journal of Molecular Sciences* 16(10):23849–23866.

[443] Gaidukov L, et al. (2018) A multi-landing pad DNA integration platform for mammalian cell engineering. *Nucleic Acids Research* 46(8):4072–4086.

[444] Estes S, Melville M (2014) Mammalian Cell Line Developments in Speed and Efficiency BT in *Mammalian Cell Cultures for Biologics Manufacturing*, eds. Zhou W, Kantardjieff A. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 11–33.

Appendix

Patent application

Mermod N., Duroy P.O., **Bosshard S.**, Le Mercier P. (2015) Improved eukaryotic cells for protein manufacturing and methods of making them. *Patent application*, WO2017109177 A1.

Publications

Kostyrko K., Neuenschwander S., Junier T., Regamey A., Iseli C., Schmid-Siegert E., **Bosshard S.**, Majocchi S., Girod P.A., Xenarios I., Mermod N. (2017) MAR-mediated transgene integration into permissive chromatin and increased expression by recombination pathway engineering. *Biotechnology and Bioengineering*, **114**(2), 384–396.

Kostyrko K., **Bosshard S.**, Urban Z., Mermod N. (2015) A role for homologous recombination proteins in cell cycle regulation. *Cell Cycle*, **14**(17), 2853-2861.

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

(51) International Patent Classification:
C12N 15/10 (2006.01)   *C12P 21/00* (2006.01)
C12N 15/90 (2006.01)   *C12N 5/071* (2010.01)

(21) International Application Number:
PCT/EP2016/082567

(22) International Filing Date:
23 December 2016 (23. 12.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/387,375   24 December 201 5 (24. 12.2015)   US

(71) Applicant: SELEXIS S.A. [CH/CH]; 18 Chemin des Aulx, 1228 Plan-les-Ouates (CH).

(72) Inventors: MERMOD, Nicolas; c/o Selexis S.A., 18 Chemin des Aulx, 1228 Plan-les-Ouates (CH). DUROY, Pierre-Oliver; c/o Selexis S.A., 18 Chemin des Aulx, 1228 Plan-les-Ouates (FR). BOSSHARD, Sandra; c/o Selexis S.A., 18 Chemin des Aulx, 1228 Plan-les-Ouates (CH). LE MERCIER, Philippe; c/o Selexis S.A., 18 Chemin des Aulx, 1228 Plan-les-Ouates (FR).

(74) Agent: IPRIS AG; Pelikanweg 2, 4054 Basel (CH).

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

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

Published:
— *with international search report (Art. 21(3))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: IMPROVED EUKARYOTIC CELLS FOR PROTEIN MANUFACTURING AND METHODS OF MAKING THEM
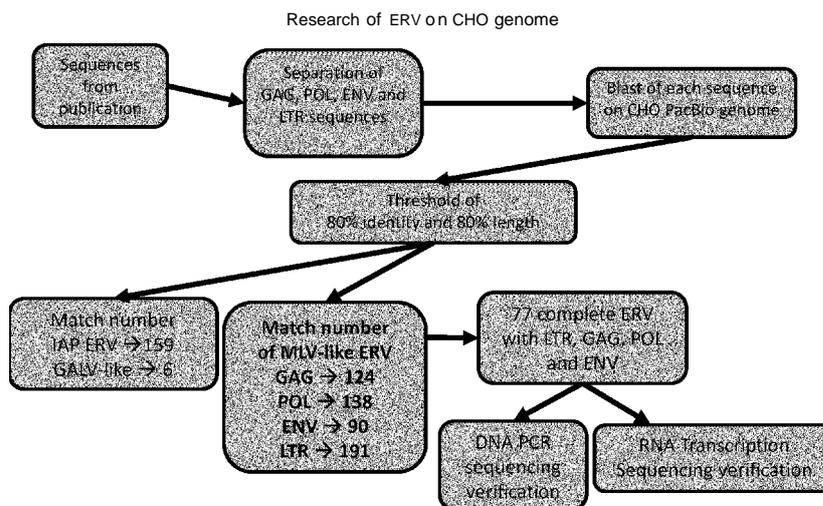
FIGURE 4A

(57) Abstract: Disclosed are mammalian cells and mammalian cell lines that have a reduced load of remnants of past viral/retroviral infections and methods of producing and using the same.

WO 2017/109177 A1

BIOTECHNOLOGY
and
BIOENGINEERING

MAR-Mediated Transgene Integration Into Permissive Chromatin and Increased Expression by Recombination Pathway Engineering

Kaja Kostyrko,[1] Samuel Neuenschwander,[2] Thomas Junier,[2] Alexandre Regamey,[3] Christian Iseli,[2] Emanuel Schmid-Siegert,[2] Sandra Bosshard,[1] Stefano Majocchi,[1] Valérie Le Fourn,[3] Pierre-Alain Girod,[3] Ioannis Xenarios,[2] Nicolas Mermod[1]

[1]Department of Fundamental Microbiology, Institute of Biotechnology, University of Lausanne, and Center for Biotechnology
UNIL-EPFL, Lausanne, Switzerland; telephone: +41-21-693-61-51; fax: +41-21-693-76-10;
e-mail: nicolas.mermod@unil.ch
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[3]Selexis SA, Geneva, Switzerland

ABSTRACT: Untargeted plasmid integration into mammalian cell genomes remains a poorly understood and inefficient process. The formation of plasmid concatemers and their genomic integration has been ascribed either to non-homologous end-joining (NHEJ) or homologous recombination (HR) DNA repair pathways. However, a direct involvement of these pathways has remained unclear. Here, we show that the silencing of many HR factors enhanced plasmid concatemer formation and stable expression of the gene of interest in Chinese hamster ovary (CHO) cells, while the inhibition of NHEJ had no effect. However, genomic integration was decreased by the silencing of specific HR components, such as Rad51, and DNA synthesis-dependent microhomology-mediated end-joining (SD-MMEJ) activities. Genome-wide analysis of the integration loci and junction sequences validated the prevalent use of the SD-MMEJ pathway for transgene integration close to cellular genes, an effect shared with matrix attachment region (MAR) DNA elements that stimulate plasmid integration and expression. Overall, we conclude that SD-MMEJ is the main mechanism driving the illegitimate genomic integration of foreign DNA in CHO cells, and we provide a recombination engineering approach that increases transgene integration and recombinant protein expression in these cells.

Introduction

Spontaneous integration of non-viral DNA vectors into the genome of eukaryotic cells is a widely exploited process in research and biotechnology. Its molecular basis, however, remains incompletely understood. It is believed to rely on cellular DNA repair mechanisms, as it is favored by the presence of free DNA ends in the vector resembling double stranded breaks (DSBs). The two major pathways responsible for DSB repair in eukaryotic cells are non-homologous end-joining (NHEJ) and homologous recombination (HR) (Jackson, 2002). NHEJ is a fast mechanism that efficiently joins DNA ends with little processing (Mao et al., 2008). In contrast, HR is a slow, multi-step process requiring resection of one of the two DNA strands and pairing to a homologous DNA template for repair. A third group of DSB repair pathways, believed to function when the main repair mechanisms are impaired, are collectively termed microhomology-mediated end joining (MMEJ). MMEJ is a still poorly characterized family of pathways, also referred to as alternative or backup non-homologous end-joining (alt- or B-NHEJ), which requires short (2–25 nt) homologies to align broken DNA strands before joining (Boboila et al., 2010; Gigi et al., 2014; Oh et al., 2014; Paul et al., 2013). Another hallmark of this process is the occurrence of large deletions and, less frequently, insertions of sequences copied from other parts of the genome, termed templated inserts (Ma et al., 2003; Merrihew et al., 1996). MMEJ shares DNA strand resection with HR, implying that it may partially rely on HR enzymes (Decottignies, 2007; Dinkelmann et al., 2009; Ma et al., 2003; Truong et al., 2013). Several mechanisms proposed

to mediate chromosomal rearrangements associated with human genetic disorders were shown to rely on MMEJ (Costantino et al., 2014; Hastings et al., 2009; Hicks et al., 2010; Lee et al., 2007; Villarreal et al., 2012). Finally, another variant of MMEJ, termed synthesis-dependent MMEJ (SD-MMEJ), was also proposed to repair DSBs in the absence of pre-existing homology (Yu and McVey, 2010). In this latter mechanism, the microhomologies required for the MMEJ pathway are synthetized de novo by an accurate non-processive DNA polymerase. While all of these mechanisms may be mechanistically different, they possess several common features, such as the annealing of single stranded DNA ends at microhomology regions and the priming of low-processivity DNA polymerization.

Plasmid integration into the genome of eukaryotic cells is an overall inefficient process, occurring in a minor proportion of cells that take up the exogenous DNA. It was shown to involve two major steps: (i) recombination between vector molecules to form multimeric transgene arrays termed concatemers and (ii) the recombination of the resulting concatemers with the genome, usually at a single or at few chromosomal loci (Folger et al., 1982; Grandjean et al., 2011; Kohli et al., 1998). The DSB repair pathways responsible for transgene concatemerization remain currently unclear. In mammalian cells, this process was attributed to HR (Folger et al., 1982; Wong and Capecchi, 1987), while NHEJ appeared to be involved in zebrafish embryos and rice (Dai et al., 2010; Kohli et al., 1998). In addition, some studies suggested that alternative pathways may also play a role in the joining of extrachromosomal DNA ends (Lundberg et al., 2001). Similarly, the mechanism mediating the recombination of the transgene with the genome remains to be fully identified. NHEJ is considered to mediate the majority of integration events in eukaryotic cells, while HR may be responsible for a smaller proportion of genomic integrations (Würtele et al., 2003). However, there is evidence that distinct repair pathways may also be implicated in this process (Iiizumi et al., 2008; Merrihew et al., 1996).

We previously reported that plasmid integration is enhanced by the presence of matrix attachment regions (MARs), which are epigenetic regulatory DNA elements that participate in the formation of chromatin boundaries and augment transcription (Galbete et al., 2009; Girod et al., 2007; Grandjean et al., 2011; Majocchi et al., 2014). MARs are thus widely used to sustain elevated transgene expression, as well as to prevent epigenetic silencing effects by blocking the propagation of heterochromatin (Allen et al., 2000; Harraghy et al., 2008; Zahn-Zabal et al., 2001). Their action to increase genomic integration and plasmid copy number suggested that stimulating recombination may constitute an additional mechanism by which MARs increase transgene expression (Girod et al., 2007; Grandjean et al., 2011). Thus, in the present study, we sought to identify the pathway(s) responsible for the integration of MAR-containing or -devoid plasmids into the genome of cultured cells.

Using siRNA-mediated knock-down approach, we show that a subset of alternative repair mechanisms resembling SD-MMEJ may be preferentially used by CHO cells for the spontaneous integration of foreign DNA into their genome. This finding was confirmed by the characterization of plasmid-to-genome junction sequences, which were found to display an SD-MMEJ pattern. Finally, we demonstrate that MAR elements and SD-MMEJ favor transgene integration into permissive chromatin loci, and that the inhibition of competing recombination pathways can be used to improve the expression of recombinant proteins.

Materials and Methods

Cells, Plasmids, and siRNA

Adherent Chinese hamster ovary (CHO) DG44 cells (Urlaub and Chasin, 1980) were cultivated in DMEM/F-12 + GlutaMAX™ supplemented with 1× HT and 10% fetal bovine serum (Gibco, Invitrogen), and with the antibiotic-antimycotic solution (Sigma–Aldrich, #A5955). Suspension-adapted CHO K1 derived cells (CHO-M) were cultured in SFM4CHO (HyClone™) medium supplemented with 8 mM L-Glutamine (PAA Laboratories GmbH) and 1× HT (Gibco).

The MAR-devoid pGEGFP, MAR 1-68-containing p1-68-GFP, pGL3-CMV-DsRed, and pSVpuro expression vectors were described previously (Supplementary Fig. S1) (Grandjean et al., 2011). The HR and NHEJ reporter plasmids were kindly provided by V. Gorbunova (University of Rochester, New York) (Mao et al., 2008). The MMEJ-specific GFP reporter assay, based on the pGEGFP vector, was constructed as described previously (Kostyrko and Mermod, 2015). Small interfering RNA duplexes, specifically designed to target the CHO cell homologs of the DNA repair proteins listed in Tables SI and SII, were designed and provided by Microsynth AG (Balgach, Switzerland) (Supplementary Table SIII). Three RNA duplexes were designed per mRNA to increase the probability of successful knock-down. It was confirmed experimentally that individual siRNAs had similar effects on mRNA levels as the siRNA mixes, and it was also controlled that the siRNA and plasmids were delivered to the cells with above 90% efficiency by using a fluorescently labelled siRNA and a GFP expression plasmid (data not shown). Three negative (non-targeting) siRNAs were designed as controls.

Recombination Assays

For HR and NHEJ recombination transient assays, adherent CHO cells were transfected with HR or NHEJ reporter plasmids digested with I-SceI, and with the pGL3-CMV-dsRed plasmid to normalize for transfection efficiency, using Fugene 6 (Promega). The pGEGFP plasmid was transfected in parallel as a positive control of GFP expression.

For siRNA-mediated knock-downs of DNA repair proteins, adherent CHO DG44 cells were transfected with equimolar mixes of three mRNA-specific or control siRNA duplexes at a final concentration of 50 nM using Lipofectamine RNAiMAX (Invitrogen), according to manufacturer's instructions (Supplementary Fig. S2A). After 2 days, the siRNA-treated cells were re-transfected with pGEGFP or p1-68-GFP, and with a puromycin resistance plasmid pSVpuro (Clontech), using Lipofectamine 2000 (Invitrogen). Prior to transfection all plasmids were linearized with PvuI and purified by ethanol precipitation. Puromycin (5 μg/mL) was added to the culture medium 24 h after transfection, and stably transfected cells were selected for 2 weeks. Stable GFP expression was analyzed by flow cytometry (CyAn flow cytometer, Beckman Coulter), whereas aliquots of each sample were used for genomic DNA extraction.

Colony Formation Assay

To assess the frequency of genomic integration events, CHO DG44 cells were transfected with siRNA duplexes against selected DNA repair proteins, using the protocol described above (Supplementary Fig. S2A). Cells were re-transfected 72 h later with pGEGFP or p1-68-GFP, and pSVpuro, using Lipofectamine 2000 (Invitrogen), once the effects of the knock-down on cell cycle progression had disappeared (Kostyrko et al., 2015). The cells were trypsinized and counted 24 h after the second transfection, and 10000 viable cells were seeded in complete medium into each well of a 6-well plate. Puromycin (5 μg/mL) was added to the medium 7 h after seeding. After 10 days of selection, puromycin-resistant colonies were stained with 0.2% methylene blue and quantified using ImageJ (U.S. National Institutes of Health, Bethesda, MD).

Transgene Copy Number Determination and Quantitative PCR

To analyze the transgene copy number, total genomic DNA was isolated from cells using the DNeasy purification kit (Qiagen). For quantitative PCR (qPCR), 6 ng of genomic DNA were analyzed using the SYBR Green I Master kit for the Light Cycler 480 machine (Roche) using AGCAAAGACCCCAACGAGAA and GGCGGCGGTCACGAA as GFP-specific primers. The beta-2-microglobulin (B2M) CHO gene was amplified as a normalization control using ACCACTCTGAAG-GAGCCCA and GGAAGCTCTATCTGTGTCAA as primers. The number of integrated transgene (GFP) copies was calculated using the B2M gene as a reference, as previously described (Pfaffl, 2001).

Characterization of Transgene Integration Sites

To assess which CHO genes were expressed in our culture conditions, the transcriptome of the suspension-adapted parental CHO K1 cells was determined by paired-end sequencing using the Illumina technology by the Next Generation Sequencing Facility of the University of Lausanne. Expressed coding sequences were annotated using the Annotation Release 101 of the Chinese hamster genome assembly (CriGri_1.0, GCF_000223135.1) (Xu et al., 2011).

