REJOINING OF DNA DOUBLE-STRAND BREAKS AND GENOME STABILITY: FROM HOST-PATHOGEN INTERACTIONS TO BREAK-INDUCED MUTAGENESIS

by

Zhuobin Liang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Molecular, Cellular, and Developmental Biology)
in the University of Michigan
2015

Doctoral Committee:

Associate Professor Thomas E. Wilson, Co-Chair
Associate Professor Lyle Simmons, Co-Chair
Associate Professor Anuj Kumar
Professor Jianming Li
Associate Professor JoAnn Sekiguchi

DEDICATION

To my beloved wife, parents, parents-in-law and brother

ACKNOWLEDGMENTS

The completion of my dissertation research would not be possible without help and support from a number of people. First, I would very much like to thank my graduate mentors Dr. Thomas Wilson and Dr. Tzvi Tzfira for the opportunities and patient guidance they provided. They instilled in me the passion and spirit of science and taught me valuable techniques for becoming an independent investigator. My sincere thanks also go to great lab colleagues whom I closely collaborated with, especially Dr. Kishore Chiruvella, Siva Nallasivam, Dr. Yoel Shiboleth and Vardit Zeevi.

Next, I am extremely grateful to the members of my thesis committee, Dr. Lyle Simmons, Dr. Anuj Kumar, Dr. Jianming Li and Dr. JoAnn Sekiguchi, for their constant support and valuable suggestions. I also greatly benefited from discussions with a number of labs including but not limited to the Mats Ljungman, Xiaochun Yu, Thomas Glover, David Ferguson, Christine Canman and Patrick O'Brien labs.

I also want to express my thanks to the departments of MCDB, Pathology, Human Genetics, Rackham Graduate School and International Center in the University of Michigan for their assistance throughout my graduate study.

Finally I want to offer my deepest gratitude to my many friends and family members for their companion and encouragement.

PREFACE

This dissertation consists of the research projects I conducted in the former lab of Dr. Tzvi Tzfira in the University of Michigan from Sept 2008 to April 2011 and in Dr. Thomas E Wilson lab from May 2011 to present. The overall objective is to investigate the complex contribution of rejoining of DNA double-strand breaks to genome stability

Parts of Chapter 1 were modified from an invited book chapter of *Repair, Mutagenesis and Other Responses to DNA Damage* published in *Cold Spring Harb Perspect Biol* 2013 May;5(5): a012757 (co-first author), and from another invited review in preparation (first author).

All the contents in Chapter 2 have been published in *Nat Commun* 4:2253 doi: 10.1038/ncomms3253 (first author).

All the contents in Chapter 3 have been published in *Plant Physiol* 2012 Jan;158(1): 132-144 (co-first author), and in *Nat Biotechnol* 2011 Dec;29(12): 1072-4 (middle author).

All the contents in Chapter 4 have been published in *PLoS Genet* 2013 June;9(6): e1003599 (second author).

Chapter 5 is currently in preparation and will submit shortly (first author).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

alt-NHEJ: alternative nonhomologous end joining

c-NHEJ: canonical nonhomologous end joining

CTD: C-terminal domain

DSB: double-strand break

ECFP: enhanced cyan fluorescent protein

EYFP: enhanced yellow fluorescent protein

GUS: β-glucuronidase

HOcs: HO cut site

HR: homologous recombination

LM-qPCR: ligation mediated quantitative PCR

MCS: multiple cloning site

MMEJ: microhomology mediated end joining

NGS: next-generation sequencing

NHEJ: nonhomologous end joining

NLS: nuclear localization signal

RFP: red fluorescent protei

T-DNA: transfer DNA

ZFN: zinc finger nuclease

ZFNcs: zinc finger nuclease cut site

ABSTRACT

An essential task for cell survival is the maintenance of genome stability despite various environmental and physiological stresses. These stresses, such as UV light and ionizing radiation, often damage DNA and result in various types of DNA lesions. If not properly resolved, DNA lesions in humans can cause autoimmune deficiency, neurodegenerative disorders and cancer. Among these lesions, DNA double-strand breaks (DSBs) are one of the most cytotoxic and greatly threaten the integrity of the genome. In addition, DSBs may act as potential hotspots for genomic integration of exogenous DNA fragments, such as the transfer DNA (T-DNA) delivered by the plant pathogen *Agrobacterium*. DSBs can be repaired by religation of two broken ends by DNA ligase IV via the nonhomologous end joining (NHEJ) pathway, of which the repair fidelity can be compromised by diverse break structures, resulting in mutagenesis.

I sought to further understand the complex contribution of NHEJ to genome stability in research projects conducted using both *Agrobacterium*-mediated plant transformation and *Saccharomyces cerevisiae* as model systems. My first project examined the process of double-stranded T-DNA formation using functional assays and demonstrated that annealing of synthetic oligonucleotides to the single-stranded T-DNA can initiate such process in plant cells. A second project developed an *Agrobacterium*-mediated transformation vector system, in which multiple expression cassettes can be assembled on a single vector using zinc finger nucleases (ZFNs) and

homing endonucleases, facilitating delivery of multiple genes in plants. A third project investigated the consequences of having a catalytically inactive DNA ligase IV in yeast NHEJ and led to discovery of an imprecise NHEJ pathway mediated by ligase Cdc9. A last project studied the impact of overhang polarity of chromosomal DSBs on the kinetics and fidelity of yeast NHEJ and results suggest that 5' overhanging DSBs can cause more frequent mutagenesis despite more efficient rejoining as compared to 3' overhanging DSBs.

Collectively, my dissertation research provides new evidence of the mechanisms governing the important process of double-stranded T-DNA formation during plant genetic transformation, as well as new insights into NHEJ mutagenesis which could lead to different human diseases.

CHAPTER 1

Introduction

Repair of DNA double-strand breaks by end-joining

DNA double-strand breaks (DSBs) are serious lesions that threaten a loss of chromosomal content. Repair of DSBs is particularly challenging because, unlike all other lesions, the DNA substrate is inherently bimolecular. Bringing two DNA molecules together is also dangerous since local mutations and chromosome rearrangements can arise if ends are inappropriately coupled.

The cell has two general strategies for repairing DSBs. Homologous recombination (HR) refers to mechanisms in which an intact homologous donor duplex is used to guide DNA synthesis across the DSB gap. Nonhomologous end joining (NHEJ) is defined as repair in which two DSB ends are joined by direct ligation. The resulting joints are characterized by little (less than about 10 bp) or no homology between the joined ends that could have been used to guide repair, which, when it exists, is referred to as "microhomology". NHEJ is thus recognized as having a high potential for error. The term NHEJ was for a long time used to refer to a specific DNA repair pathway characterized by its dependence on Ku, DNA ligase IV and associated factors (**Table 1-1**), referred to here as "canonical NHEJ" (c-NHEJ). However, some homology-independent repair occurs in the absence of these proteins, variably referred to as "backup NHEJ", "microhomology-mediated end joining" (MMEJ) or "alternative NHEJ" (alt-NHEJ, the main terminology used here).

1.1 Canonical NHEJ core proteins

Ku and DNA-PKcs. The prototypical c-NHEJ protein is Ku, a heterodimer of two related proteins Ku70 and Ku80. Ku is a DSB-specific end-binding protein by virtue of the oriented threading of a DNA end into a hole in the protein dimer that allows its further translocation onto the DNA duplex (**Figure 1-1**) (Walker *et al.*, 2001). This configuration means that degradation of Ku is required for its removal from DNA following, and possibly during, repair, which is promoted by the ubiquitylation of Ku80 by the E3 ligase RNF8 (Postow *et al.*, 2008; Feng and Chen, 2012).