To identify the plasmid integration sites in polyclonal populations, CHO K1 cells were electroporated with the MAR-devoid pGEGFP or the MAR-containing p1-68-GFP plasmids and with the pSVpuro puromycin resistance construct using the Neon® transfection system (Invitrogen). After 3 weeks of puromycin selection, total genomic DNA was isolated from polyclonal cells using the Genomic-tip G/20 kit (Qiagen). The DNA was sequenced using the Single Molecule Real-Time (SMRT) technology (Pacific Biosciences) at the Next Generation Sequencing Facility of the University of Lausanne. CHO cells transfected with p1-68-GFP were sequenced using 20 SMRT cells, and those transfected with pGEGFP required the use of 60 SMRT cells to obtain a similar number of integration site sequences. Transgene integration sites were identified by a custom identification pipeline. PacBio filtered subreads were obtained using the tool DEXTRACTOR (Myers, unpublished) using the standard settings. Plasmid sequences were identified in PacBio filtered subreads with the help of the alignment tool BLASR (Chaisson and Tesler, 2012). A raw score of at least

−500 was chosen as cut-off based on results using PacBio reads from untransfected CHO cells. Flanking regions of matching plasmid sequences were extracted and mapped onto the CHO K1 genome using BLASR. 14 CHO genomic integration sites were identified in the p1-68-GFP-transfected population and 10 in the pGEGFP-transfected population. Two sets, one of 14 and one of 10, different, randomly picked genomic scaffolds of the same length (±10%) as the sample scaffolds were selected as controls. The Annotation Release 101 of the Chinese hamster genome assembly (CriGri_1.0, GCF_000223135.1) was used to identify the CHO genes in the vicinity of the integration sites. The presence of genes near the plasmid integration position in each of the identified scaffolds was compared with an analogous position on a corresponding control scaffold. An exact binomial test was used to calculate statistical significance between these datasets. Based on this analysis, integration within 5 kb from an open reading frame (ORF) was considered as intragenic, whereas integration within 35 kb from an ORF was defined as gene-proximal.

Suspension-adapted CHO K1 cells were stably transfected in multiple transfection cycles with plasmid vectors containing the human MAR X-29 and encoding the light and heavy chains of the trastuzumab and adalinumab therapeutic antibodies, as previously described (Le Fourn et al., 2014), with prior PvuI cleavage of the vectors. Clones expressing the highest amount of the recombinant proteins were selected for whole genome sequencing (Illumina), performed by Fasteris SA (Plan-Les-Ouates, Switzerland). Integration sites were first predicted by the in silico identification of paired reads displaying linked plasmid and genomic sequences, and the predicted junctions were subsequently validated by PCR amplification and Sanger sequencing. Identification of CHO genes near the plasmid integration sites was performed as described for the polyclonal populations.

Analysis of Immunoglobulin-Expressing CHO Cells

To assess the impact of DNA repair protein knock-down on recombinant protein expression, CHO K1 cells were electroporated with a negative control siRNA and siRNAs against MDC1, Ligase I, Rad51, and Rad52 using the Neon® transfection system (Invitrogen) (Supplementary Fig. S2B). Two days post transfection the cells were electroporated with PvuI-linearized human immunoglobulin (IgG1) expression vectors containing the MAR 1–68 and a puromycin resistance plasmid (pSVpuro) (Supplementary Fig. S1), using the Neon® transfection system (Invitrogen). After 3 weeks of antibiotic selection the IgG titer in cell culture supernatants was measured by sandwich ELISA and the specific productivity was calculated as described previously (Le Fourn et al., 2014).

Results

Plasmid Integration Does Not Rely on NHEJ or the Canonical HR Pathway

To assess the possible implication of NHEJ and HR in plasmid concatemer formation and spontaneous integration into the cell genome, we silenced the components of these major DSB repair pathways in CHO DG44 cells using short interfering RNA (siRNA)

(Supplementary Fig. S3 and Table SI). Efficient reduction of the target mRNA and/or protein levels by siRNA transfection was validated experimentally, to insure decreased levels by at least twofold (Supplementary Figs. S4 and S5).

To evaluate if the knock-down of these genes affects DNA recombination, we used previously described HR and NHEJ fluorescent reporter assays based on the repair of transiently transfected plasmids with a I-SceI-induced DSB in the GFP coding sequence (Mao et al., 2008; Seluanov et al., 2004). These assays enable to evaluate the efficiency of extrachromosomal break repair, and thereby may provide an estimation of HR and NHEJ involvement in plasmid concatemer formation. We observed that DSB repair of the HR reporter plasmid was impaired by the knock-down of the Rad51 HR protein, whereas it was rather increased in cells treated with siRNAs targeting NHEJ factors (Supplementary Fig. S6A). This indicated that Rad51 may contribute to the repair of DSBs in episomal plasmids. Interestingly, the knock-down of the remaining HR factors had no detectable effect on GFP reconstitution in this assay, although there was a very significant difference between the overall effect of knocking-down NHEJ and HR genes, in line with the previously reported competition between these pathways (Neal et al., 2011). In contrast, the occurrence of GFP expression from the NHEJ reporter was not altered by any of the NHEJ-targeting siRNAs (Supplementary Fig. S6B), implying that NHEJ is not prominently used to rejoin episomal DSBs in CHO cells or that alternative end-joining pathways may be more active than NHEJ.

We further assessed the recombination mechanisms involved in plasmid concatemer formation and genomic integration by stably transfecting the siRNA-treated CHO cells with plasmids carrying the GFP reporter and a puromycin resistance gene (Supplementary Fig. S2A). The average number of integrated GFP copies was measured in antibiotic-resistant polyclonal populations, so as to assess the efficiency of plasmid concatemerization prior to genomic integration. Since expression from individual plasmids can be influenced by the surrounding chromatin environment, the level of GFP fluorescence and its normalization to the transgene copy number was used to estimate plasmid integration within transcription permissive or non-permissive areas of the genome. Finally, we measured the efficiency of plasmid genomic integration by quantifying puromycin-resistant colonies arising from cells that had successfully integrated transgenes into their genome, focusing on siRNAs that affected GFP expression or plasmid concatemerization, as well as representative targets from each DSB repair pathway.

The average GFP expression and plasmid copy number were not affected by the down regulation of NHEJ activities such as DNA-PKcs, Ligase IV or Xrcc4, nor was the expression per transgene copy or the number of antibiotic resistant colonies (Fig. 1A–D). This indicated that NHEJ activities are not limiting for plasmid concatemerization and integration within the cell genome.

Stable GFP expression and/or transgene copy numbers were increased by the knock-down of HR proteins, notably MDC1, Rad51, Rad52, Rad54, and Brca1 (Fig. 1A and B). The knockdown of these proteins had overall little effect on gene expression when normalized to the copy number, indicating that the increased expression observed upon HR gene knockdown resulted mostly from an increased copy number rather than from preferential plasmid integration into transcription-permissive chromatin (Fig. 1C). These observations indicated that HR activities may oppose a mechanism that mediates plasmid concatemerization prior to genomic integration. However, the knock-down of proteins having an effect on plasmid concatemerization and/or GFP expression, such as MDC1 and Rad51, strongly decreased the number of puromycin-resistant colonies (Fig. 1D), indicating that these components of the HR pathway may mediate transgene genomic integration. Interestingly, the frequency of integration was not affected by the knock-down of other components of HR, such as Rad52, Rad54, or Brca1, despite their effect on transgene concatemerization and expression. These findings implied that some HR activities are required for genomic integration whereas others are not, suggesting the occurrence of non-canonical HR-related integration mechanisms.

MMEJ-Type Mechanisms Mediate Plasmid Concatemerization and Genomic Integration

Given that neither the NHEJ nor the canonical HR pathway may be involved in plasmid concatemerization prior to genomic integration, we speculated that this could involve MMEJ-related mechanisms active in eukaryotic cells with impaired NHEJ and/or HR, but that may share early 5′ strand resection events with the HR pathway (Supplementary Fig. S3 and Table SII) (Decottignies, 2007; Ma et al., 2003).

Knock-down of most MMEJ proteins had a moderate effect on plasmid integration or expression, possibly because these pathways may be masked by other repair mechanisms in the absence of induced DNA damage, as was the case here (Fig. 2A–C). Nevertheless, we observed a small decrease in GFP copy number upon the knock-down of DNA polymerase θ (Pol theta), suggesting that this polymerase might be involved in plasmid concatemerization, although the potential involvement of other DNA polymerases cannot be excluded. Interestingly, the knock-down of Ligase I had an opposite effect (Fig. 2A and B). Moreover, the depletion of this ligase strongly inhibited plasmid genomic integration (Fig. 2D). A recent study suggested the existence of two branches of the MMEJ-related end-joining pathways, one of which may depend on Ligase I whereas the other would require Ligase III (Oh et al., 2014; Paul et al., 2013). We thus speculated that upon Ligase I knock-down, the Ligase III-dependent branch could prevail, which may favor plasmid concatemer formation. In contrast, the pathway responsible for plasmid genomic integration may be dependent on Ligase I, as it is suppressed by it's depletion.

We have recently constructed a MMEJ-specific GFP reporter assay, based on principles analogous to the HR and NHEJ reporter plasmids used above (Kostyrko and Mermod, 2015). Interestingly, the use of this reporter in CHO cells revealed that the majority of episomal DSBs were not re-joined by a simple MMEJ pathway. Instead, the joined sequences of most repaired vectors rescued from the transfected cells resembled the recently proposed alternative DNA synthesis dependent (SD)-MMEJ mechanism (Yu and McVey, 2010). This pathway relies on a non-processive DNA polymerase, such DNA polymerase θ, to copy short homologous sequences (2–9 bp) from a different part of the repaired molecule, which can
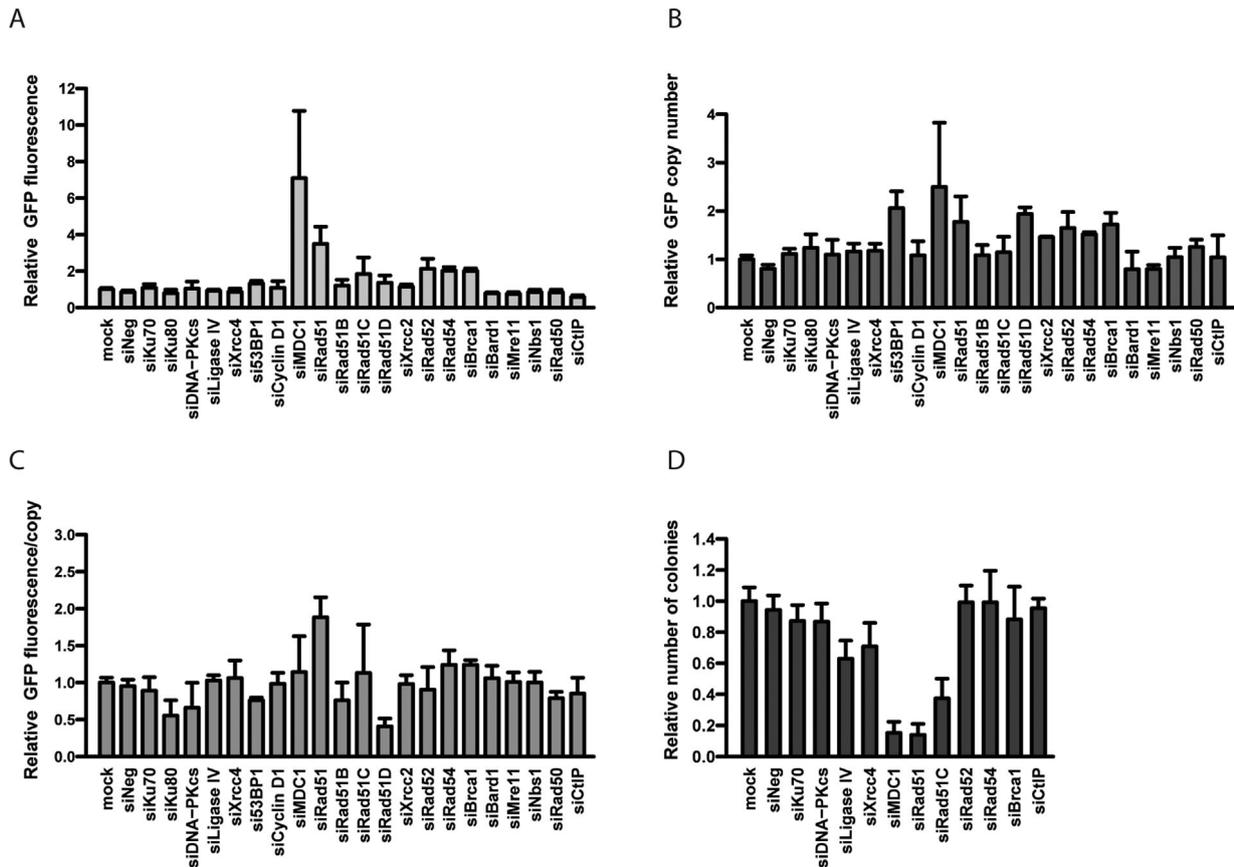
Figure 1. Effect of HR and NHEJ components knock-down on plasmid genomic integration and expression. CHO cells treated with indicated siRNAs were re-transfected with a GFP expression plasmid and puromycin resistance vector. Puromycin-resistant stable polyclonal CHO populations were assessed for average GFP fluorescence (**A**), GFP copy number (**B**), GFP expression per transgene copy (**C**), and the occurrence of puromycin-resistant colonies (**D**). Values represent mean fold change over control cells not treated with siRNAs (mock); s.e.m error bars, $n \geq 3$.

then be used to rejoin the DSB (Yousefzadeh et al., 2014; Yu and McVey, 2010). As a result, the junction sequence consists of a short duplication (direct or inverted) of a sequence found nearby on the repaired DNA fragment (Supplementary Fig. S3). Seventy percent of the analyzed repair products had no pre-existing microhomology indicative of MMEJ, but they displayed direct or inverted repeat sequences associated SD-MMEJ, up- or downstream of the repaired junction (Kostyrko and Mermod, 2015). We thus concluded that plasmid-to-plasmid joining relies mostly on a SD-MMEJ pathway potentially involving DNA polymerase θ and Ligase III, and that the simple MMEJ mechanism is seldom used.

MAR Elements Promote Plasmid Integration by Stimulating SD-MMEJ Pathways

We previously showed that transgene integration in CHO cells is enhanced three- to fourfold in the presence of matrix attachment regions (MARs), which are DNA elements that form chromatin domain boundaries (Girod et al., 2007; Majocchi et al., 2014). A human MAR, termed MAR 1–68, was found to increase both the number of transgene copies as well as the frequency of genomic integration events in CHO cells, which has been previously ascribed

to HR-related mechanisms (Grandjean et al., 2011). However, which HR-related recombination mechanism may be activated by MAR elements was not assessed.

To unambiguously identify the recombination mechanism activated by such elements, we combined the addition of the human MAR 1–68 in the GFP vector with the siRNA knock-down approach used earlier. As shown previously, inclusion of the MAR 1–68 enhanced GFP expression and copy number by approximately five- and threefold, respectively, when compared to the MAR-devoid control (Fig. 3A and B). This indicated that the MAR acted in part to activate plasmid concatemerization, whereas it concomitantly increased expression per gene copy (Fig. 3C). The presence of the MAR also increased by around twofold the proportion of cells having recombined the transgenes into their genome (Fig. 3D), indicating that it also activated genomic integration.

In the presence of the MAR, the silencing of NHEJ factors had no effect on transgene expression or copy number, as before (Fig. 3A and B). In contrast, the knock-down of many HR and cell cycle control factors yielded very high transgene expression, but without further increasing the transgene copy number. Consistently, we observed an enhancement of expression per gene copy upon the knock-down of most HR factors, which was markedly
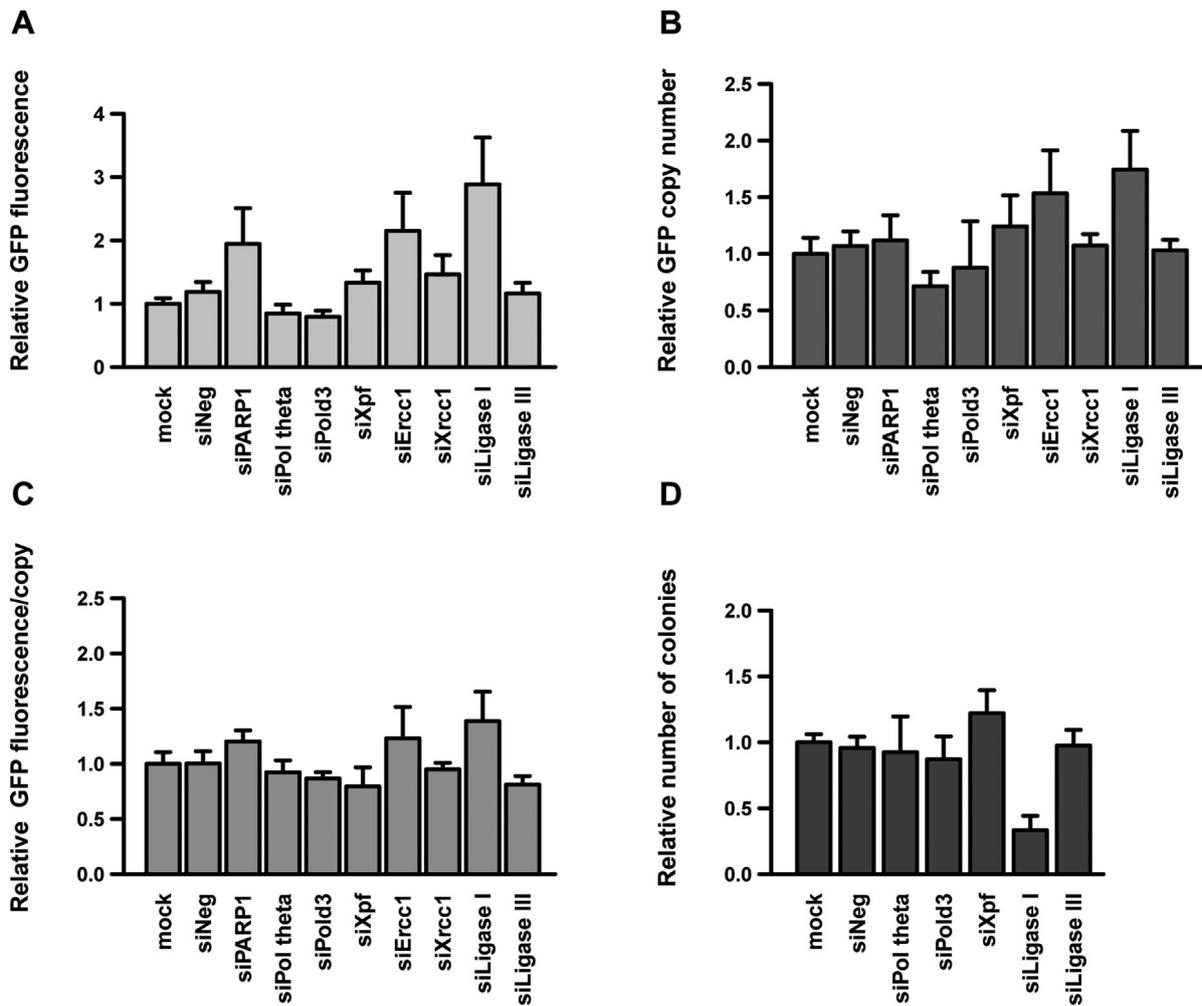
Figure 2. Effect of MMEJ components knock-down on plasmid genomic integration and expression. CHO cells were treated with siRNAs against the indicated MMEJ genes and processed as described in the legend to Figure 1. The average GFP fluorescence (**A**), GFP copy number (**B**), GFP expression per transgene copy (**C**), and frequency of genomic integration events (**D**) were assessed and represented as in Figure 1 ($n \geq 3$).

higher than the increase already mediated by the MAR (Fig. 3C). A strong inhibition of the frequency of plasmid genomic integration was again noted upon the knock-down of MDC1 and especially Rad51 (Fig. 3D). This indicated that these factors and the MAR may act synergistically to promote transgene genomic integration. However, upon the knock down of Rad51 and other HR proteins, the MAR-containing plasmids may have integrated preferentially into expression-permissive portions of the genome. We therefore speculated that the MAR acts to promote one or several MMEJ-related pathways that may direct transgenes into expression-favoring chromatin structures.

In the presence of the MAR, the knock-down of MMEJ factors had mostly similar effects on GFP expression and copy number as observed earlier for the MAR-devoid plasmid, with a small decrease upon the knock-down of DNA polymerase θ, and an increase in the absence of Ligase I (Fig. 3A and B). Interestingly, the presence of the MAR seemed to counteract the effect of Ligase I down-regulation on transgene genomic integration, possibly by reducing the inhibitory

effect of the reduced ligase level, or by stimulating a distinct recombination mechanism (Figs. 2C and 3D). Overall, we concluded that the MAR may activate both concatemerization and genomic integration processes by stimulating SD-MMEJ-related repair pathways, and that these pathways may concur with the MAR to favor integration into expression-permissive genomic loci.

The MAR and SD-MMEJ Pathways Mediate Transgene Integration Near Cellular Genes

To further assess which of the alternative recombination pathways may mediate favorable genomic integration events, we analyzed the genomic integration loci and the DNA sequence of the genome-plasmid junctions. This was performed on three CHO clones transfected multiple times with immunoglobulin (IgG) expression vectors containing the human MAR X-29 and selected for high stable expression of the therapeutic protein. To do so, we used a whole genome sequencing approach on these clones and
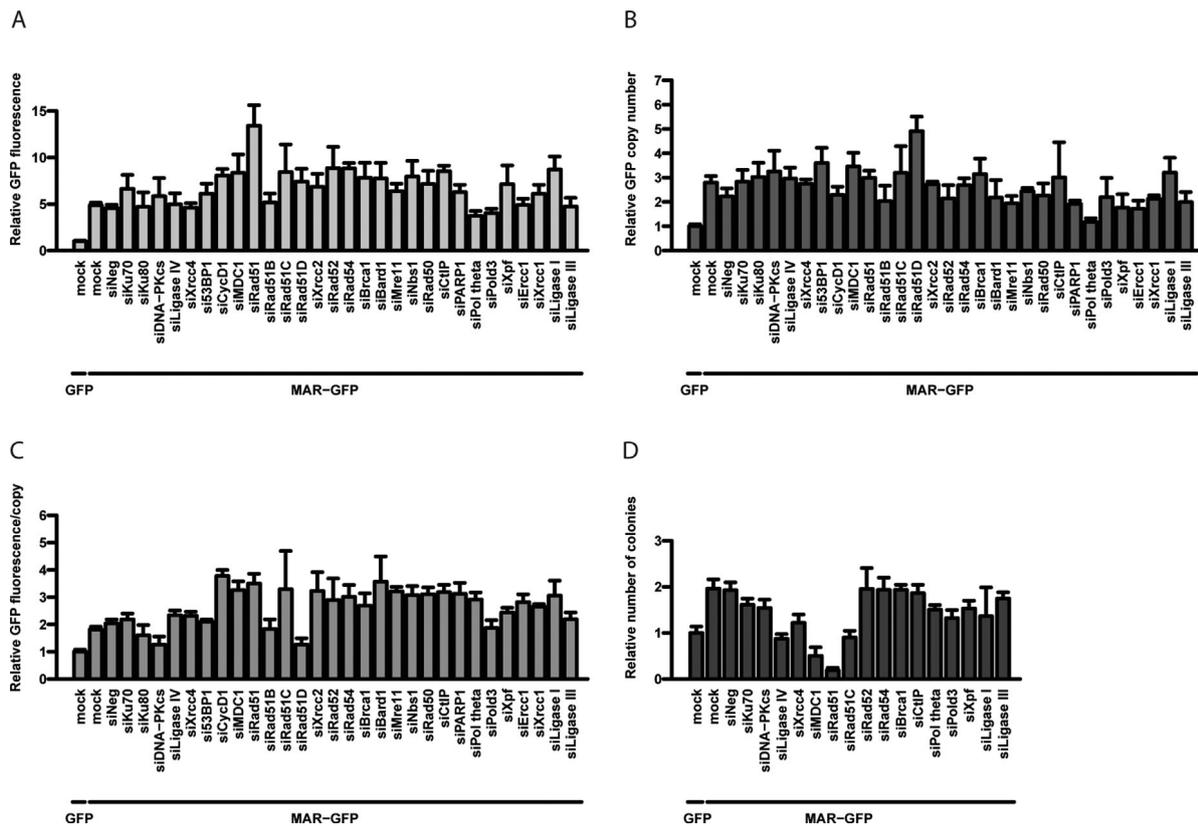
Figure 3. Effect of a MAR element and recombination gene knock-down on plasmid genomic integration and expression. The effect of the inclusion of a MAR element on stable GFP expression (**A**), GFP copy number (**B**), GFP expression per transgene copy (**C**), and the frequency of genomic integration events (**D**), were assessed as described for Figures 1 and 2, except that siRNA-treated cells were re-transfected with GFP or MAR-GFP vectors, as indicated ($n \geq 3$).

devised a software to identify paired sequence reads pertaining to the plasmid and the CHO genome. Six integration sites in one clone (BS01) and two in the other clones (BS03 and Cp33/64) were predicted in silico and validated experimentally by PCR amplification and DNA sequencing. The occurrence of the predicted number of plasmid integration loci was further validated by FISH for two of the analyzed clones (Supplementary Fig. S7).

From the five integration sites where the junction sequences were validated experimentally on both sides of the transgenes, two had large deletions (913 bp in BS01 and 320 bp in Cp33/64), as expected from MMEJ-related mechanisms (Supplementary Table SIV and Fig. 4). In 5 of the 15 experimentally validated junctions, we noted the presence of short (1–3 bp) or long (60–100 bp) templated inserts, suggesting the involvement of a DNA polymerase in the repair process, a hallmark of the SD-MMEJ mechanism (Fig. 5). All analyzed junction sequences fitted well to the SD-MMEJ model, although 5 out of 15 junctions also covered pre-existing micro-homologies (≥ 2 nt), and thus could also be explained by simple MMEJ. Interestingly, no integration site could be explained by HR. Although NHEJ cannot be fully excluded, as it does not strictly require extensive homology, the SD-MMEJ mechanism more readily explains the presence of extended deletions and templated inserts. Moreover, no junction lacking any type of microhomology was

observed. Overall, these results confirmed that the genomic integration of MAR-containing plasmids predominantly involves a SD-MMEJ pathway.

Out of the 10 integration events, eight had occurred within or near cellular genes, whereas only two were intergenic. Seven out of these eight gene-proximal integrations were found in or close to an expressed gene (Supplementary Table SV), suggesting that most integration events had occurred in transcriptionally active genomic loci. These results further suggested that the MAR-containing plasmids preferably integrate within- or in close proximity- to expressed CHO genes. To assess whether this indeed resulted from the presence of the MAR in the IgG expression vector or from the selection of highly expressing clones, we directly compared the integration loci of MAR-containing and MAR-devoid GFP expression vectors in polyclonal cell populations.

Analysis of the integration sites identified from the whole genome sequencing of these cells revealed that, in presence of the MAR, plasmids indeed often integrated close to cellular genes (10/14 loci) (Supplementary Table SV and Fig. S8). This result was significantly different from random ($P = 0.05$), indicating that the MAR may stimulate genomic integration into chromatin regions permissive for transgene expression. In the cells transfected with the MAR-devoid plasmid, integration in the vicinity of genes was not significantly enriched (Supplementary Table SV and Fig. S8B).
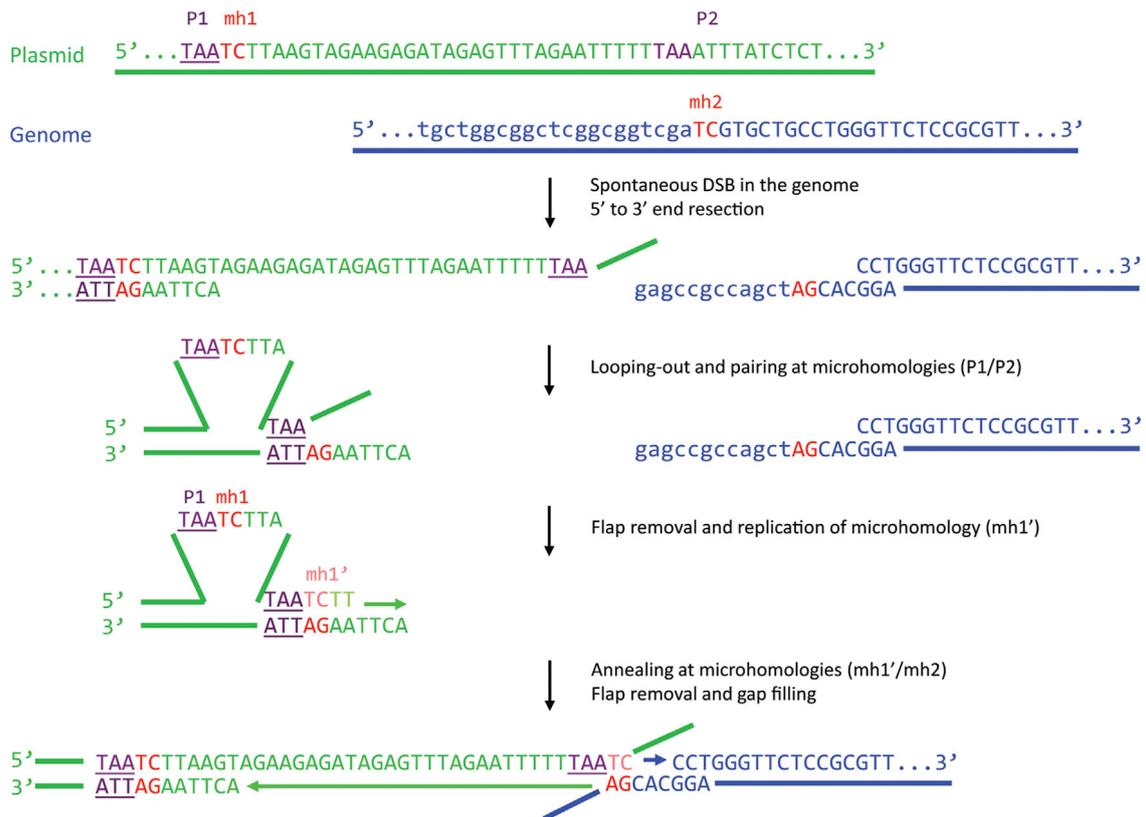
Figure 4. Example of a plasmid-to-genome junction and underlying SD-MMEJ mechanism. The integration site and junction sequence used in this example is taken from Supplementary Table SIV (clone BS01, integration site #2, right junction). P1/P2, primer repeats; mh1/mh2, microhomology repeats. Adapted from Yu and McVey (2010).

Furthermore, these cells required threefold more sequencing reads to identify a comparable number of integration loci as obtained from the cells transfected with the MAR, further indicating that genomic integration events were less frequent in the absence of the MAR.

Interestingly, all cellular genes near the integration loci of MAR-devoid plasmids were transcribed in the parental CHO cells (Supplementary Table SV and Fig. S9). This suggested that, in the absence of the MAR, the cells had to integrate the transgenes into transcriptionally active chromatin in order to express the selection gene at a sufficient level to survive antibiotic selection. This may explain the strong decrease in cell survival upon the knock-down of Rad51, as this protein was recently reported to be primarily responsible for DSB repair in transcriptionally active chromatin (Aymard et al., 2014). In contrast, presence of the MAR seemed to alleviate the need to integrate transgenes into transcribed genomic sequences, as only half of the CHO genes close to integration sites were found to be transcriptionally active. This indicated that the MAR itself may ensure high expression of transgenes integrated in non-transcribed DNA, likely due to its previously reported transcription-enhancing properties (Galbete et al., 2009; Majocchi et al., 2014). Taken together, these results suggested that MAR elements may promote transgene integration into gene-rich chromatin regions by stimulating an SD-MMEJ mechanism.