Ku is one part of the DNA-dependent protein kinase (DNA-PK). The DNA-PK catalytic subunit (DNA-PKcs), related to the Ataxia-Telangiectasia Mutated (ATM) checkpoint kinase (Lempiainen and Halazonetis, 2009), is also a core c-NHEJ protein, although there are situations when the requirement for DNA-PKcs is less stringent than for Ku (Gapud and Sleckman, 2011; Gu *et al.*, 2000). For example, DNA-PKcs deficiency does not retard growth as in Ku-deficient mice (Gu *et al.*, 1997). When activated, DNA-PKcs extensively phosphorylates itself, other c-NHEJ proteins, and other targets (Douglas *et al.*, 2005; Wang *et al.*, 2004; Yu *et al.*, 2008).

DNA ligase IV. Lig4 is the DNA ligase required for, and specific to, c-NHEJ. It catalyzes the same ATP-dependent transfer of phosphate bonds that results in strand ligation in all eukaryotic DNA repair (Ellenberger and Tomkinson, 2008). However, unusual biochemical properties of Lig4 modify this core reaction in ways important to c-NHEJ. Lig4 was the only ligase with the mechanistic flexibility to ligate one strand independently of another (Ma *et al.*, 2004) or incompatible DSB ends as well gaps of several nucleotides (Gu *et al.*, 2007), properties consistent with the joining of the wide variety of DSB structures relevant to c-NHEJ *in vivo*.

Structurally, little is known about Lig4, including how mechanistic flexibility might be realized at the protein level. One possibility is that the Lig4 catalytic domain is structurally as well as functionally flexible (Perry *et al.*, 2010), although SAXS experiments suggest it may in fact have an extended shape with limited flexibility (Ochi *et al.*, 2012). Lig4 shares a catalytic domain structure with DNA ligase I (Lig1) that allows the latter protein to completely encircle DNA during catalysis (**Figure 1-2A**) (Pascal *et al.*, 2004). The simplest assumption is that this architecture will be recapitulated in Lig4, an idea supported by the similar folding of the isolated Lig4 adenylation domain compared to Lig1 (**Figure 1-2A**) (Ochi *et al.*, 2012).

XRCC4 and XLF. XRCC4 is a non-enzymatic Lig4 partner protein, with the two co-purifying as a complex. XRCC4 has a homodimeric structure with paired globular head domains, an elongated coiled-coil (**Figure 1-2B**) and a structurally ill-defined and less conserved CTD that nonetheless has strong influences on c-NHEJ *in vivo* (Palmbos *et al.*, 2005; Koch *et al.*, 2004). The Lig4-XRCC4 interaction is mediated by tandem BRCA1 C-terminal (BRCT) repeats at the Lig4 C-terminus, and especially the inter-BRCT linker, that intimately wrap around the XRCC4 coiled-coil in a clamp-like fashion (**Figure 1-2B**) (Wu *et al.*, 2009; Dore *et al.*, 2006). The catalytic and XRCC4-binding domains account for nearly all of Lig4 except for a small region between them that carries a basic patch implicated in DNA binding (Hammel *et al.*, 2011).

XRCC4 and XLF together form long super-helical filaments (**Figure 1-2D**) (Andres and Junop, 2011; Wu *et al.*, 2011; Ropars *et al.*, 2011; Hammel *et al.*, 2010). Further contacts support interactions between parallel filaments (**Figure 1-2E**) (Hammel *et al.*, 2011). Binding of the Lig4 BRCT region likely influences the nature and extent of these various modes of XRCC4-XLF multimerization, in part by preventing the XRCC4 C-terminus (not present in **Figure 1-2D**) from

interacting with the XRCC4-XLF interface (Hammel *et al.*, 2010; Ochi *et al.*, 2012)**.** The importance of XRCC4-XLF higher order structures appears to be DNA binding, which might include co-linear protein and DNA filaments as well as a channel whose base is the XRCC4/XLF head domains and sides are the coiled-coil stalks, with DNA binding perpendicularly to the XRCC4-XLF filament (**Figure 1-2E**) (Hammel *et al.*, 2011; Andres *et al.*, 2012).

Mre11-Rad50-Nbs1. Unlike Ku and Lig4, the MRN complex of proteins formed by Mre11, Rad50 and Nbs1 (**Figure 1-3**) is involved in most aspects of DSB repair, including ATM-dependent checkpoint signaling. MRN was first identified as a c-NHEJ factor in budding yeast, where it is as required as Ku or Lig4 (Milne *et al.*, 1996). In contrast, MRN is not required for NHEJ in fission yeast (Manolis *et al.*, 2001) and early efforts gave conflicting but often non-supportive observations regarding a role for MRN in vertebrate NHEJ (Di Virgilio and Gautier, 2005; Huang and Dynan, 2002; Yamaguchi-Iwai *et al.*, 1999). However, a series of more recent studies using refined genetic tools have established a less penetrant but consistent contribution of MRN to some mammalian NHEJ (Dinkelmann *et al.*, 2009; Deriano *et al.*, 2009; Xie *et al.*, 2009; Rass *et al.*, 2009; Helmink *et al.*, 2009). Conditional Mre11 loss in mouse B lymphocytes caused NHEJ deficiencies during immunoglobulin class switch recombination that could not be explained by impairment of ATM activation (Dinkelmann *et al.*, 2009). In a distinct approach, siRNA-mediated depletion of Mre11 reduced end joining in a reporter assay in both $XRCC4^{+/+}$ and $XRCC4^{-/-}$ cells, suggesting MRN roles in both c-NHEJ and alt-NHEJ (Xie *et al.*, 2009). Such studies support a checkpoint-independent NHEJ role of MRN that might include structural stabilization of DSBs and/or end processing.

DNA polymerases. The yeast PolX family DNA polymerase, Pol4, is absolutely required for c-NHEJ events in which gaps must be filled on both strands of 3' overhanging DSBs (Daley *et al.*, 2005). This pattern demonstrates the critical importance of DSB configuration, since the template for extending a 3' overhang is the other side of the DSB, requiring a polymerase that can engage a disrupted template (**Figure 1-4C**). Structural studies of mammalian Polλ and Polμ, themselves also c-NHEJ enzymes, provide insight into this function (Garcia-Diaz *et al.*, 2005; Moon *et al.*, 2007; Nick McElhinny *et al.*, 2005; Ramsden and Asagoshi, 2012). The catalytic domains of these PolX polymerases are bipartite with binding pockets for the 5' and 3' termini of the broken strand whose gap is being filled (**Figure 1-4A**). Loop 1 of the enzymes can be seen positioned near the point on the template strand that might be broken in a DSB (**Figure 1-4A**), the same loop that replaces the template strand in terminal deoxynucleotidyl transferase (TdT) (Delarue *et al.*, 2002). Biochemical analyses support the idea that Loop1 and other PolX residues promote catalysis by intrinsically stabilizing the weak association of primer terminus to a disrupted template (Juarez *et al.*, 2006; Bebenek *et al.*, 2010; Moon *et al.*, 2007; Martin *et al.*, 2012). 5' overhanging DSBs in yeast do not require Pol4 (Daley *et al.*, 2005). Polλ and Polμ are similarly required for only subsets of c-NHEJ *in vivo* implying that other polymerases must act (Capp *et al.*, 2006; Capp *et al.*, 2007). Yeast replicative polymerases Pol2 and Pol3 have been suggested to participate (Tseng *et al.*, 2008; Chan *et al.*, 2008) and c-NHEJ translesion synthesis has been observed (Covo *et al.*, 2009), although the role of bypass polymerases is not well explored.