MARs and HR or SD-MMEJ Knock-Down Improve Recombinant Protein Expression

The transient knock-down of MDC1, Rad51, Rad52, and Ligase I was found to mediate the highest and most homogeneous GFP fluorescence from polyclonal pools of cells stably transfected with the MAR-GFP vector (Fig. 6A). To ascertain whether the knock-down of these specific HR and/or SD-MMEJ activities may be used in conjunction with MAR elements as a general approach to boost the expression of recombinant proteins, we similarly assessed vectors encoding a therapeutic IgG1 immunoglobulin, using a suspension-adapted CHO K1 cell line derivative suitable for the production of therapeutics. CHO cells treated with siRNAs against Rad51, Rad52, MDC1, or Ligase I were subsequently re-transfected with MAR-containing vectors for the human IgG1 light and heavy chains (Supplementary Fig. S2B). Polyclonal populations were then assessed for specific antibody secretion, which revealed that prior treatment with Rad52, MDC1, or Ligase I siRNAs increased stable IgG expression by approximately twofold relative to the untreated cells (Fig. 6B). The high productivity levels observed from these polyclonal populations, up to over six picograms per cell per day (PCD), are usually only observed from monoclonal populations obtained from the screening of hundreds of individual cell clones, to identify the most productive ones. Interestingly, Rad51 depletion in
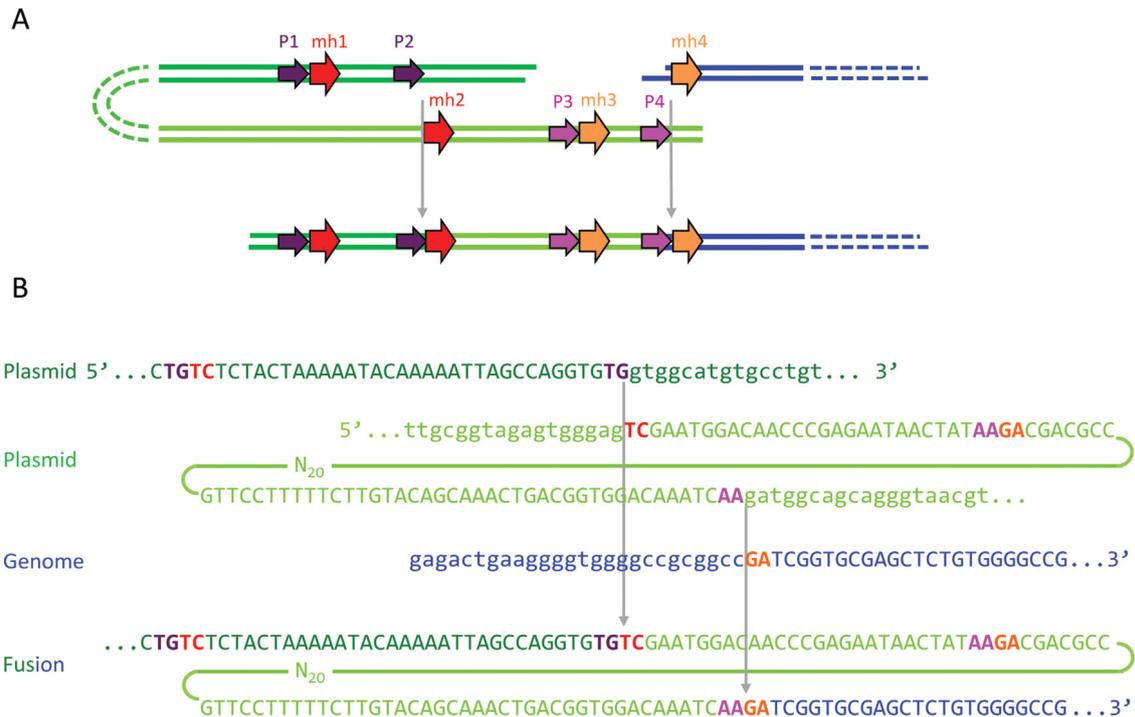
Figure 5. Example of a plasmid-to-genome junction and SD-MMEJ mechanism requiring a templated insertion. (**A**) A scheme showing the mechanism of plasmid (dark green) joining with the genome (blue). Another fragment of the plasmid (light green) serves as an adaptor providing the microhomologies required for joining and becomes incorporated into the junction as a templated insert. (**B**) Sequences of plasmid and genome fragments shown in panel A. The integration site and junction sequence used in this example is taken from Supplementary Table SIV (clone BS01, integration site #1, right junction). P1/P2 and P3/P4, primer repeats; mh1/mh2 and mh3/mh4, microhomology repeats.

CHO K1 cells had a weaker effect on transgene expression than in CHO DG44 cells. This could be due to the combination of mechanical stress associated with growth as cell suspension in shake flasks, antibiotic selection and the deleterious effect of Rad51 knock-down. Consistently, we observed that CHO K1 cells treated with Rad51 siRNA grew much slower than the cells treated with other siRNAs, and only a small number of cells survived selection (data not shown). We hypothesize that the population of CHO K1 cells that recovered from selection represented cells which retained some Rad51 activity, and which thus did not have a large increase of plasmid concatemerization and overall expression.

In conclusion, the increase in expression mediated by Rad52, MDC1, and Ligase I knock-down could be observed for distinct recombinant proteins, and from the use of distinct CHO cell lines and vectors. We concluded that the production of therapeutic proteins in CHO cells may be significantly improved by transiently altering their DSB repair properties during transfection and by incorporating MAR elements in the vector.

Discussion

Eukaryotic cells have developed many defense mechanisms that detect and repair DNA double stranded breaks, one of the most deleterious types of DNA damage. The two canonical pathways responsible for DSB repair are HR and NHEJ. However, recent evidence indicated that these two mechanisms may not suffice to repair all DSBs, and that several alternative pathways, collectively termed MMEJ or alt-NHEJ, also exist in eukaryotic cells (Gigi et al., 2014; Truong et al., 2013). These later processes are often obscured by the main repair mechanisms, which may predominate in normal cells. Furthermore, their components are still poorly characterized and there was no simple assay to specifically detect them, rendering their study difficult (Kostyrko and Mermod, 2015). However, they are now attracting increasing attention, notably in oncology, since these "illegitimate" recombination pathways were shown to be more prevalent in tumor cells and to cause chromosomal rearrangements leading to cancer (Bentley et al., 2004; Simsek et al., 2011; Tobin et al., 2012; Zhang and Jasin, 2011).

Here, we found that NHEJ and HR are not the main pathways responsible for non-specific recombination in CHO cells, as required for plasmid genomic integration in these cells. Rather, we found that the absence of several HR factors augmented plasmid concatemerization, implying that HR proteins may compete with one or more DSB repair pathways that mediate this process. In contrast, specific HR proteins, such as Rad51, were required for efficient transgene recombination with the genome, whereas the silencing of downstream HR proteins had no effect. This suggested the involvement of other mechanisms, distinct from the canonical NHEJ or HR pathways but nevertheless requiring DNA homology, such as MMEJ-related pathways. Consistently, the knock-down of Ligase I, a protein reported to play a role in alternative DSB repair pathways, was found to alter plasmid genomic integration.
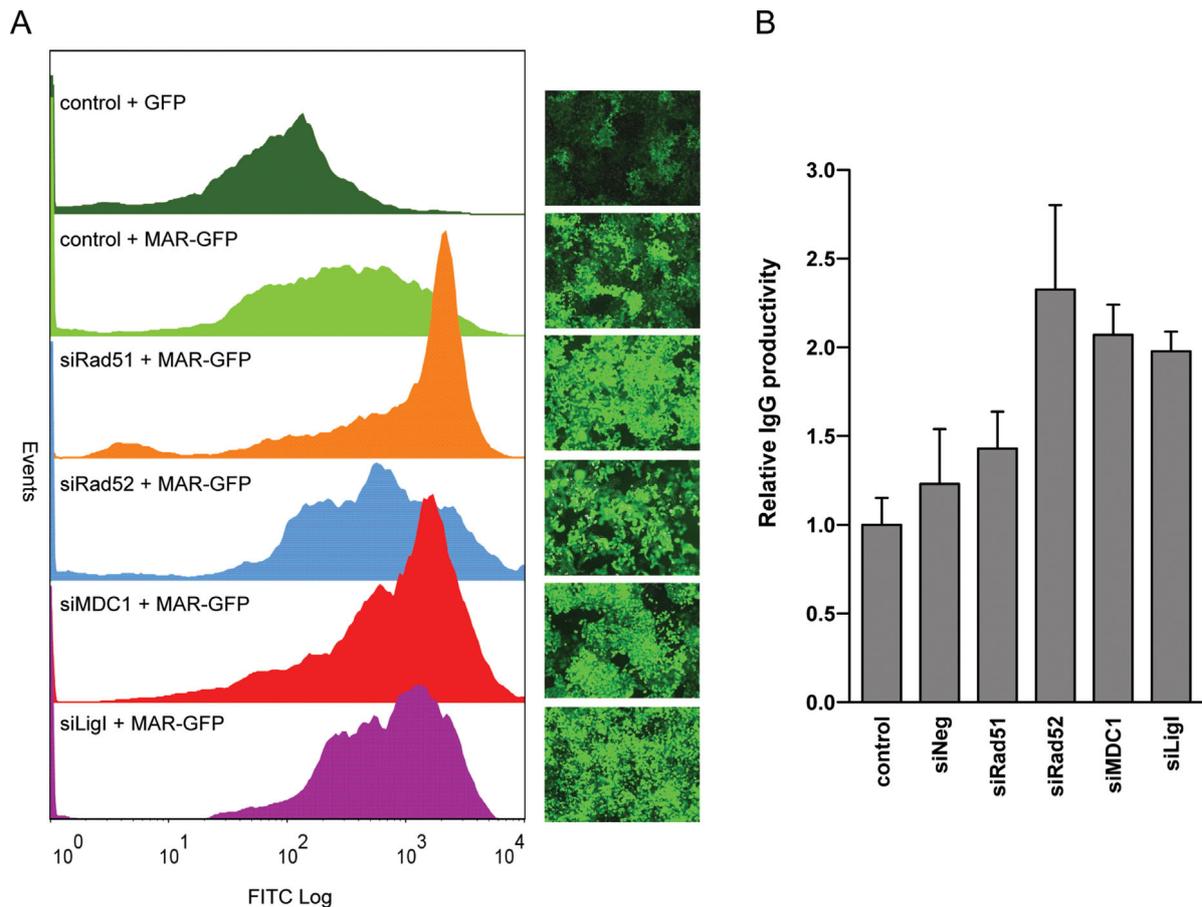
Figure 6. Engineering of the transgene integration process for improved expression. (**A**) Adherent CHO cells transfected with the indicated siRNAs or left untreated (control), were re-transfected with a GFP or MAR-GFP vector, as indicated, and selected for antibiotic resistance. GFP fluorescence profiles of polyclonal cell pools and corresponding fluorescence microscopy pictures are shown. (**B**) Specific IgG productivity in polyclonal, suspension-adapted CHO cells treated as for panel A, except that they were re-transfected with the MAR-containing IgG1 expression vectors. Values represent the average fold change in IgG secretion (in picograms/cell/day) as compared to the cells not treated with siRNAs (control); s.e.m error bars, $n=3$.

As the majority of rejoined plasmid extremities displayed microhomology patterns and templated inserts, we attribute these end-joining events to the SD-MMEJ mechanism proposed by Yu and McVey (2010). Indeed, both plasmid-to-plasmid and plasmid-to-genome fusion sequences were also present as direct or inverted repeats near the junctions, occasionally accompanied by templated inserts. However, the knock down of specific SD-MMEJ activities had distinct effects on plasmid concatemer formation and on genomic integration, suggesting the occurrence of multiple SD-MMEJ pathways. One of these pathways, which may rely on DNA polymerase θ and Ligase III, appears to mediate plasmid concatemerization. The other SD-MMEJ pathway, which may involve the activity of Ligase I, appears to mediate the recombination of plasmid concatemer with the genome, as indicated by the finding that the lack of this ligase nearly abolished genomic integration of the GFP vector. Taken together, these results imply that concatemer formation and integration of MAR-devoid plasmids may be mediated by sets of proteins belonging to distinct branches of the SD-MMEJ pathways, as proposed in Figure 7.

Interestingly, we observed that the knock down of Rad51 had similar effects as the silencing of Ligase I, implying that they contribute to the same pathway mediating genomic integration of exogenous DNA. The mechanism mediating microhomology search of the SD-MMEJ pathway remains mostly uncharacterized, but it may involve DSB repair components that are common to other mechanisms. In this model, the Ligase I-dependent SD-MMEJ pathway may lie downstream of the search for a homologous DNA strand by Rad51, as in canonical HR (Fig. 7). However, the lack of extended homology may preclude the productive cooperation of Rad51 with its accessory proteins, preventing extended strand invasion and the successful completion of HR. End-joining would then rather be performed by Ligase I-dependent SD-MMEJ, as a salvage repair pathway, since it only requires short homology regions as shared by the plasmid and cell genome. When such microhomologies are not available, they may be provided by an adaptor DNA stretch copied from nearby plasmid or genome sequences, leading to the insertion of a templated insert separating the joined sequences (Fig. 5). We hypothesize that the enzyme involved in the synthesis of the templated insert may be DNA
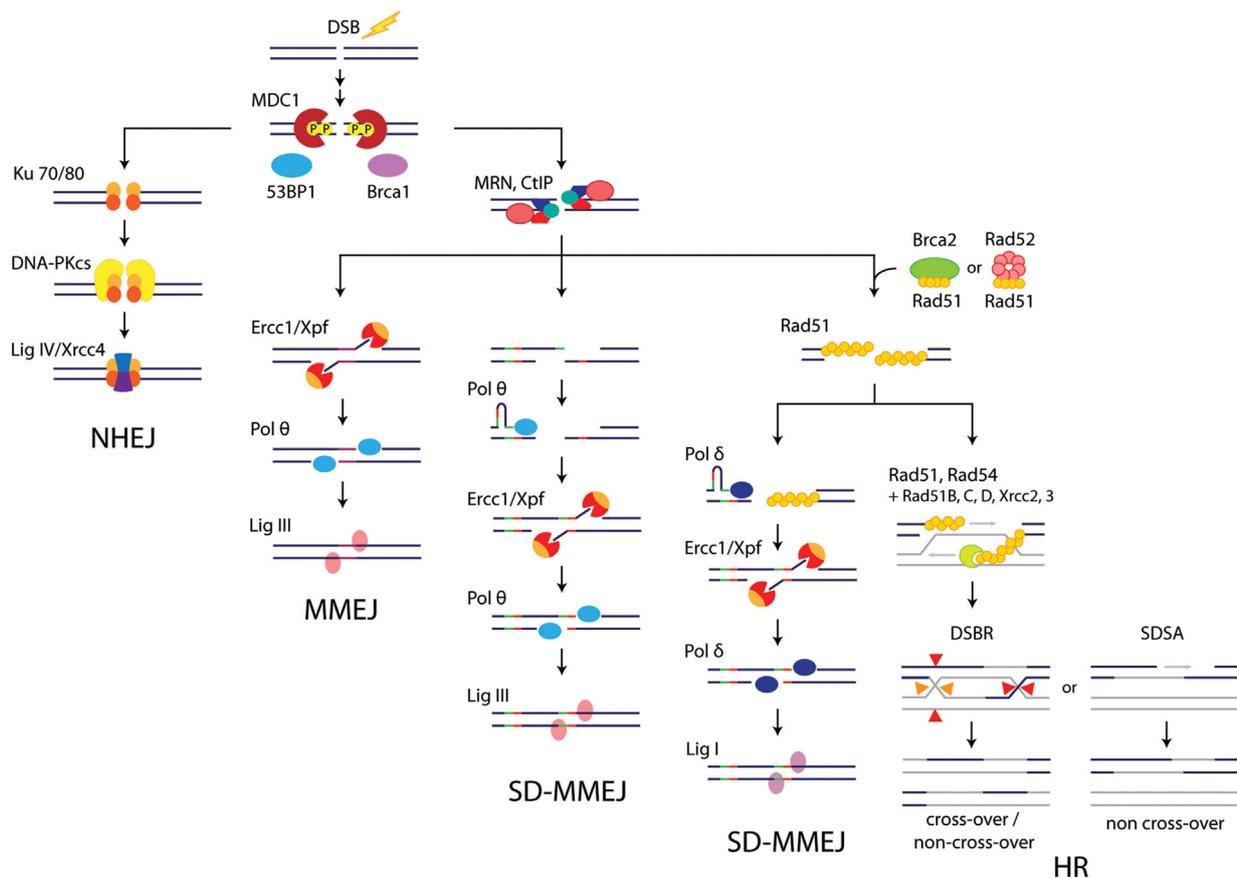
Figure 7. Revised model of the major CHO cell DSB repair pathways. A novel model describing the possible interplay of the NHEJ, HR, MMEJ, and two distinct SD-MMEJ pathways involved in DSB repair in CHO cells, as modified from Supplementary Figure S3. Although the junction sequences resulting from both SD-MMEJ pathways are similar, the Ligase I-dependent SD-MMEJ requires the homology-searching Rad51 protein, and it may provide a fallback mechanism in the absence of extensive homology, as required to complete HR. Activities that initiate MMEJ and Ligase III-dependent SD-MMEJ remain to be identified, but the presence or absence of pre-existing microhomologies may dictate the choice between these two pathways.

polymerase δ, which together with Ligase I participates in DNA replication and long patch base excision repair (BER) (Stucki et al., 1998). In human cells, a break-induced replication (BIR) mechanism responsible for the repair of one-ended DSBs was also recently shown to rely on POLD3, a DNA polymerase δ subunit, and to involve microhomologies (Costantino et al., 2014; Lydeard et al., 2007). Nevertheless, whether the Rad51-dependent SD-MMEJ pathway proposed here may be related to BIR remains to be established.

Inclusion of a MAR element also increased plasmid concatemerization, suggesting that it can act to activate the processing of linearized plasmid extremities by a Ligase I-independent SD-MMEJ mechanism. In addition, the MAR presence increased genomic integration and dampened the inhibitory effect of Ligase I downregulation, whereas it did not abolish the requirement for Rad51. These findings suggest a preferential use of the Ligase I-dependent SD-MMEJ mechanism for the genomic integration of the MAR-devoid plasmid, whereas the presence of the MAR may stimulate the use of a repair pathway downstream of Rad51 that may involve a distinct ligase, for example, Ligase III.

The molecular mechanisms by which MARs may promote SD-MMEJ-mediated recombination could involve their AT-rich cores, which possess a high potential for double helix denaturation (Bode et al., 1992; Platts et al., 2006), or their enrichment in topoisomerase II cleavage sites and so-called fragile sites that may be the hot spots of DNA breakage and repair (Jackson et al., 2003; Sperry et al., 1989; Svetlova et al., 2001). Consistently, these sites were previously reported to be preferred targets of plasmid integration (Rassool et al., 1991). MAR elements were also proposed to mediate DNA replication initiation in mammalian somatic cells (Debatisse et al., 2004). Thus, they might associate with DNA replication machinery components also involved in MMEJ-related mechanisms (e.g., Pold3, Ligase I), thereby contributing to the repair of DSBs arising at replication forks (Truong et al., 2013).

In this study, we identified SD-MMEJ as the primary mechanism driving plasmid integration in the genome of CHO cells. We propose the occurrence of two distinct SD-MMEJ branches relying on different subsets of proteins, both of which are stimulated by MAR elements, to increase transgene copy number and to preferentially target plasmid DNA into potentially expression-permissive, gene-

rich regions of the genome. Finally, we use this knowledge to transiently modify the DNA recombination properties of CHO cells to improve the expression of a therapeutic antibody, demonstrating that this approach can be used to engineer cells for more efficient recombinant protein expression.

References

Allen GC, Spiker S, Thompson WF. 2000. Use of matrix attachment regions (MARs) to minimize transgene silencing. Plant Mol Biol 43:361–376.

Aymard F, Bugler B, Schmidt CK, Guillou E, Caron P, Briois S, Iacovoni JS, Daburon V, Miller KM, Jackson SP, Legube G. 2014. Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. Nat Struct Mol Biol 21:366–374.

Bentley J, Diggle CP, Harnden P, Knowles MA, Kiltie AE. 2004. DNA double strand break repair in human bladder cancer is error prone and involves microhomology-associated end-joining. Nucleic Acids Res 32:5249–5259.

Boboila C, Yan C, Wesemann DR, Jankovic M, Wang JH, Manis J, Nussenzweig A, Nussenzweig M, Alt FW. 2010. Alternative end-joining catalyzes class switch recombination in the absence of both Ku70 and DNA ligase 4. J Exp Med 207:417–427.

Bode J, Kohwi Y, Dickinson L, Joh T, Klehr D, Mielke C, Kohwi-Shigematsu T. 1992. Biological significance of unwinding capability of nuclear matrix-associating DNAs. Science 255:195–197.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Theory and application. BMC Bioinformatics 13:238.

Costantino L, Sotiriou S, Rantala J, Magin S, Mladenov E, Helleday T, Haber J, Iliakis G, Kallioniemi O, Halazonetis T. 2014. Break-induced replication repair of damaged forks induces genomic duplications in human cells. Science 343:88–91.

Dai J, Cui X, Zhu Z, Hu W. 2010. Non-homologous end joining plays a key role in transgene concatemer formation in transgenic zebrafish embryos. Int J Biol Sci 6:756–768.

Debatisse M, Toledo F, Anglana M. 2004. Replication initiation in mammalian cells: Changing preferences. Cell Cycle 3:19–21.

Decottignies A. 2007. Microhomology-mediated end joining in fission yeast is repressed by pku70 and relies on genes involved in homologous recombination. Genetics 176:1403–1415.

Dinkelmann M, Spehalski E, Stoneham T, Buis J, Wu Y, Sekiguchi JM, Ferguson DO. 2009. Multiple functions of MRN in end-joining pathways during isotype class switching. Nat Struct Mol Biol 16:808–813.

Folger KR, Wong EA, Wahl G, Capecchi MR. 1982. Patterns of integration of DNA microinjected into cultured mammalian cells: Evidence for homologous recombination between injected plasmid DNA molecules. Mol Cell Biol 2: 1372–1387.

Galbete JL, Buceta M, Mermod N. 2009. MAR elements regulate the probability of epigenetic switching between active and inactive gene expression. Mol Biosyst 5:143–150.

Gigi V, Lewis S, Shestova O, Mijušković M, Deriano L, Meng W, Luning Prak ET, Roth DB. 2014. RAG2 mutants alter DSB repair pathway choice in vivo and illuminate the nature of šalternative NHEJć. Nucleic Acids Res 42:6352–6364.

Girod P-A, Nguyen D-Q, Calabrese D, Puttini S, Grandjean M, Martinet D, Regamey A, Saugy D, Beckmann JS, Bucher P, Mermod N. 2007. Genome-wide prediction of matrix attachment regions that increase gene expression in mammalian cells. Nat Methods 4:747–753.

Grandjean M, Girod P-A, Calabrese D, Kostyrko K, Wicht M, Yerly F, Mazza C, Beckmann JS, Martinet D, Mermod N. 2011. High-level transgene expression by homologous recombination-mediated gene transfer. Nucleic Acids Res 39:e104.

Harraghy N, Gaussin A, Mermod N. 2008. Sustained transgene expression using MAR elements. Curr Gene Ther 8:353–366.

Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 5: e1000327.

Hicks WM, Kim M, Haber JE. 2010. Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. Science 329: 82–85.

Iiizumi S, Kurosawa A, So S, Ishii Y, Chikaraishi Y, Ishii A, Koyama H, Adachi N. 2008. Impact of non-homologous end-joining deficiency on random and targeted DNA integration: Implications for gene targeting. Nucleic Acids Res 36: 6333–6342.

Jackson JA, Trevino AV, Herzig MC, Herman TS, Woynarowski JM. 2003. Matrix attachment region (MAR) properties and abnormal expansion of AT island minisatellites in FRA16B fragile sites in leukemic CEM cells. Nucleic Acids Res 31:6354–6364.

Jackson SP. 2002. Sensing and repairing DNA double-strand breaks. Carcinogenesis 23:687–696.

Kohli A, Leech M, Vain P, Laurie DA, Christou P, Brill WJ. 1998. Transgene organization in rice engineered through direct DNA transfer supports a two-phase integration mechanism mediated by the establishment of integration hot spots. Proc Natl Acad Sci USA 95:7203–7208.

Kostyrko K, Mermod N. 2015. Assays for DNA double-strand break repair by microhomology-based end-joining repair mechanisms. Nucleic Acids Res 44(6):e56.

Kostyrko K, Bosshard S, Urban Z, Mermod N. 2015. A role for homologous recombination proteins in cell cycle regulation. Cell Cycle 14:2853–2861.

Le Fourn V, Girod P-A, Buceta M, Regamey A, Mermod N. 2014. CHO cell engineering to prevent polypeptide aggregation and improve therapeutic protein secretion. Metab Eng 21:91–102.

Lee JA, Carvalho CMB, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131:1235–1247.

Lundberg R, Mavinakere M, Campbell C. 2001. Deficient DNA end joining activity in extracts from fanconi anemia fibroblasts. J Biol Chem 276:9543–9549.

Lydeard JR, Jain S, Yamaguchi M, Haber JE. 2007. Break-induced replication and telomerase-independent telomere maintenance require Pol32. Nature 448:820–823.

Ma J-L, Kim EM, Haber JE, Lee SE. 2003. Yeast mre11 and rad1 proteins define a ku-Independent mechanism to repair double-Strand Breaks lacking overlapping end sequences. Mol Cell Biol 23:8820–8828.

Majocchi S, Aritonovska E, Mermod N. 2014. Epigenetic regulatory elements associate with specific histone modifications to prevent silencing of telomeric genes. Nucleic Acids Res 42:193–204.

Mao Z, Bozzella M, Seluanov A, Gorbunova V. 2008. Comparison of nonhomologous end joining and homologous recombination in human cells. DNA Repair (Amst) 7:1765–1771.

Merrihew RV, Marburger K, Pennington SL, Roth DB, Wilson JH. 1996. High-frequency illegitimate integration of transfected DNA at preintegrated target sites in a mammalian genome. Mol Cell Biol 15:10.

Neal JA, Dang V, Douglas P, Wold MS, Lees-Miller SP, Meek K. 2011. Inhibition of homologous recombination by DNA-dependent protein kinase requires kinase activity, is titratable, and is modulated by autophosphorylation. Mol Cell Biol 31:1719–1733.

Oh S, Harvey A, Zimbric J, Wang Y, Nguyen T, Jackson PJ, Hendrickson EA. 2014. DNA ligase III and DNA ligase IV carry out genetically distinct forms of end joining in human somatic cells. DNA Repair (Amst) 21:97–110.

Paul K, Wang M, Mladenov E, Bencsik-Theilen A, Bednar T, Wu W, Arakawa H, Iliakis G. 2013. DNA ligases I and III cooperate in alternative non-homologous end-joining in vertebrates. PLoS ONE 8:e59505.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29:e45.

Platts AE, Quayle AK, Krawetz SA. 2006. In-silico prediction and observations of nuclear matrix attachment. Cell Mol Biol Lett 11:191–213.

Rassool FV, McKeithan TW, Neilly ME, van Melle E, Espinosa R III, Le Beau MM. 1991. Preferential integration of marker DNA into the chromosomal fragile site at 3p14: An approach to cloning fragile sites. Proc Natl Acad Sci USA 88:6657–6661.

Seluanov A, Mittelman D, Pereira-Smith OM, Wilson JH, Gorbunova V. 2004. DNA end joining becomes less efficient and more error-prone during cellular senescence. Proc Natl Acad Sci USA 101:7624–7629.

Simsek D, Brunet E, Wong SY-W, Katyal S, Gao Y, McKinnon PJ, Lou J, Zhang L, Li J, Rebar EJ, Gregory PD, Holmes MC, Jasin M. 2011. DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation. PLoS Genet 7:e1002080.

Sperry AO, Blasquez VC, Garrard WT. 1989. Dysfunction of chromosomal loop attachment sites: Illegitimate recombination linked to matrix association regions and topoisomerase II. Proc Natl Acad Sci USA 86:5497–5501.

Stucki M, Pascucci B, Parlanti E, Fortini P, Wilson SH, Hübscher U, Dogliotti E. 1998. Mammalian base excision repair by DNA polymerases δ and ε. Oncogene 17:835–843.

Svetlova EY, Razin SV, Debatisse M. 2001. Mammalian recombination hot spot in a DNA loop anchorage region: A model for the study of common fragile sites. J Cell Biochem 81:170–178.

Tobin LA, Robert C, Nagaria P, Chumsri S, Twaddell W, Ioffe OB, Greco GE, Brodie AH, Tomkinson AE, Rassool FV. 2012. Targeting abnormal DNA repair in therapy-resistant breast cancers. Mol Cancer Res 10:96–107.

Truong LN, Li Y, Shi LZ, Hwang PY-H, He J, Wang H, Razavian N. 2013. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. PNAS 110:7720–7725.

Urlaub G, Chasin LA. 1980. Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. Proc Natl Acad Sci USA 77:4216–4220.

Villarreal DD, Lee K, Deem A, Shim EY, Malkova A, Lee SE. 2012. Microhomology directs diverse DNA Break repair pathways and chromosomal translocations. PLoS Genet 8:e1003026.

Würtele H, Little KCE, Chartrand P. 2003. Illegitimate DNA integration in mammalian cells. Gene Ther 10:1791–1799.

Wong EA, Capecchi MR. 1987. Homologous recombination between coinjected DNA sequences peaks in early to mid-S phase. Mol Cell Biol 7:2294–2295.

Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S, Andersen MR, Neff N, Passarelli B, Koh W, Fan HC, Wang J, Gui Y, Lee KH, Betenbaugh MJ, Quake SR, Famili I, Palsson BO, Wang J. 2011. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotechnol 29:735–741.

Yousefzadeh MJ, Wyatt DW, Takata K-I, Mu Y, Hensley SC, Tomida J, Bylund GO, Doublié S, Johansson E, Ramsden a D, McBride KM, Wood RD. 2014. Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. PLoS Genet 10:e1004654.

Yu AM, McVey M. 2010. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. Nucleic Acids Res 38:5706–5717.

Zahn-Zabal M, Kobr M, Girod P-A, Imhof M, Chatellard P, de Jesus M, Wurm F, Mermod N. 2001. Development of stable cell lines for production or regulated expression using matrix attachment regions. J Biotechnol 87:29–42.

Zhang Y, Jasin M. 2011. An essential role for CtIP in chromosomal translocation formation through an alternative end-joining pathway. Nat Struct Mol Biol 18:80–84.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

A role for homologous recombination proteins in cell cycle regulation

Kaja Kostyrko, Sandra Bosshard, Zuzanna Urban, and Nicolas Mermod*

Institute of Biotechnology; University of Lausanne; and Center for Biotechnology UNIL-EPFL; Lausanne, Switzerland

Eukaryotic cells respond to DNA breaks, especially double-stranded breaks (DSBs), by activating the DNA damage response (DDR), which encompasses DNA repair and cell cycle checkpoint signaling. The DNA damage signal is transmitted to the checkpoint machinery by a network of specialized DNA damage-recognizing and signal-transducing molecules. However, recent evidence suggests that DNA repair proteins themselves may also directly contribute to the checkpoint control. Here, we investigated the role of homologous recombination (HR) proteins in normal cell cycle regulation in the absence of exogenous DNA damage. For this purpose, we used Chinese Hamster Ovary (CHO) cells expressing the Fluorescent ubiquitination-based cell cycle indicators (Fucci). Systematic siRNA-mediated knockdown of HR genes in these cells demonstrated that the lack of several of these factors alters cell cycle distribution, albeit differentially. The knock-down of MDC1, Rad51 and Brca1 caused the cells to arrest in the G2 phase, suggesting that they may be required for the G2/M transition. In contrast, inhibition of the other HR factors, including several Rad51 paralogs and Rad50, led to the arrest in the G1/G0 phase. Moreover, reduced expression of Rad51B, Rad51C, CtIP and Rad50 induced entry into a quiescent G0-like phase. In conclusion, the lack of many HR factors may lead to cell cycle checkpoint activation, even in the absence of exogenous DNA damage, indicating that these proteins may play an essential role both in DNA repair and checkpoint signaling.

Introduction

DNA double strand breaks (DSBs), one of the most deleterious types of DNA lesions, can result from ionizing radiation or chemical agents, or from natural cellular processes such as DNA replication or maturation of the immune system genes. If left unrepaired, they constitute a major threat to genetic integrity and stability, possibly leading to cell death or carcinogenesis.[1] In response to DSBs, cells activate a network of DNA repair and signaling pathways, collectively termed the DNA damage response (DDR).[2-4] To allow time for DNA repair, the DDR machinery activates cell cycle checkpoints that arrest cell cycle progression until genome integrity is restored. The DDR-activated checkpoints include the G1/S, the intra-S and the G2/M transitions. The G1/S checkpoint, the one most sensitive to DNA damage, is defective in most human cancer cells.[5,6]

The Mre11/Rad50/Nbs1 (MRN) complex is among the first sensors of DSBs, subsequently activating Ataxia telangiectasia mutated (ATM).[7] ATM, a key protein kinase in the DDR network, is responsible for phosphorylation of many downstream DNA repair and cell cycle factors, including tumor suppressor p53, mediator of DNA-damage checkpoint 1 (MDC1), cell cycle checkpoint kinase 2 (Chk2), and breast cancer susceptibility protein 1 (Brca1).[8,9] The activation of these factors results in signaling cascades ultimately leading to cell cycle arrest. ATM-dependent phosphorylation of histone H2AX also induces global changes in the chromatin structure, leading to the recruitment of DNA repair proteins to the sites of damage.