Nucleases and related activities. The Mre11 nuclease does not have clear roles in trimming overhangs in yeast where the most precise correlation of DSB and joint structures can

be performed (Wilson TE, unpublished observations). However, it has been suggested to support microhomology pairing (Zhang and Paull, 2005; Williams *et al.*, 2008) and the mammalian Mre11 nuclease does promote DSB end processing that leads to nucleotide deletions (**Figure 1-5**) (Xie *et al.*, 2009). Mre11 nuclease deficiency confers a milder NHEJ defect than loss of MRN during class switch recombination, although the exact mode of end processing in the nuclease-dependent events is not known (Dinkelmann *et al.*, 2009).

Another surprising development is that some cleavage of DSB termini might be executed to promote joining fidelity rather than the necessity of ligation. The tyrosyl-DNA phosphodiesterase (Tdp1) is a general 3' phosphoesterase capable of removing lesions such as 3' phosphoglycolates (Zhou *et al.*, 2009). Tdp1 leaves a terminal 3' phosphate that must be removed for repair to continue. An interesting model developed in yeast proposes that Tdp1 removes intact bases from otherwise undamaged 5' overhanging DSBs such that the new 3' phosphate lesion prevents undesirable filling of the overhang and insertional mutagenesis (**Figure 1-4E**) (Bahmed *et al.*, 2010).

1.2 Alternative NHEJ

In every system studied there is a residual amount of NHEJ observed when c-NHEJ is disabled, referred to as alt-NHEJ. Of interest is the increased propensity of alt-NHEJ to create mutations, since joints often harbor local deletions with relatively long stretches of microhomology, itself often called MMEJ. However, it is more meaningful to characterize NHEJ pathways by enzymology than by joint structure since the DNA outcomes of different pathways can be the same (**Figure 1-5**). Limited precise religation of overhangs is observed even in the

absence of Lig4, which is influenced by features such as overhang length in a manner that suggests an equilibrium between DSB and ligatable SSB states (Daley and Wilson, 2005). Similarly, MMEJ events can arise by both c-NHEJ and alt-NHEJ, albeit to a greater extent with alt-NHEJ.

The enzymatic mechanisms of alt-NHEJ are less well defined than for c-NHEJ. It is not even clear how many distinct mechanisms alt-NHEJ encompasses. Like c-NHEJ, alt-NHEJ demands DSB synapsis. PARP-1 has been implicated in this function in alt-NHEJ, reminiscent of the role of DNA-PKcs in c-NHEJ (Mansour *et al.*, 2010; Audebert *et al.*, 2008; Robert *et al.*, 2009). Not only is alt-NHEJ independent of DNA-PK it is suppressed by Ku, implying a competition of factors for DSB ends that is typically won by c-NHEJ (Audebert *et al.*, 2004; Wang *et al.*, 2006). In addition to its possible synapsis function, PARP-1 may also serve as a platform for directly or indirectly recruiting alt-NHEJ repair factors (Table 1-1) (Audebert *et al.*, 2004; Audebert *et al.*, 2006; Sallmyr *et al.*, 2008; Della-Maria *et al.*, 2011). PARP-1 action is pleiotropic, though, and has sometimes been seen to support NHEJ, including that mammalian SIRT6 can stimulate PARP-1 resulting in promotion of both NHEJ and HR (Mao *et al.*, 2011).

As an obligatorily Lig4-independent pathway, other enzyme(s) must catalyze strand ligation in alt-NHEJ. Studies in mammalian cells have implicated DNA Ligase III (Lig3) as the major alt-NHEJ ligase (Wang *et al.*, 2005; Audebert *et al.*, 2004; Sallmyr *et al.*, 2008; Simsek *et al.*, 2011; Chiruvella *et al.*, 2012). As a co-factor of Lig3, XRCC1 is also suggested to be involved (Audebert *et al.*, 2004; Saribasak *et al.*, 2011), although more recent studies suggest that XRCC1 may be dispensable (Boboila *et al.*, 2012; Han *et al.*, 2012). A possible contribution to such discrepancies is that Lig1 (the only other mammalian ligase) has also been shown to

support some level of alt-NHEJ (Simsek *et al.*, 2011; Liang *et al.*, 2008). It is unknown whether Lig3 and Lig1 are simply redundant or if their usage is somehow regulated.

Microhomology is an important feature of many alt-NHEJ joints that enhances the stability of PARP-1-mediated DNA synapsis (Audebert *et al.*, 2008). One mechanism creates this base-pairing potential by locally templated extensions of the 3' DSB strand (Yu and McVey, 2010; Simsek *et al.*, 2011), but more generally internal microhomologies are exposed by resection of DSB ends, evident as deletions in final joints. Alt-NHEJ likely uses limited resection based on the size of these deletions, consistent with two-step models of DSB resection that transition from an initial local resection to a faster and more processive extended resection to support fully efficient HR (**Figure 1-5**) (Symington and Gautier, 2011; Grabarz *et al.*, 2012). Thus, alt-NHEJ and HR appear to share a common initial resection mechanism promoted by the Mre11 nuclease and CtIP (Dinkelmann *et al.*, 2009; Rass *et al.*, 2009; Xie *et al.*, 2009; Lee-Theilen *et al.*, 2011; Zhang and Jasin, 2011).

1.3 Disposition of DSBs between c-NHEJ and HR

c-NHEJ use relative to HR must be regulated, a relationship often seen as a competition. However, loss of most HR proteins does not increase c-NHEJ efficiency (Karathanasis and Wilson, 2002). The exception is mutants that abolish the redundant mechanisms that resect the 5' terminated strand to create a substrate for Rad51 binding in HR, which do increase c-NHEJ yield (Ira *et al.*, 2004; Symington and Gautier, 2011). Loss of Ku in turn allows more rapid resection while overexpression delays the onset of resection (Clerici *et al.*, 2008). These observations emphasize that c-NHEJ and HR are sequential reactions, with c-NHEJ becoming impossible

once DSB resection begins and HR being necessary only if c-NHEJ fails (**Figure 1-5**). Disposition of DSBs between pathways is therefore determined by the rate of c-NHEJ relative to the onset of 5' resection. This fact leads to the concept that c-NHEJ and HR cooperate rather than compete to drive the greatest likelihood of successful repair.

HR works best in late S and G2 when the sister chromatid is available for repair and c-NHEJ-incompatible one-ended DSBs arise at collapsed replication forks. Consistently, 5' resection is under cell cycle control, with increased cyclin-dependent kinase (CDK) activity in S phase leading to potentiation of resection (**Figure 1-5**) by mechanisms that include phosphorylation of pro-resection proteins such as CtIP (Ira *et al.*, 2004; Aylon and Kupiec, 2005; Huertas and Jackson, 2009). Importantly, c-NHEJ is still possible in S/G2 and remains a predominant repair pathway for DSBs in mammalian G2 (Beucher *et al.*, 2009; Karanam *et al.*, 2012). Regulation simply dictates the window of opportunity afforded to c-NHEJ, effectively placing it on a cell-cycle dependent timer.