Several specialized pathways act to repair DNA breaks in higher eukaryotic cells. One of the main pathways responsible for DSB repair is non-homologous end-joining (NHEJ). NHEJ is a fast process, based on a simple ligation of the 2 broken DNA ends, active throughout the entire cell cycle.[10] In the absence of functional NHEJ, cells were shown to use a highly error-prone, backup mechanism termed microhomology mediated end joining (MMEJ).[11,12] The third pathway, considered to be the most precise of all DSB repair mechanisms, is based on homologous recombination (HR).[13] HR requires extensive homology for

repair, and thus is primarily used in late S and G2 phases of the cell cycle, when the genetic material has been replicated and sister chromatids are available as repair template.

A key role in eukaryotic HR is played by the Rad51 recombinase, which coats ssDNA ends resulting from the initial processing of the DSB.[14,15] The DNA-bound Rad51 then searches for sequence homology along a cDNA strand and mediates pairing between the 2 strands. The Rad51 protein is essential, as the targeted knock-out of its gene leads to embryonic lethality in mice.[16] Other proteins involved in HR include CtIP, Brca2, Rad52, Rad54 and the 5 Rad51 paralogs: Rad51B, Rad51C, Rad51D, Xrcc2, and Xrcc3.[17–21] Rad51B, Rad51C, Rad51D and Xrcc2 together form the BCDX2 complex, which was proposed to facilitate the formation and stabilization of the Rad51 nucleofilament.[22] Rad51C also participates in the formation of a second complex with Xrcc3 termed CX3, which was reported to play an essential role in the final resolution of recombination intermediates.[23] The MRN complex, MDC1 and Brca1, which are components of the DDR response, also play a role in the initial steps of HR.[24-26]

It was recently proposed that HR proteins may also directly contribute to cell cycle control, in addition to their role in DNA repair.[27-29] The knock-down of Rad51 was shown to induce G2/M arrest, suggesting that this protein is required for the progression from the G2 phase to mitosis,[30-32] and Brca1 was reported to play a role in the regulation of the G2/M and intra-S checkpoints.[29,33] Rad51C was also proposed to contribute to cell cycle regulation, although there are conflicting reports as to its exact role. Rodrigue and others observed that the knock-down of Rad51C in human cells leads to arrest at the G2/M checkpoint, similarly to Rad51.[27] In another study, knock-down of Rad51C caused cells to escape the intra-S and G2/M checkpoints, thus allowing entry into mitosis.[28] Several other DSB repair proteins have also been proposed to participate in cell cycle progression.[34,35]

Here, we systematically assessed the role of HR factors in cell cycle regulation in the absence of exogenous DNA damage by silencing HR genes of Chinese Hamster Ovary (CHO) cells expressing the fluorescent ubiquitination-based cell cycle indicator (Fucci) probes.[36] We show that the knock-down of many HR factors, including Rad51, MDC1, Brca1, several Rad51 paralogs, CtIP, and Rad50 significantly affected cell cycle progression, albeit differentially. The knock-down of MDC1, Rad51 and Brca1 caused the cells to arrest at the G2/M checkpoint, suggesting that these factors may be required for the transition through the G2 phase and entry into mitosis. In contrast, the absence of the remaining HR proteins increased the proportion of G1/G0 phase cells, indicating that their deficiency may cause the cells to escape the G2/M checkpoint, divide and subsequently become arrested in the G1 phase. We also observed that knock-down of Rad51B, Rad51C, CtIP and Rad50 increased the proportion of G0 cells, suggesting that the absence of these factors may cause cells to enter a quiescent state. We conclude that many HR proteins may regulate cell cycle progression in addition to their known role in DSB repair.

Results

Characterization of CHO Fucci cells

The fluorescent ubiquitination-based cell cycle indicator (Fucci) system enables the simultaneous observation of multiple cell cycle phases in living cells.[36] It is based on the expression, ubiquitination and degradation of the cell cycle-dependent human Cdt1 and Geminin proteins fused to fluorescent markers, the monomeric Kusabira Orange 2 (mKO2) and monomeric Azami Green (mAG), respectively. Depending on the fluorophore levels, 4 main cell subpopulations can be visualized, namely early G1 phase cells, G1/G0 cells, early S cells, and late S, G2 and M phase cells (**Fig. 1**). The freshly divided early G1 phase cells are non-fluorescent, but they start to express and accumulate the mKO2-hCdt1(30/120) chimeric protein as they progress through the G1 phase. In the early S phase, as the mKO2-hCdt1(30/120) becomes ubiquitinated and degraded, expression
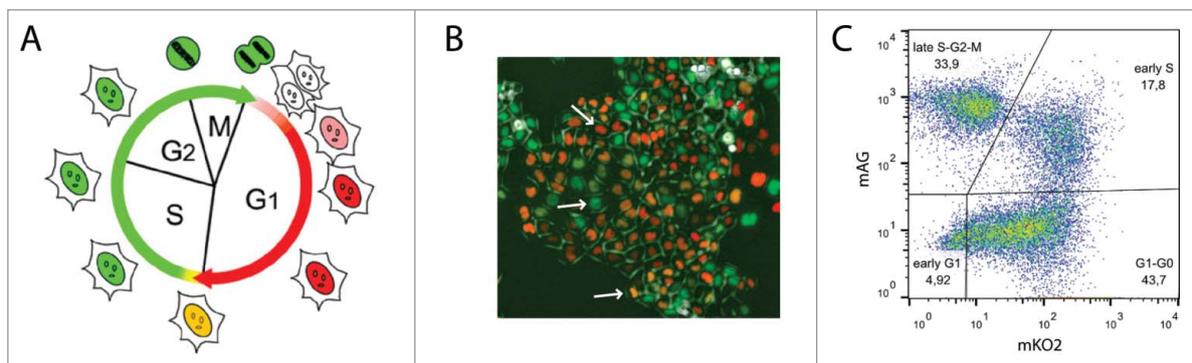
Figure 1. Characterization of CHO Fucci cells. (**A**) Scheme of the Fucci cell cycle indicator assay.[36] (**B**) Typical fluorescence image of non-synchronized CHO Fucci cells. Arrows point to cells in G1/G0 (top), late S, G2, and M (middle) and early S phase cells (bottom). (**C**) Flow cytometry analysis of CHO Fucci cells with 4 subpopulations of cells: mKO2-positive G1/G0 cells (mKO2), mAG-positive late S, G2 and M cells (mAG), double positive (mKO2+mAG+) early S cells, and a double negative (mKO2-, mAG-) early G1 cells.

of the mAG-hGem(1/110) fusion protein yields yellow cells. Expression of mAG-hGem(1/110) and green fluorescence is then maintained from late S until the end of M phase.

To confirm that the CHO Fucci cells express the fluorescent probes in a cell cycle dependent manner, we synchronized the culture by serum deprivation, to induce a cell cycle arrest in the G1/G0 phase, or by contact inhibition, which can cause G1 arrest.[37,38] Over 90% of serum-deprived cells displayed red fluorescence indicative of the G1/G0 phase (**Figs. S1A, S2A**). In cultures grown to confluence, a lower proportion (50%) of the red fluorescent G0/G1 cells was obtained, indicating that this method did not efficiently synchronize cells into the same phase of the cell cycle (**Figs. S1B, S2B**). We subsequently released the cells from cycle arrest and monitored the cell cycle phase distribution 16–24 h and 40–48 h following release. Within 20 h, the serum deprived cells had progressed through the S phase into the G2 and M phases, as evidenced by the accumulation of green fluorescent cells. Most serum deprived cells remained synchronized until 48 h post release, whereas the cells grown to confluence became desynchronized. Overall, the CHO Fucci cells appeared to express the fluorescent probes in a cell cycle dependent manner and therefore may serve as a model to study cell cycle regulation.

Knock-down of HR factors differentially influences cell cycle distribution

To assess the role of recombination factors in cell cycle regulation, we subsequently treated the CHO Fucci cells with a panel of short interfering RNAs (siRNAs) targeting HR protein mRNAs. The efficiency of siRNA knock-down was assessed by qPCR (**Fig. S3A**). Rad51 and Rad51D knock-down was also confirmed by western blot (**Fig. S3B**). A non-targeting siRNA, as well as an siRNA targeting Cyclin D1, which is required for the progression through the G1/S checkpoint,[39] were used as a negative and positive control, respectively.

The siRNA treatments had little effect on the proportion of freshly divided early G1 phase cells, except for a decrease of this sub-population, albeit not statistically significant, with Rad51C siRNA (**Fig. 2A**). However, the number of G1/G0 cells was significantly increased upon treatment with the Cyclin D1 siRNA, as expected from a G1/S checkpoint arrest (**Fig. 2B**). Interestingly, the knock-down of 3 Rad51 paralogs – Rad51B, -C, and -D, as well as Rad50, one of the MRN components, and to a lower extent CtIP, had a similar effect. The accumulation of G1/G0 cells upon the knock-down of these factors suggests that the progression through the G1 phase may be perturbed in their absence, possibly involving the G1/S checkpoint. This is surprising, as these proteins are thought to operate primarily in the late S and G2 phases of the cell cycle, when HR is most active. However, it is also possible that these factors may be necessary for the G2/M checkpoint activation in response to DNA damage. In this scenario, the absence of these factors would cause the cells to circumvent arrest and enter mitosis despite the presence of unrepaired breaks, which later on would activate the G1/S checkpoint.

As expected, the proportion of early S phase cells was significantly decreased by the knock-down of Cyclin D1, due to the defective G1/S transition (**Fig. 2C**). We also noted a decrease in this subpopulation in the presence of Xrcc2 siRNA, which, however, did not correlate with an increase in the number of G1/G0 phase cells. Distinctly, the knock-down of MDC1, Rad51 and Brca1 resulted in a significant accumulation of green-fluorescent late S, G2 or M phase cells (**Fig. 2D**). This further supported the view that these proteins are required for the progression through the G2/M checkpoint.[29–31,40,41] The percentage of green fluorescent cells was also slightly increased in the presence of Xrcc2 siRNA. In contrast, the silencing of Cyclin D1 as well as Rad51B, -C, -D, Rad50 and CtIP, resulted in a significant decrease in this subpopulation, which may result from the aforementioned lack of G2/M checkpoint activation upon DNA damage and in the following accumulation of G1/G0 phase cells.

In conclusion, several HR proteins appear to be involved in cell cycle regulation, albeit differentially. MDC1, Rad51 and Brca1 seem to be essential for the progression from S and G2 phases into mitosis, while Rad51 paralogs, Rad51B, -C and -D, as well as the DNA end resection enzymes, Rad50 and CtIP may be required for activating the G2/M checkpoint in response to damage and/or progression through the G1/S checkpoint.

Knock-down of specific HR proteins induces entry into G0 phase and cell cycle arrest

In addition to increasing the number of mKO2-positive cells, we also observed that the knock-down of Cyclin D1 and several HR factors increased the level of mKO2 fluorescence (**Fig. S4**). A detailed analysis of the cells knocked-down for Rad51B, Rad51C, CtIP, and Rad50 revealed a sub-population of cells with distinctly higher mKO2 fluorescence patterns (**Fig. 3A**). A recently published report identified low- and high mKO2-expressing cells as cycling G1 and quiescent G0 cells, respectively.[42] Consistently, we observed an increase in the number of these bright red fluorescent cells upon serum starvation (**Fig. S5**), confirming the view that this sub-population represents non-cycling G0 cells.

We therefore set out to quantify the G1 and G0 sub-populations in cells transfected with HR siRNAs. In controls, as well as in most siRNA-treated samples, the G0 phase cells constituted only approximately 5–10% of the population (**Fig. 3B**). However, treatment with Cyclin D1 siRNA increased the number of quiescent cells to 30%. This is consistent with previous studies showing that Cyclin D1 deficiency causes entry in the G0 phase.[43] Interestingly, we also observed a very significant increase in the number of G0 cells upon the knock-down of Rad51B, Rad51C, CtIP, and Rad50. This was especially striking in the presence of Rad51B and Rad51C siRNAs, where G0 cells comprised up to 40% of the entire population. This implied that the absence of these HR factors may constitute a signal to enter the quiescent state.

We next sought to investigate whether the altered cell cycle distribution observed upon siRNA knock-down of HR factors results from a cell cycle arrest or from a delayed cycle progression. We focused our attention on Rad51 and Rad51C, the 2 HR proteins with pronounced, but distinct effects on the cell cycle. We synchronized the siRNA-treated cells in early S phase by sorting
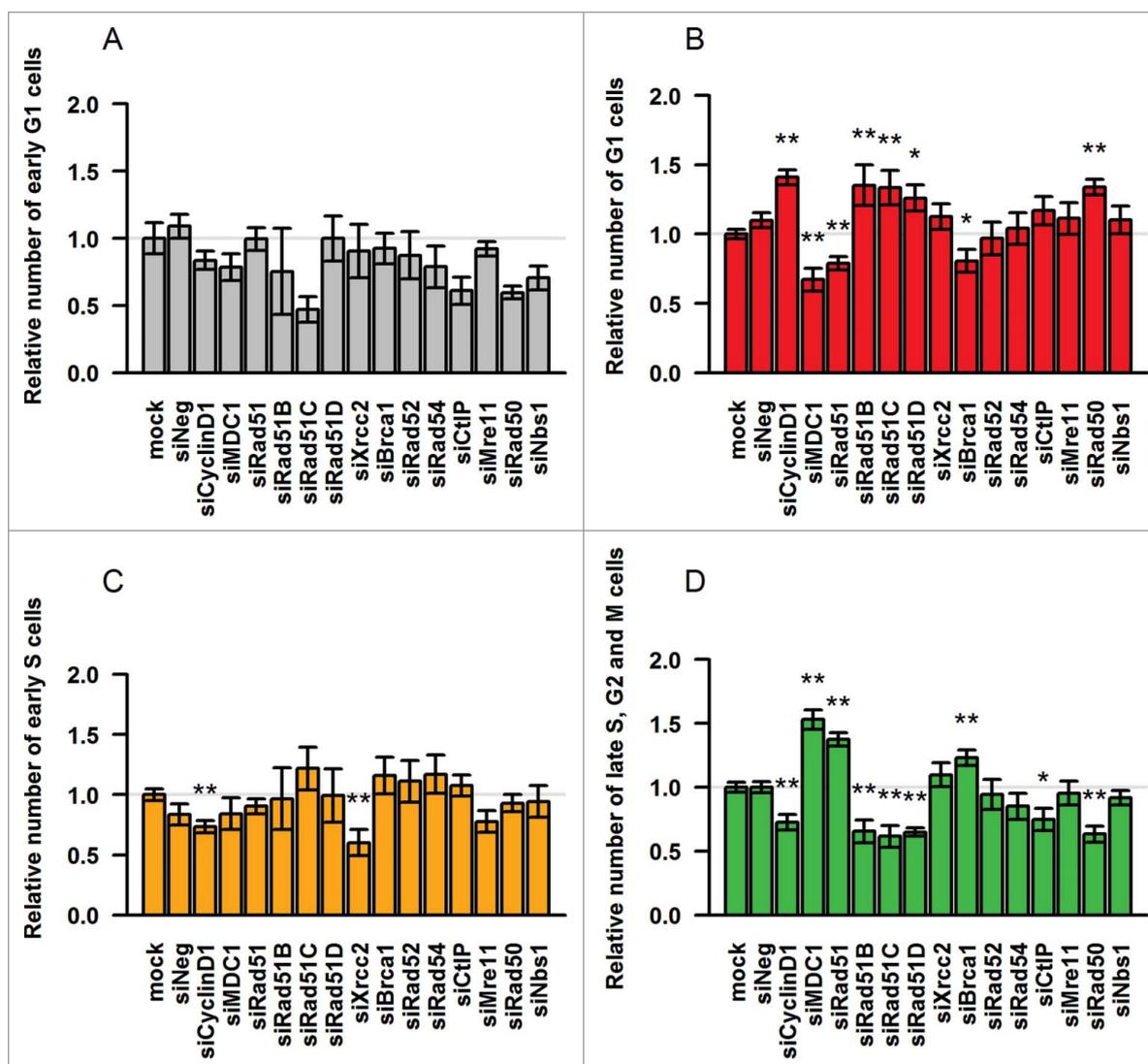
Figure 2. Knock-down of HR factors affects cell cycle distribution of CHO Fucci cells. Graphs show relative numbers of cells in (**A**) early G1 phase, (**B**) G1/G0 phase, (**C**) early S phase, (**D**) late S, G2 and M phases. Results are shown as fold change over the data obtained from mock-treated cells (mock). Mean of $\geq$ 3 experiments, error bars show s.e.m. Asterisks indicate significant differences between siRNA-treated samples and mock control. Statistical significance relative to mock was determined by unpaired Student's t-test with Benjamini-Hochberg correction; significance level $P < 0.05$ (*), $P < 0.01$ (**).

double positive mKO2+mAG+ cells, and subsequently analyzed their cell cycle distribution (**Fig. 4A**) and the presence of cell doublets and quadruplets (**Fig. S6A**) 1 or 2 d after sorting. Since the doubling time of CHO DG44 cells is approximately 12–14 h,[37,44] the cells should have completed a first cell cycle and be nearly completing a second one during the 18 h between sorting and the first measurements. Accordingly, the majority of cells should be green fluorescent at the beginning of the analysis, 40 h post transfection, and green fluorescence should decrease thereafter upon mitosis completion. This was indeed the case for untreated cells, and for cells transfected with the non-targeting siRNA (**Fig. 4B,C; Fig. S6A, B**). Both populations also showed a similar cell cycle distribution over time, with another cell division around 64 h to 72 h post transfection. The time between

these mitosis was also consistent with the normal CHO cell cycle duration.

Cells treated with the Cyclin D1 siRNA initially displayed a cell cycle pattern similar to that of the cells treated with the non-targeted control siRNAs, with one division 44–48 h post transfection (**Fig. 4D; Fig. S6C**). However, the next division to an 8-cell stage was delayed to 64–68 h, after which the accumulation of G1 and G0 phase red-fluorescent cells occurred (**Fig. 4C and D; Figs. S6B and S6C**). In Rad51-depleted cells, the first division was notably delayed compared to the controls (48–64 h), despite the cell enlargement, after which most cells appeared not to divide anymore, except for a small portion of cells (approx. 20%) which underwent a second division at around 64–68 h after transfection. (**Fig. 4E; Fig. S6D**). The percentage
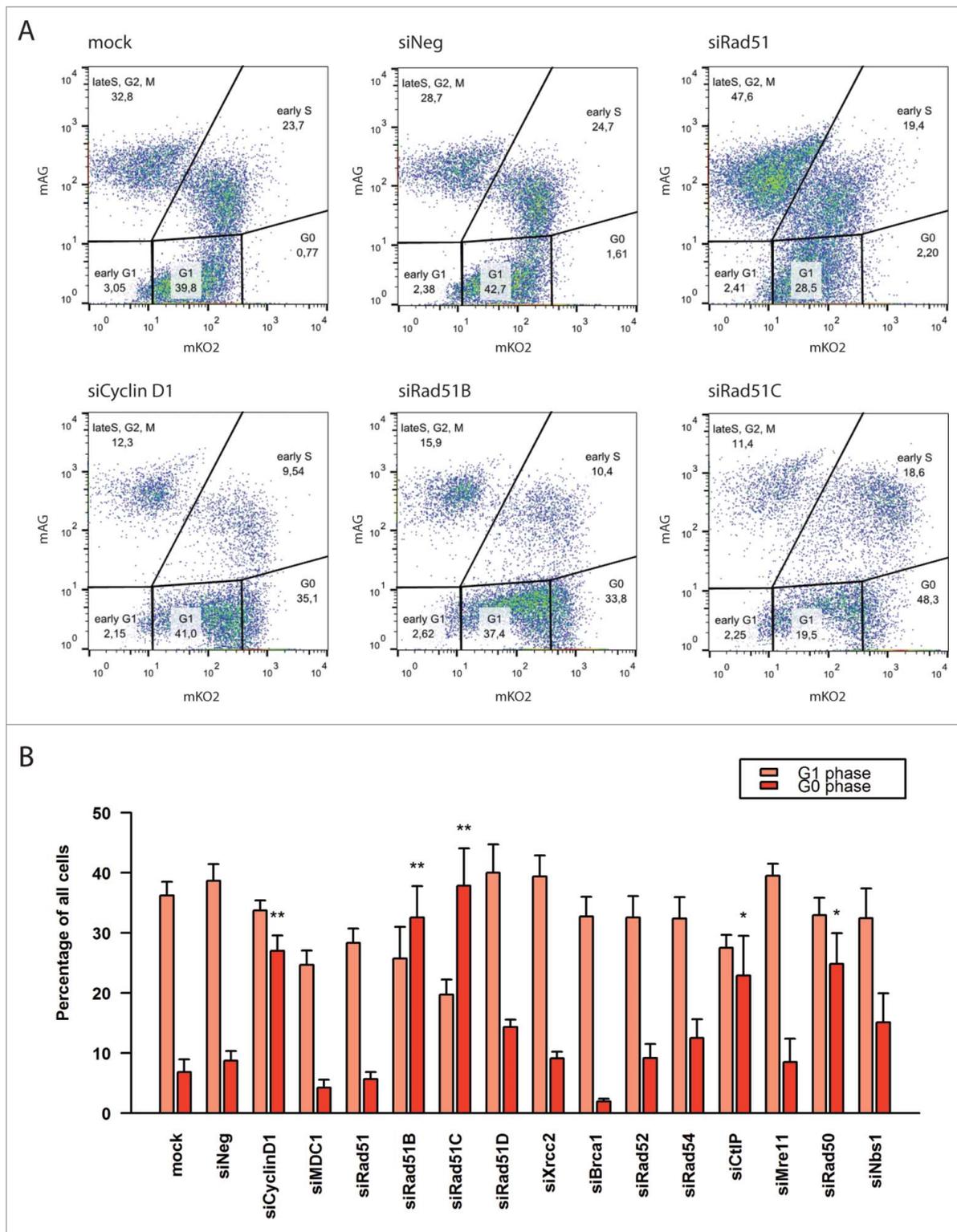
Figure 3. Knock-down of Cyclin D1, Rad51B, Rad51C, CtIP and Rad50 induces G0-like quiescence. (**A**) FACS plots of siRNA-treated cells. (**B**) Percentages of cells in G1 and G0 phases. The mean of values from ≥ 3 experiments is displayed, and error bars indicate the standard errors of the mean. Statistical significance relative to mock was determined by unpaired Student's t-test with Benjamini-Hochberg correction; significance level $P < 0.05$ (*), $P < 0.01$ (**).
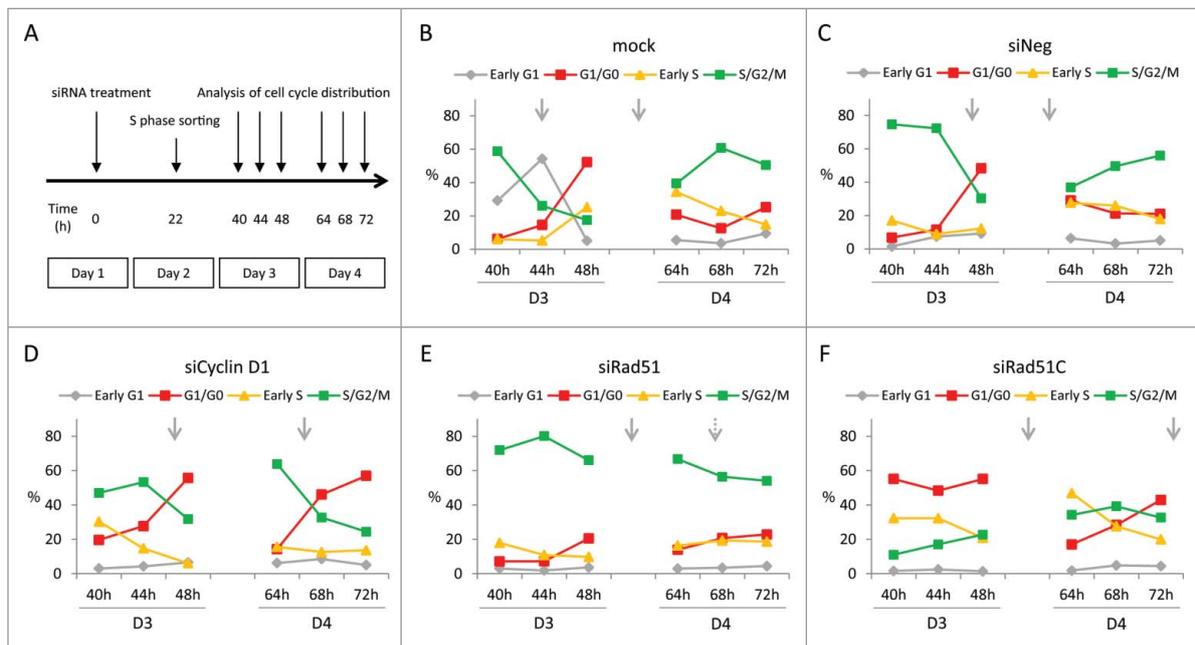
Figure 4. Effect of Rad51 and Rad51C knock-down on cell cycle progression of CHO Fucci cells synchronized in early (S)phase. (**A**) General outline of the procedure. (**B**–**F**) Percentages of cells in a given cell cycle phase at day 3 (D3) and day 4 (D4) post siRNA transfection. Arrows indicate the estimated average time of cell division, dotted arrow indicates a cell division of a subpopulation of cells.

of green fluorescent cells stayed high throughout the time course, implying that most cells were arrested in late S, G2 or M phase. Thus, a near complete Rad51 deficiency may have caused the cells to arrest in the late phases of the cycle, likely at the G2/M checkpoint, whereas a milder Rad51 depletion, as due to lower knock-down efficiency, may have resulted in delayed cell cycle progression.

Cells treated with Rad51C siRNA also divided later than the controls (between 48 and 64 h) pointing to a delay in the cell cycle (**Fig. 4F**; **Fig. S6E**). However, Rad51C-deficiency resulted

in a steadily elevated proportion of cells in G1 and G0 phases, resembling the effect of the Cyclin D1 knock-down. Indeed, cluster analysis demonstrated that Rad51C-depleted cells grouped together with Cyclin D1 siRNA-treated cells (data not shown). However, the majority of these red-fluorescent cells appeared to be quiescent, reaching 50% of the entire population (**Fig. 5**). This was more than observed in the presence of Cyclin D1 siRNA (10–20%). After the delayed first cycle, most of the Rad51C-depleted cells underwent a second division within a normal time of 12–14 h, although some cells remained arrested
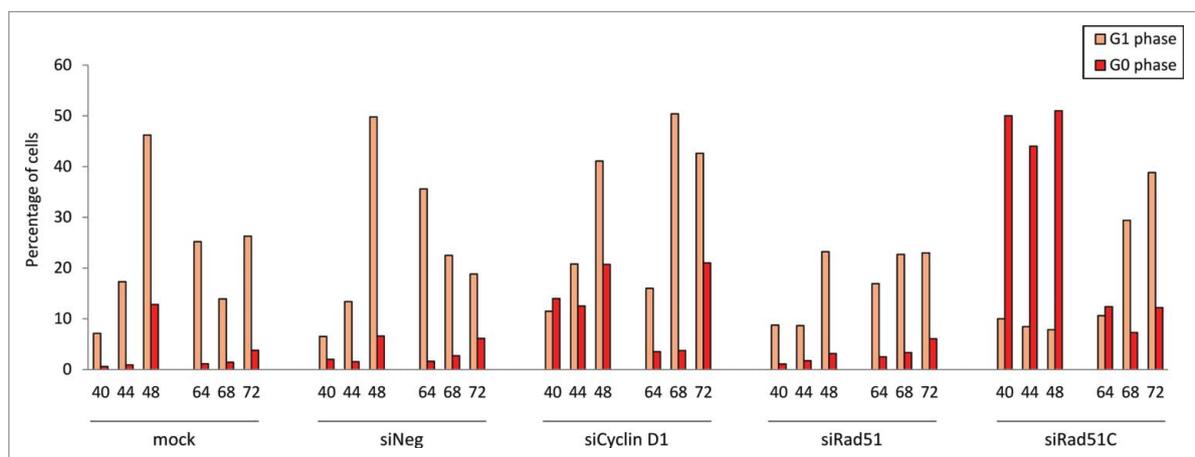
Figure 5. Effect of HR protein depletion on the percentages of quiescent cells in CHO Fucci cells synchronized in early S phase. Numbers below the bars represent time (in hours) post siRNA transfection. The siRNA targets are indicated below the plot. Mock indicates cells treated with the transfection reagent only.

in a prolonged G1, G0-like phase (**Fig. 5**; **Fig. S6E**). This indicated that, out of all the cells initially arrested in a quiescent G0-like phase, some were later released from the block and re-entered the cell cycle. At 64 h post transfection, most of the cells had reverted to a G1 profile, indicating that the cell Rad51C depletion-mediated cell cycle arrest is reversible.

Overall, we concluded that Rad51 knock-down causes an arrest at the late phases of the cell cycle, most probably at the G2/M checkpoint, or delayed cell cycle progression in cells circumventing the arrest. In contrast, Rad51C depletion causes entry into a quiescent G0-like phase, most likely due to the activation of the G1/S checkpoint.

Discussion

The DNA damage response encompasses many functionally interconnected pathways, including cell cycle checkpoint signaling and DNA repair. Many proteins participating in cell cycle regulation are also known to control DNA repair, whereas the converse was not known to be true for the DNA repair factors. Recent evidence, however, suggested that some DSB repair proteins may also be implicated in cell cycle regulation in response to exogenous DNA damage.[28,29,34] To further explore this connection between DNA repair proteins and cell cycle machinery, we analyzed the effect of HR protein depletion on the CHO cell cycle in the absence of induced DNA damage. We found that the knock-down of several HR factors altered cell cycle progression, suggesting that they may play a direct role in cell cycle control. However, the lack of proteins affected the cell cycle differentially, despite their being part of the same DNA repair pathway.

In the late S and G2 phases, the presence of unrepaired DNA breaks results in the activation of the G2/M cell cycle checkpoint by DNA damage sensing factors, e.g. ATM, ATR, p53, and Chk1.[45,46] We anticipated that the absence of crucial HR factors might lead to the accumulation of unrepaired DSBs, thereby stimulating the DDR to arrest the cells before division. Surprisingly, the knock-down of the majority of HR proteins, including 3 Rad51 paralogs, CtIP, and Rad50, failed to arrest the cells at the G2/M checkpoint. Instead, the presence of these factors seemed to be necessary for the progression from G1 to S phase. Only the loss of MDC1, Rad51 and Brca1 led to an accumulation of late S, G2 and M cells, likely due to the activation of the intra-S and/or G2/M checkpoints. This could indicate that the knock-down of these 3 genes, which products are involved in the early steps of the HR pathway, led to the accumulation of enough endogenous DNA damage to trigger the G2/M checkpoint. The absence of the remaining HR factors would potentially inhibit the repair process incompletely, instead rendering it more error-prone. These cells would still enter mitosis, but due to the accumulation of imprecisely or incompletely repaired DSBs, they may become arrested at the next G1/S checkpoint, explaining the accumulation of cells in the G1/G0 phase. In a recent study, Shibata and co-workers estimated that only about 15% of DSBs occurring in the G2 phase are repaired by the HR pathway, while the remaining breaks are efficiently repaired by other mechanisms.[47] This, together with our results obtained in the absence of induced DNA damage, may indicate that the effects of HR protein knock-down described here are not merely due to the accumulation of unrepaired DSBs. Instead, these HR proteins may play a more direct role in the cell cycle, for instance by interacting with cell cycle signaling factors, cyclins or cyclin-dependent kinases.

Our results obtained with Rad51B and Rad51C siRNAs contrast recently published observations that the knock-down of these Rad51 paralogs blocks progression through the G/M checkpoint in HeLa cells.[27] However, it is possible that the absence of these proteins in human cells may have a different impact on HR and/or the cell cycle than it does in CHO cells, which divide 2 times faster than HeLa cells and display different kinetics of DSB repair.[48,49]

In the present study, we also observed that apart from a defect in the G1/S transition, the knock-down of Rad51B, Rad51C, CtIP and Rad50 also caused entry into a non-proliferative G0 phase resembling the effect of the Cyclin D1 knock-down. This suggests that the absence of these proteins may constitute a signal for the cells to withdraw from the cell cycle. This may in part explain the characteristic enhanced proliferation of cancer cells, in which HR proteins are often overexpressed. Taken together, these results indicate that many HR components are required for normal cell cycle progression, at least in CHO cells. Thus, their expression to levels that are sufficiently high to handle spontaneous DSB would act as one of the regulatory cues that control progression through cell cycle checkpoints. It will therefore be of interest to decipher how these proteins transmit the signals to the cell cycle control machinery, and what their molecular targets may be, to further understand their function in the cell cycle regulation network.