Notably, normal Ku levels inhibit resection in G1, when CDK activity is low, but not S/G2 (Clerici *et al.*, 2008). This may reveal in part a less potent binding of Ku in S/G2 (Zhang *et al.*, 2009). Ku disappears from unrepaired DSBs in a manner that depends on MRN (Wu *et al.*, 2008). This disappearance may result from an unknown active Ku removal mechanism that is necessary for resection to begin (Zhang *et al.*, 2009; Shao *et al.*, 2012; Langerak *et al.*, 2011) but could also be a simple consequence of MRN-dependent resection, where the dynamic on-off rate of Ku might be sufficient to allow the entry of activated resection enzymes in S/G2 (**Figure 1-5**) (Mari *et al.*, 2006).

T-DNA integration into genomic double-strand breaks

Genetic modification and improvement of crop plants is fundamental for modern agriculture. *Agrobacterium*-mediated genetic transformation, a process by which the bacterium's transferred DNA (T-DNA) molecule randomly integrates into the plant genome, is one of the most commonly used tools for genetic modification of plants. However, many gaps in our knowledge have limited the full potential utilization of this tool with respect to controlling and optimizing the integration process. More specifically, our ability to target the T-DNA to specific genomic locations, routinely produce crop plants with defined T-DNA insertions and with controlled and reproducible transgene expression patterns, is still greatly limited. Fundamental scientific questions exist regarding the mechanisms of T-DNA processing in plant cells as well as genome integration that might be exploited to improve targeted genetic modification.

1.4 Gene targeting in plants

To translate the enormous amount of data obtained from the various large genome-sequencing projects (Shultz *et al.*, 2006; Yu *et al.*, 2002; Yuan *et al.*, 2003; Yuan *et al.*, 2005; Arabidopsis Genome, 2000) into new knowledge and, potentially, new agricultural traits, reliable and efficient methods are needed for the precise expression and modification of genome sequences in plant cells. In the past few decades, we have witnessed significant progress in the development of transgenic technologies, and a constantly increasing number of plant and crop species are now transformable via *Agrobacterium* and/or other methods (Lorence and Verpoorte, 2004) (Taylor and Fauquet, 2002). This progress is reflected in the

growing number of transgenic crops now sold in markets worldwide. Despite this impact, there has been almost no improvement in controlling the quality of transformation events. Thus, for example, the integration of new DNA molecules, often delivered as T-DNA molecules by *Agrobacterium*-mediated genetic transformation, typically occurs randomly into different genomic locations. This leads to several complications, including positional effects (Matzke and Matzke, 1998), poor stability of the transgene's expression over several generations (Stewart *et al.*, 2000), and downregulation of non-target genes (Malik *et al.*, 2006). Thus, the development of novel methods for site-specific DNA integration and gene targeting are highly desirable as they will allow more precise, predictable and stable expression (or silencing) of transgenes and thus full exploitation of the "transgenic revolution". Indeed, gene targeting (often achieved by homologous recombination (HR)) has had far-reaching implications in biology. It is a powerful technique in prokaryotes, lower eukaryotes, and some higher eukaryotes including the moss Physcomitrella (Schaefer and Zryd, 1997), Drosophila (Rong and Golic, 2000), chicken (Bezzubova *et al.*, 1997) and mouse (Capecchi, 1989). In plants, however, the ability to modify the genome in a precise manner is still underdeveloped because low frequencies of HR have prevented gene targeting from becoming a practical and routine technique (Hanin and Paszkowski, 2003; Weinthal *et al.*, 2010).

1.5 Voyage of *Agrobacterium* T-DNA in the host cell

The most common vector used today for the genetic transformation of various model and crop plants is *Agrobacterium tumefaciens* (Banta and Montenegro, 2008). The natural host

range of *Agrobacterium* is limited to certain dicotyledonous plants, but under controlled laboratory conditions it can transform not only an ever-increasing number of plant species (Banta and Montenegro, 2008), but also eukaryotes from other families, ranging from lower eukaryotes such as yeast (Bundock *et al.*, 1995) (Piers *et al.*, 1996) and other fungi (de Groot *et al.*, 1998; Godio *et al.*, 2004; Michielse *et al.*, 2005) to higher eukaryotes, such as human cells (Kunik *et al.*, 2001). This bacterium is capable of transforming its host by delivering a well-defined fraction of its own genome (designated transferred DNA or T-DNA) as a single-stranded (ss) DNA molecule (designated T-strand) into the host cell where numerous bacterial and host proteins are involved for the transportation of T-DNA into the cell nucleus (**Figure 1-6** and **Table 1-2**).

1.6 DNA repair and T-DNA integration

After entry of the host nucleus, the T-DNA can integrate into the host genome by mechanism(s) that are still largely unknown, but two major models exist. One model suggests that the T-strand can integrate into the host genome without the extensive uncoating of the proteins bound to it, such as *Agrobacterium* VirE2 and various plant factors, while the other model emphasizes that T-strand is rapidly converted into double stranded form which integrates using HR or NHEJ-mediated integration machinery **(Figure 1-7).**

The T-DNA molecule does not code for the machinery needed for its integration into the plant genome. Thus the T-DNA sequence can potentially be replaced by any other sequence of interest without affecting the transformation and integration process, which is likely to be

governed largely by plant proteins (Mysore *et al.*, 2000; Friesner and Britt, 2003; Li *et al.*, 2005; Endo *et al.*, 2006; Anand *et al.*, 2007; Lacroix *et al.*, 2008) **(Table 1-2)**. Genome structure and transcription activity have also been suggested to affect T-DNA integration (Brunaud *et al.*, 2002; Alonso *et al.*, 2003; Tzfira *et al.*, 2003; Schneeberger *et al.*, 2005; Li *et al.*, 2006; Gelvin and Kim, 2007; Kim *et al.*, 2007). More importantly, T-DNA molecules can be captured by genomic DSBs (Chilton and Que, 2003; Tzfira *et al.*, 2003; Salomon and Puchta, 1998), which led to the suggestion that DSBs may act as 'hotspots' for T-DNA integration (**Figure 1-7**) (Ziemienowicz *et al.*, 2008; Tzfira *et al.*, 2004). Recent data from our lab suggest that double-stranded (ds) intermediates (dsT-DNAs) may be the dominant substrate for the NHEJ-mediated integration process (**Figure 1-8**) (Liang and Tzfira, 2013; Dafny-Yelin *et al.*, 2015). While it is possible that single stranded T-DNA (i.e. T-strand) molecules serve as substrate for the HR-mediated integration machinery (**Figure 1-9**) (Tinland, 1996; Mayerhofer *et al.*, 1991; Gheysen *et al.*, 1991), recent data suggest that double-stranded (ds) intermediates (dsT-DNAs) may be the dominant substrate in the NHEJ-mediated integration process (Liang and Tzfira, 2013; Singer *et al.*, 2012; Dafny-Yelin *et al.*, 2015). Furthermore, dsT-DNAs are an important substrate for transient gene expression studies (Sheludko, 2008) and recombination-based gene targeting (Wang *et al.*, 2011) in plant cells. The mechanism by which T-strands are complemented to dsT-DNAs is still unknown, but data obtained from our current showed that short DNA or RNA fragments may be required to initiate this process (Liang and Tzfira, 2013; Dafny-Yelin *et al.*, 2015).

Genetic modification of plants by T-DNA is an example of a broad class of genomic structural alterations that occur in normal biology and disease, of which other examples include

mating type switching in yeast (Haber, 1998), V(D)J recombination in the vertebrate immune system (Gellert, 2002) and copy number variants in all organisms (Alkan *et al.*, 2011; Freeman *et al.*, 2006). T-DNA integration is subject to similar genetic controls as these other recombination events, most importantly that the primary substrate for genetic recombination is the DSB (Haber, 2000; Szostak *et al.*, 1983). Thus, both NHEJ and HR are capable of joining DSB ends to create novel products, the basis of the hypothesis that T-DNA integration occurs via recombinogenic repair at random spontaneous genomic DSBs (Ziemienowicz *et al.*, 2008; Tzfira *et al.*, 2004). Moreover, dsT-DNA itself has two DSB ends and is likely processed as DNA damage by the cell. With two distinct pathways of DSB repair it is important to understand which mechanism acts in different situations. Current evidence indicates that cell cycle stage is the strongest determinant of whether NHEJ or HR is used for repair of spontaneous chromosomal DSBs (Ira *et al.*, 2004; Shrivastav *et al.*, 2008). In late S and G2, HR is activated because a sister chromatid is available to assist in repair. In contrast, in G1 or quiescent G0 cells HR is down-regulated since a repair donor is unavailable with NHEJ providing the most accurate repair. Perceived differences in the relative importance of the two pathways in different organisms are often explained by this model. For example, budding yeast show efficient HR because they spend a large fraction of the cell cycle in S/G2. In plants, like human cells, NHEJ predominates in most non-dividing cells, consistent with evidence that random T-DNA integration in plants occurs mainly by NHEJ (Ziemienowicz *et al.*, 2008; Tzfira *et al.*, 2004; Puchta, 2005; Friesner and Britt, 2003).

Outline of subsequent chapters

I will present recent findings of the process by which *Agrobacterium* T-DNA is converted to the double-stranded form in plant cells prior to its genomic integration in **Chapter 2**. Then I will describe advancements of our developed techniques to deliver multiple genes and their application to production of anti-malaria drugs in transgenic tobacco in **Chapter 3**. After that, I will describe catalytically inactive DNA ligase IV in promoting imprecise rejoining in yeast in **Chapter 4**. Then I will report new evidence that overhang polarity of chromosomal DSBs can impact repair kinetics and fidelity of yeast NHEJ in **Chapter 5**. Finally, I will summarize the conclusions of my thesis research and discuss further directions in **Chapter 6**. The outcomes of this work provide important new insights into the relationship between NHEJ and genome stability, as well as new experimental techniques for further investigation in the field.

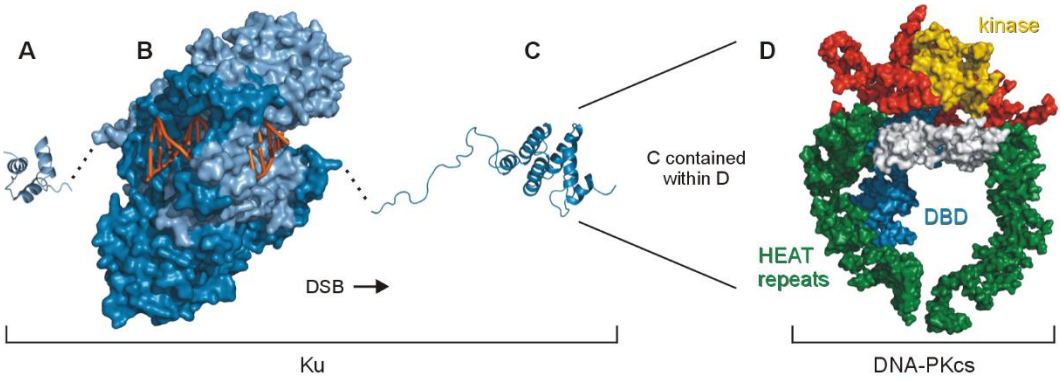


Figure 1-1. Ku and DNA-PKcs. (A) Structural representation of the CTD of human Ku70 (PDB 1JJR (Zhang *et al.*, 2001)). **(B)** The human Ku70-Ku80 heterodimer bound to DNA (PDB 1JEY (Walker *et al.*, 2001)). **(C)** The CTD of human Ku80 (PDB 1RW2 (Zhang *et al.*, 2004)). Connections between structures are indicated as dashed black lines. Ku70, light blue; Ku80, blue; DNA, orange. **(D)** Low resolution structure of human DNA-PKcs plus the Ku80 CTD (PDB 3KGV (Sibanda *et al.*, 2010)). Kinase domain, yellow; HEAT repeats, green; "brow", white; putative DBD, blue; remainder, red.

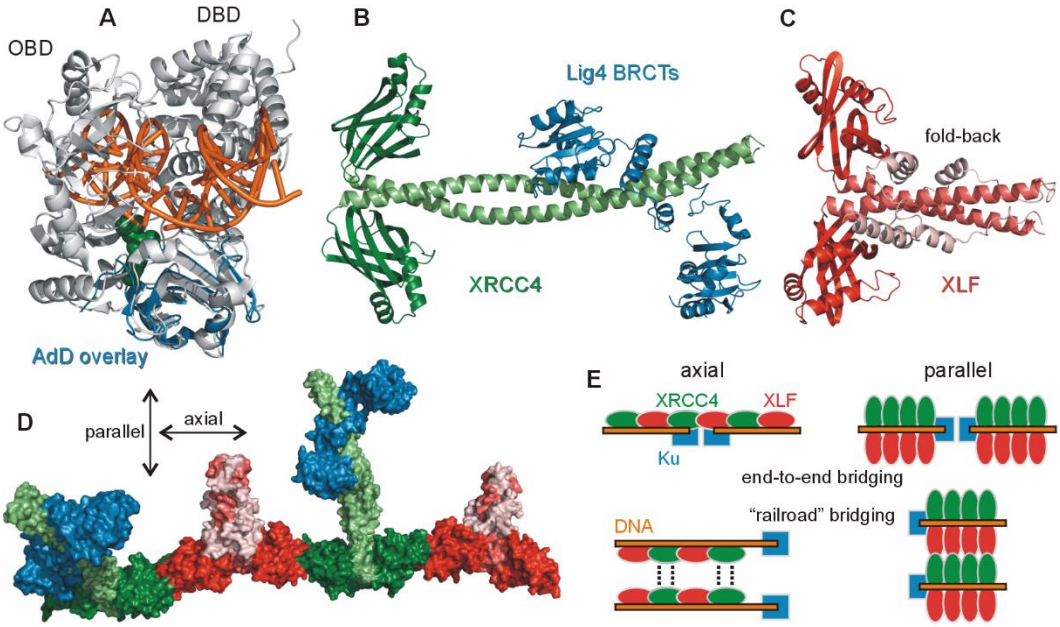


Figure 1-2. DNA ligase IV assembly. (A) The adenylation domain (AdD) of Lig4 (blue; PDB 3VNN (Ochi *et al.*, 2012)) is superimposed on a structural representation of Lig1 bound to a DNA nick (light grey; PDB 1X9N (Pascal *et al.*, 2004)) as a surrogate model of how Lig4 might bind DNA. OBD, oligonucleotide/oligosaccharide binding domain; 5' AMP, green; DNA, orange. **(B)** The human XRCC4 homodimer bound to the Lig4 tandem BRCT repeat region (PDB 3II6 (Wu *et al.*, 2009)). **(C)** Human XLF homodimer (PDB 2QM4 (Li *et al.*, 2008)). **(D)** Surface representation of the XRCC4-XLF axial filament with bound Lig4 BRCT region, created by superimposing PDB 3II6 onto PDB 3RWR (Andres and Junop, 2011). Lig4, blue; XRCC4, shades of green; XLF, shades of red. **(E)** Idealized models of DNA engagement and end bridging by XRCC4-XLF multimers, colored the same (D). "Axial" and "parallel" refer to the orientation of XRCC4-XLF interactions that drive the assembly.