Materials and Methods

CHO cells expressing Fucci probes

Adherent CHO DG44 cells[50] were cultivated in DMEM/F12+GlutaMAX^{TM} supplemented with 1x HT and 10% fetal bovine serum (FBS) (Gibco, Invitrogen), and with antibiotic-antimycotic solution (Sigma-Aldrich, #A5955). CHO Fucci cells were constructed using lentiviral vectors carrying the red and green fluorescent ubiquitination-based cell cycle indicator (Fucci) cassettes.[36] The red Fucci cassette contains a monomeric version of Kusabira Orange 2 (mKO2) reporter gene fused to a truncated human Cdt1 (hCdt1, amino acids 30–120). The mKO2-hCdt1(30/120) protein is expressed in G1 phase and degraded at the onset of the S phase. The green Fucci cassette contains the monomeric version of Azami green (mAG) reporter gene fused to the 110 amino acid N-terminus of human Geminin (hGem amino acids 1–110). The mKO2-hGem(1/110) protein accumulates through S, G2 and M phases of the cell cycle and is degraded in the metaphase/anaphase transition of mitosis. The lentiviral constructs were kindly provided by M. Lutolf (EPFL, Lausanne, Switzerland). Briefly,

the cells were transduced with a 1:1 ratio of mKO2 and mAG vectors at MOI 50. Three weeks after transduction double positive (mKO2+mAG+) clones were single-cell sorted by fluorescence-activated cell sorting (FACS) (FACSAria II sorter, Becton-Dickinson, Allschwil Switzerland). A single clone with similar levels of mKO2 and mAG fluorescence intensity was selected for subsequent experiments.

Cell synchronization

For starvation synchronization CHO cells were grown for 72 h in medium supplemented with 0.2% FBS (Gibco, Invitrogen). For synchronization through contact inhibition cells were grown for 3–5 d until complete confluency. Both methods synchronize the cells in G1/G0 phase. To reinitiate cell cycle progression cells were replated at lower density in complete medium. For early S phase synchronization 10,000 double positive (mAG+mKO2+) cells were sorted by FACS (FACSAria II sorter, Becton-Dickinson, Allschwil Switzerland) into each well of a 12-well plate.

siRNA transfection

Small interfering RNA duplexes were specifically designed to target the Chinese hamster homologs of HR genes. The siRNAs were designed and provided by Microsynth AG (Balgach, Switzerland). Three RNA duplexes were designed per gene to increase the probability of successful knock-down. Three negative (non-targeting) siRNAs were also designed as controls. For siRNA-mediated knock-down, CHO-Fucci cells were transfected with equimolar amounts of 3 siRNA duplexes at a final concentration of 50 nM using Lipofectamine RNAiMAX, according to the manufacturer's instructions (Invitrogen). After 72 h cells were analyzed using the Axio Observer.A1 microscope (Zeiss, Jena, Germany).

Flow cytometry

For flow cytometry cells were harvested 72 h following siRNA transfection, resuspended in 0.5 ml of PBS with 2% FBS (Gibco, Invitrogen), and analyzed using the CyAn analyzer (Beckman Coulter, Nyon, Switzerland). Acquired data was analyzed using the FlowJo software (Tree Star Inc., Ashland, OR, USA). For the time point experiment cells were harvested every 4 h at 40, 44, 48, 64, 68 and 72 h post transfection, fixed in PBS with 4% PFA (Merck) (1:2 v/v) and analyzed by flow cytometry.

Supplemental Material

Supplemental data for this article can be accessed on the publisher's website.

Author Contributions

NM conceived the first ideas leading to this project. KK and NM designed the experiments and wrote the manuscript. KK, SB, ZU performed and analyzed the experiments.

References

1. Shiloh Y, Lehmann AR. Maintaining integrity. Nat Cell Biol 2004; 6:923-8; PMID:15459720; http://dx.doi.org/10.1038/ncb1004-923
2. Hiom K. Coping with DNA double strand breaks. DNA Repair (Amst) 2010; 9:1256-63; PMID:21115283; http://dx.doi.org/10.1016/j.dnarep.2010.09.018
3. Lord CJ, Ashworth A. The DNA damage response and cancer therapy. Nature 2012; 481:287-94; PMID:22258607; http://dx.doi.org/10.1038/nature10760
4. Jackson SP, Bartek J. The DNA-damage response in human biology and disease. Nature 2009; 461:1071-8; PMID:19847258; http://dx.doi.org/10.1038/nature08467
5. Löbrich M, Jeggo PA. The impact of a negligent G2/M checkpoint on genomic instability and cancer induction. Nat Rev Cancer 2007; 7:861-9; PMID:17943134; http://dx.doi.org/10.1038/nrc2248
6. Lapenna S, Giordano A. Cell cycle kinases as therapeutic targets for cancer. Nat Rev Drug Discov 2009; 8:547-66; PMID:19568282; http://dx.doi.org/10.1038/nrd2907

7. Lee J-H, Paull TT. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. Science 2005; 308:551-4; PMID:15790808; http://dx.doi.org/10.1126/science.1108297
8. Shiloh Y. The ATM-mediated DNA-damage response: taking shape. Trends Biochem Sci 2006; 31:402-10; PMID:16774833; http://dx.doi.org/10.1016/j.tibs.2006.05.004
9. Shiloh Y, Ziv Y. The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. Nat Rev Mol Cell Biol 2013; 14:197-210; http://dx.doi.org/10.1038/nrm3546
10. Pastwa E, Błasiak J. Non-homologous DNA end joining. Acta Biochem Pol 2003; 50:891-908
11. Iliakis G. Backup pathways of NHEJ in cells of higher eukaryotes: cell cycle dependence. Radiother Oncol 2009; 92:310-5; PMID:19604590; http://dx.doi.org/10.1016/j.radonc.2009.06.024
12. Kabotyanski EB, Gomelsky L, Han J, Stamato TD, Roth DB. Double-strand break repair in Ku86- and XRCC4-deficient cells. Nucleic Acids Res 1998; 26:27-31; PMID:9399794; http://dx.doi.org/10.1093/nar/26.23.5333

13. San Filippo J, Sung P, Klein H. Mechanism of eukaryotic homologous recombination. Annu Rev Biochem 2008; 77:229-57; PMID:18275380; http://dx.doi.org/10.1146/annurev.biochem.77.061306.125255
14. Shibata A, Conrad S, Birraux J, Geuting V, Barton O, Ismail A, Kakarougkas A, Meek K, Taucher-Scholz G, Löbrich M, et al. Factors determining DNA double-strand break repair pathway choice in G2 phase. EMBO J 2011; 30:1079-92; PMID:21317870; http://dx.doi.org/10.1038/emboj.2011.27
15. Baumann P, West SC. Role of the human RAD51 protein in homologous recombination and double-stranded-break repair. Trends Biochem Sci 1998; 23:247-51; PMID:9697414; http://dx.doi.org/10.1016/S0968-0004(98)01232-8
16. Tsuzuki T, Fujii Y, Sakumi K, Tominaga Y, Nakao K, Sekiguchi M, Matsushiro A, Yoshimura Y, Morita T. Targeted disruption of the Rad51 gene leads to lethality in embryonic mice. Proc Natl Acad Sci U S A 1996; 93:6236-40; PMID:8692798; http://dx.doi.org/10.1073/pnas.93.13.6236
17. Suwaki N, Klare K, Tarsounas M. RAD51 paralogs: roles in DNA damage signalling, recombinational repair and tumorigenesis. Semin Cell Dev Biol 2011;

22:898-905; PMID:21821141; http://dx.doi.org/10.1016/j.semcdb.2011.07.019

18. Swagemakers SMA, Essers J, de Wit J, Hoeijmakers JHJ, Kanaar R. The Human Rad54 Recombinational DNA Repair Protein Is a Double-stranded DNA-dependent ATPase. J Biol Chem 1998; 273:28292-7; PMID:9774452; http://dx.doi.org/10.1074/jbc.273.43.28292

19. Moynahan ME, Pierce AJ, Jasin M. BRCA2 is required for homology-directed repair of chromosomal breaks. Mol Cell 2001; 7:263-72; PMID:11239455; http://dx.doi.org/10.1016/S1097-2765(01)00174-5

20. Feng Z, Scott SP, Bussen W, Sharma GG, Guo G, Pandita TK, Powell SN. Rad52 inactivation is synthetically lethal with BRCA2 deficiency. Proc Natl Acad Sci U S A 2011; 108:686-91; PMID:21148102; http://dx.doi.org/10.1073/pnas.1010959107

21. Sartori AA, Lukas C, Coates J, Mistrik M, Fu S, Bartek J, Baer R, Lukas J, Jackson SP. Human CtIP promotes DNA end resection. Nature 2007; 450:509-14; PMID:17965729; http://dx.doi.org/10.1038/nature06337

22. Masson JY, Tarsounas MC, Stasiak AZ, Stasiak A, Shah R, McIlwraith MJ, Benson FE, West SC. Identification and purification of two distinct complexes containing the five RAD51 paralogs. Genes Dev 2001; 15:3296-307; PMID:11751635; http://dx.doi.org/10.1101/gad.947001

23. Liu Y, Tarsounas M, O'regan P, West SC. Role of RAD51C and XRCC3 in genetic recombination and DNA repair. J Biol Chem 2007; 282:1973-9; PMID:17114795; http://dx.doi.org/10.1074/jbc.M609066200

24. Lamarche BJ, Orazio NI, Weitzman MD. The MRN complex in double-strand break repair and telomere maintenance. FEBS Lett 2010; 584:3682-95; PMID:20655309; http://dx.doi.org/10.1016/j.febslet.2010.07.029

25. Zhang J, Ma Z, Treszezamsky A, Powell SN. MDC1 interacts with Rad51 and facilitates homologous recombination. Nat Struct Mol Biol 2005; 12:902-9; PMID:16186822; http://dx.doi.org/10.1038/nsmb991

26. Jasin M. Homologous repair of DNA damage and tumorigenesis: the BRCA connection. Oncogene 2002; 21:8981-93; PMID:12483514; http://dx.doi.org/10.1038/sj.onc.1206176

27. Rodrigue A, Coulombe Y, Jacquet K, Gagne J-P, Roques C, Gobeil S, Poirier G, Masson J-Y. The RAD51 paralogs ensure cellular protection against mitotic defects and aneuploidy. J Cell Sci 2013; 126:348-59; PMID:23108668; http://dx.doi.org/10.1242/jcs.114595

28. Badie S, Liao C, Thanasoula M, Barber P, Hill MA, Tarsounas M. RAD51C facilitates checkpoint signaling by promoting CHK2 phosphorylation. J Cell Biol 2009; 185:587-600; PMID:19451272; http://dx.doi.org/10.1083/jcb.200811079

29. Yarden RI, Pardo-Reoyo S, Sgagias M, Cowan KH, Brody LC. BRCA1 regulates the G2/M checkpoint by activating Chk1 kinase upon DNA damage. Nat Genet 2002; 30:285-9; PMID:11836499; http://dx.doi.org/10.1038/ng837

30. Shtam TA, Kovalev RA, Varfolomeeva EY, Makarov EM, Kil YV, Filatov MV. Exosomes are natural carriers of exogenous siRNA to human cells in vitro. Cell Commun Signal 2013; 11:88; PMID:24245560; http://dx.doi.org/10.1186/1478-811X-11-88

31. Du L-Q, Wang Y, Wang H, Cao J, Liu Q, Fan F-Y. Knockdown of Rad51 expression induces radiation- and chemo-sensitivity in osteosarcoma cells. Med Oncol 2011; 28:1481-7; PMID:20625943

32. Le Cigne A, Menil-Philippot V, Fleury F, Takahashi M, Thiriet C. Transient expression of RAD51 in the late G2-phase is required for cell cycle progression in synchronous Physarum cells. Genes to Cells 2014; 19:755-65; PMID:25200281; http://dx.doi.org/10.1111/gtc.12174

33. Wiltshire T, Senft J, Wang Y, Konat GW, Wenger SL, Reed E, Wang W. BRCA1 contributes to cell cycle arrest and chemoresistance in response to the anticancer agent irofulven. Mol Pharmacol 2007; 71:1051-60; PMID:17229870

34. Xu Z-Y, Loignon M, Han F-Y, Panasci L, Aloyz R. Xrcc3 induces cisplatin resistance by stimulation of Rad51-related recombinational repair, S-phase checkpoint activation, and reduced apoptosis. J Pharmacol Exp Ther 2005; 314:495-505; PMID:15843498

35. Langerak P, Russell P. Regulatory networks integrating cell cycle control with DNA damage checkpoints and double-strand break repair. Philos Trans R Soc Lond B Biol Sci 2011; 366:3562-71; PMID:22084383

36. Sakaue-Sawano A, Kurokawa H, Morimura T, Hanyu A, Hama H, Osawa H, Kashiwagi S, Fukami K, Miyata T, Miyoshi H, et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. Cell 2008; 132:487-98; PMID:18267078; http://dx.doi.org/10.1016/j.cell.2007.12.033

37. Grosjean F, Batard P, Jordan M, Wurm FM. S-phase synchronized CHO cells show elevated transfection efficiency and expression using CaPi. Cytotechnology 2002; 38:57-62; PMID:19003087; http://dx.doi.org/10.1023/A:1021197830091

38. Rosner M, Schipany K, Hengstschläger M. Merging high-quality biochemical fractionation with a refined flow cytometry approach to monitor nucleocytoplasmic protein expression throughout the unperturbed mammalian cell cycle. Nat Protoc 2013; 8:602-26; PMID:23449254

39. Baldin V, Lukas J, Marcote MJ, Pagano M, Draetta G. Cyclin D1 is a nuclear protein required for cell cycle progression in G1. Genes Dev 1993; 7:812-21; PMID:8491378

40. Lou Z, Minter-Dykhouse K, Franco S, Gostissa M, Rivera MA, Celeste A, Manis JP, van Deursen J, Nussenzweig A, Paull TT, et al. MDC1 maintains genomic stability by participating in the amplification of ATM-dependent DNA damage signals. Mol Cell 2006; 21:187-200; PMID:16427009

41. Townsend K, Mason H, Blackford AN, Miller ES, Chapman JR, Sedgwick GG, Barone G, Turnell AS, Stewart GS. Mediator of DNA damage checkpoint 1 (MDC1) regulates mitotic progression. J Biol Chem 2009; 284:33939-48; PMID:19826003

42. Tomura M, Sakaue-Sawano A, Mori Y, Takase-Utsugi M, Hata A, Ohtawa K, Kanagawa O, Miyawaki A. Contrasting quiescent G0 phase with mitotic cell cycling in the mouse immune system. PLoS One 2013; 8:1-10; http://dx.doi.org/10.1371/journal.pone.0073801

43. Hsu L-C, Huang X, Seasholtz S, Potter DM, Gollin SM. Gene amplification and overexpression of protein phosphatase 1alpha in oral squamous cell carcinoma cell lines. Oncogene 2006; 25:5517-26; PMID:16619035; http://dx.doi.org/10.1038/sj.onc.1209563

44. Grandjean M, Girod P-A, Calabrese D, Kostyrko K, Wicht M, Yerly F, Mazza C, Beckmann JS, Martinet D, Mermod N. High-level transgene expression by homologous recombination-mediated gene transfer. Nucleic Acids Res 2011; 39:e104; PMID:21652640

45. Taylor WR, Stark GR. Regulation of the G2/M transition by p53. Oncogene 2001; 20:1803-15; PMID:11313928; http://dx.doi.org/10.1038/sj.onc.1204252

46. Lin JJ, Dutta A. ATR pathway is the primary pathway for activating G2/M checkpoint induction after re-replication. J Biol Chem 2007; 282:30357-62; PMID:17716975

47. Shibata A, Barton O, Noon AT, Dahm K, Deckbar D, Goodarzi AA, Löbrich M, Jeggo PA. Role of ATM and the damage response mediator proteins 53BP1 and MDC1 in the maintenance of G(2)/M checkpoint arrest. Mol Cell Biol 2010; 30:3371-83; PMID:20421415

48. Kampinga HH, Hiemstra YS, Konings AW, Dikomey E. Correlation between slowly repairable double-strand breaks and thermal radiosensitization in the human HeLa S3 cell line. Int J Radiat Biol 1997; 72:293-301; PMID:9298109

49. Posakony JW, England JM, Attardi G. Mitochondrial growth and division during the cell cycle in HeLa cells. J Cell Bol 1977; 74:468-91

50. Urlaub G, Chasin LA. Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. Proc Natl Acad Sci U S A 1980; 77:4216-20; PMID:6933469