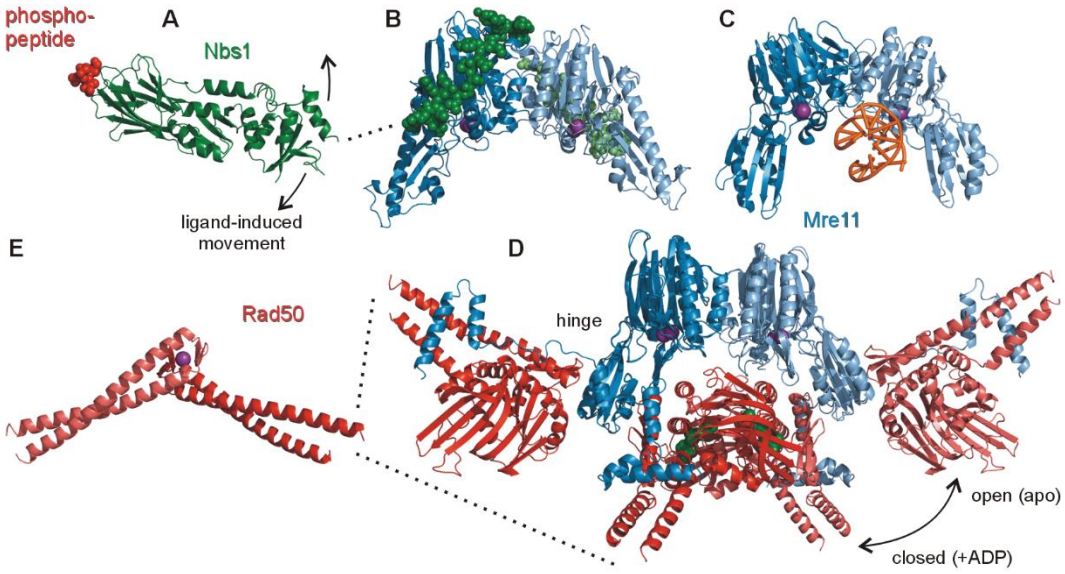


Figure 1-3. Mre11-Rad50-Nbs1 (MRN) complex. (A) Fission yeast Nbs1 FHA domain bound to a Ctp1 phosphopeptide (PDB 3HUF (Williams *et al.*, 2009)). **(B)** Fission yeast Mre11 globular domain bound to an Nbs1 internal peptide (PDB 4FBW (Schiller *et al.*, 2012)). **(C)** *Pyrococcus furiosus* Mre11 globular domain bound to DNA (PDB 3DSD (Williams *et al.*, 2008)). **(D)** Superimposed structures of the *Thermotoga maritima* Mre11 globular domain in the open (PDB 3QG5 (Lammens *et al.*, 2011)) and closed (PDB 3THO (Mockel *et al.*, 2012)) conformations showing the large Rad50 domain movement induced by adenosine nucleotide binding. **(E)** *Pyrococcus furiosus* Rad50 Zn hook motif (PDB 1L8D (Hopfner *et al.*, 2002)). Ctp1 peptide, red; Nbs1, green; Mre11, shades of blue; Rad50, shades of red; ADP, green; Zn and Mn ions, purple; DNA, orange. Connections between structures are indicated as dashed black lines.

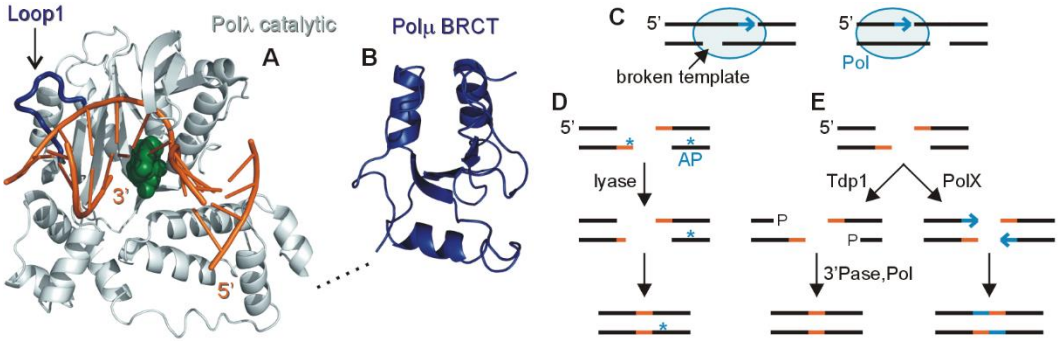


Figure 1-4. PolX polymerases. (A) Human Polλ catalytic domain bound to a 1-base gap and incoming nucleotide (light grey; PDB 1XSN (Garcia-Diaz *et al.*, 2005)). Loop 1, dark blue; ddTTP, green; DNA, orange. **(B)** Human Polμ BRCT domain (PDB 2DUN, RIKEN Structural Genomics/Proteomics Initiative, similar to PDB 2HTF (DeRose *et al.*, 2007)). Connection between structures is indicated as a dashed black line. **(C)** Line diagrams depicting the different requirements imposed on a DNA polymerase (blue) filling a DSB gap (blue arrow) at 3' (left panel) vs. 5' (right panel) overhangs with respect to placement of the template strand break. **(D)** dRP lyase activity as an example of end processing, illustrating different requirements for handling terminal vs. internal base damage. AP, abasic site. **(E)** The Tdp1 fidelity control mechanism that prevents PolX-dependent insertional mutagenesis by transiently cleaving and blocking termini with a 3' phosphate.

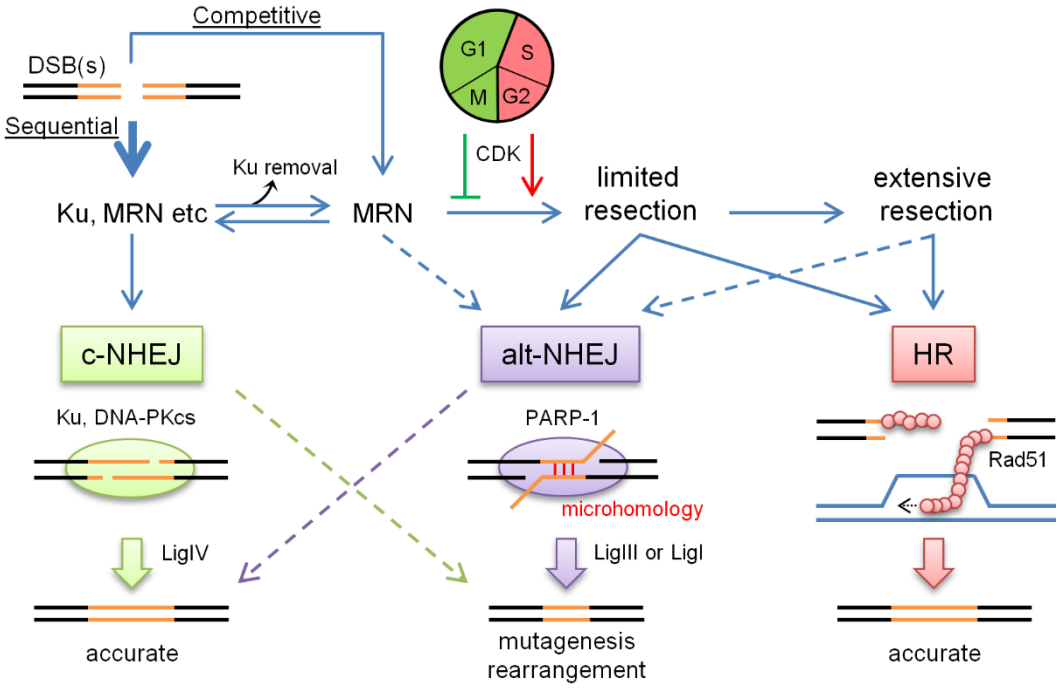


Figure 1-5. Disposition of DSBs between repair pathways. Diagram illustrating the relationships between c-NHEJ, alt-NHEJ and HR and the factors that influence the disposition of DSBs between these repair pathways. See text for further discussion.

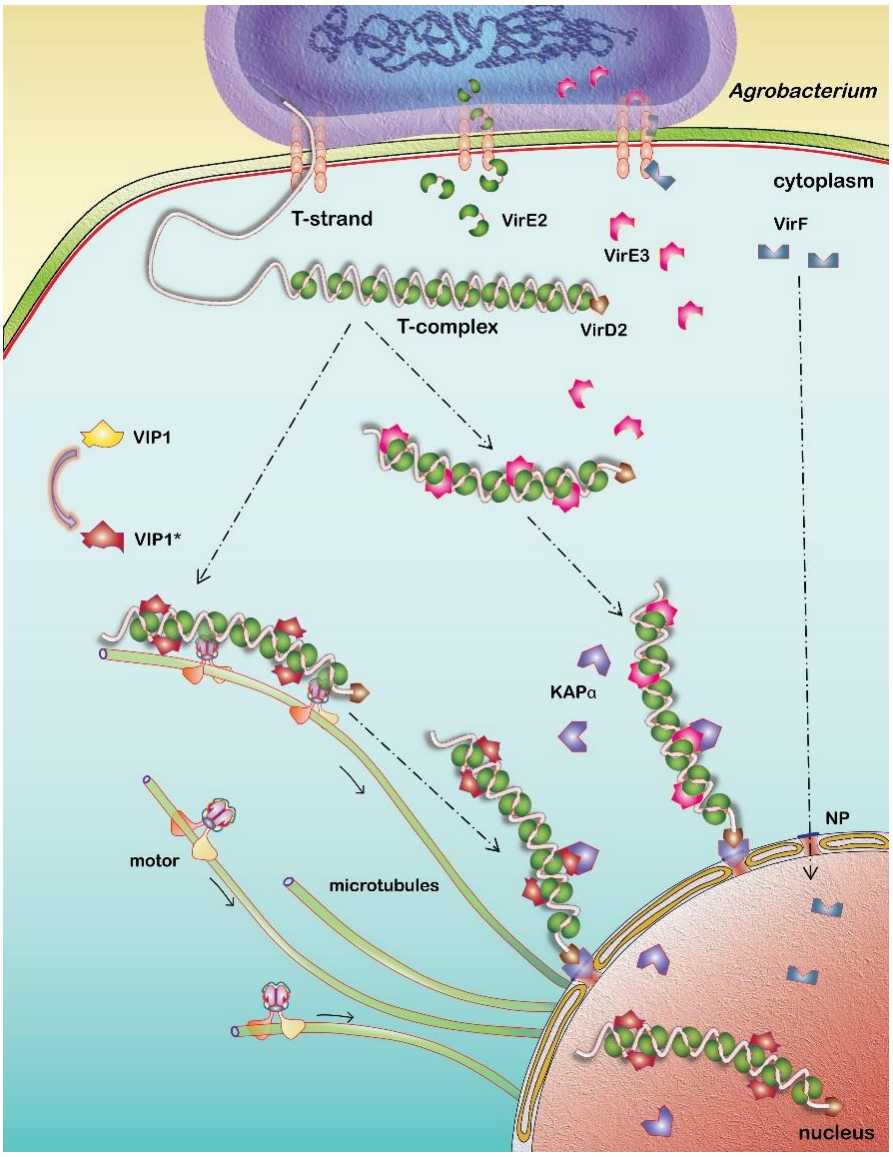


Figure 1-6. Transportation of the T-DNA from *Agrobacterium* to the host nucleus. More information of the plant factor involved in this process is listed in **Table 1-2**.

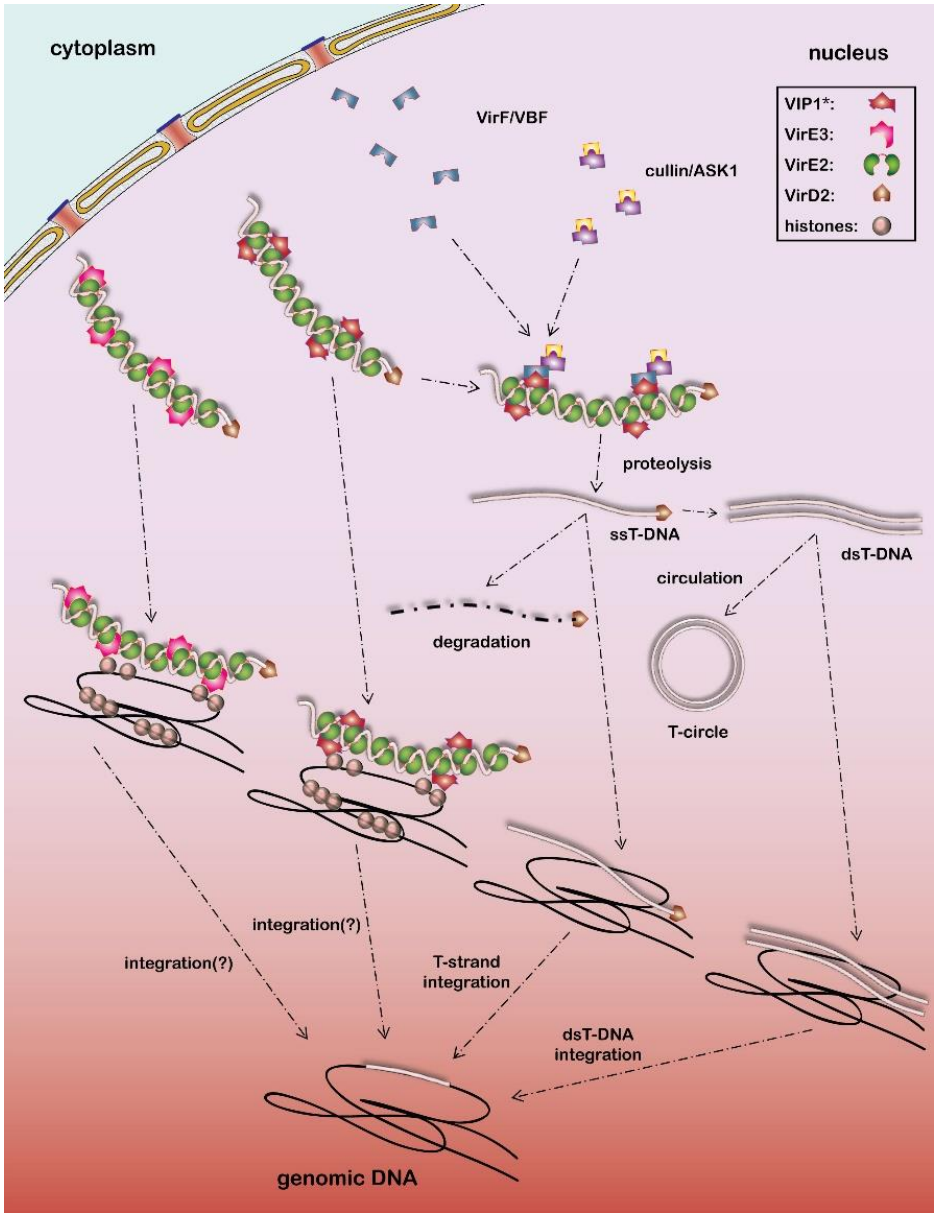


Figure 1-7. Hypothetical models of T-DNA integration to the host genome.

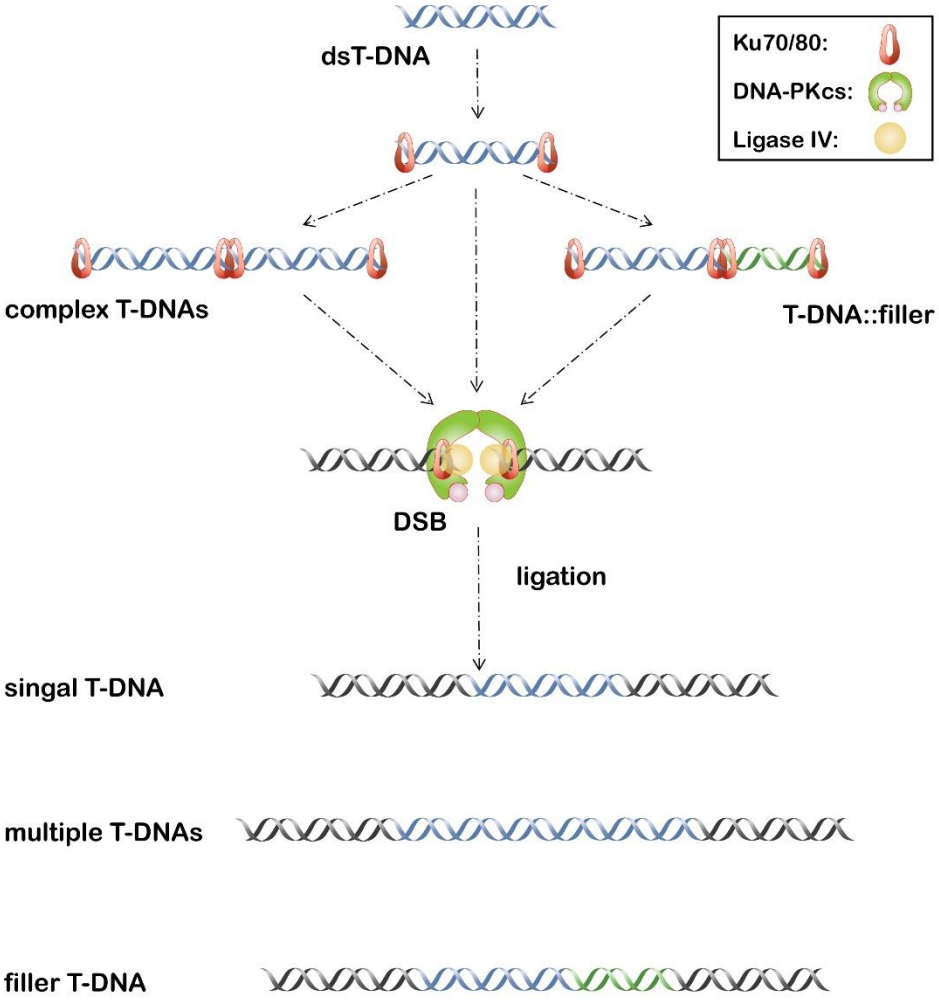


Figure 1-8. The NHEJ-mediated integration machinery for T-DNA integration.

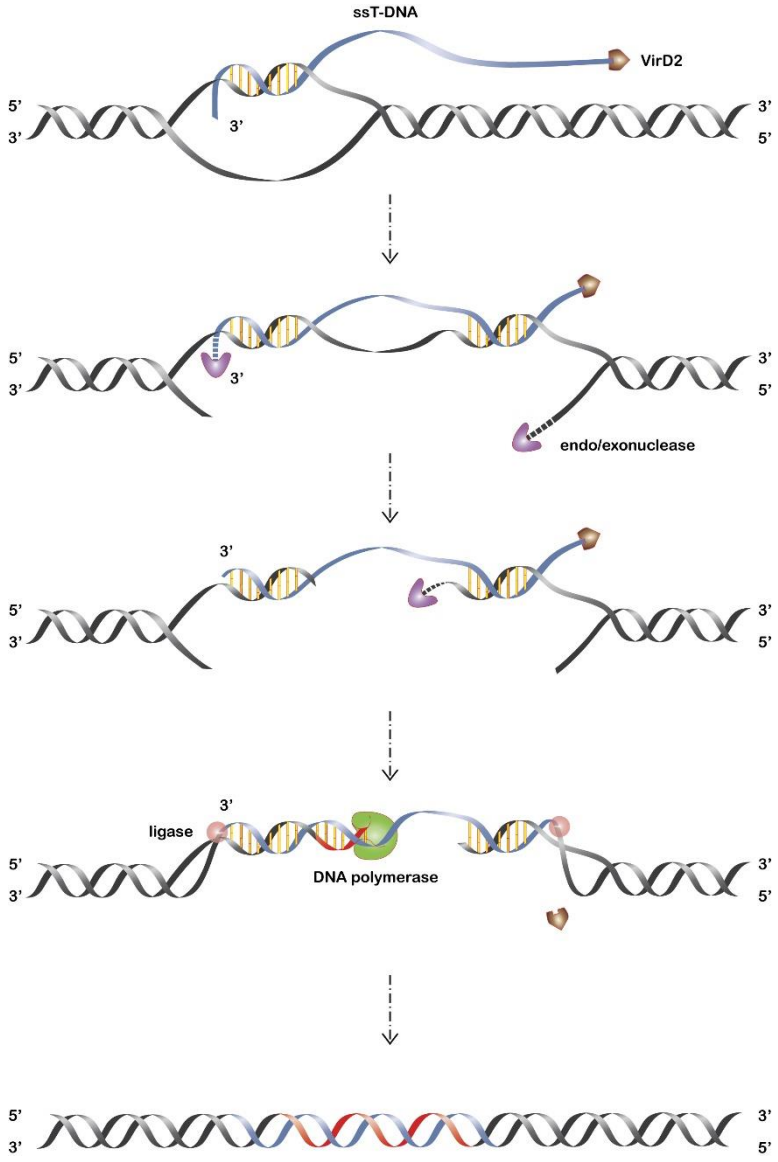


Figure 1-9. The HR-mediated integration machinery for T-DNA integration.

Table 1-1. Major eukaryotic NHEJ proteins. A summary of NHEJ proteins discussed in the text, with a focus on structural and enzymatic components. This list is not exhaustive as other proteins have been suggested to influence NHEJ. "+", described role in c-NHEJ or alt-NHEJ; "-", not involved by definitions of c-NHEJ and alt-NHEJ; "+/-", conflicting reports or uncertain role.

Category	Gene	c-NHEJ	alt-NHEJ	Description
DNA-PK	Ku70	+	-	DSB recognition, dRP lyase
	Ku80	+	-	Ku70 partner protein
	DNA-PKcs	+	-	DSB-dependent protein kinase
Ligase	Lig4	+	-	Ligase catalytic subunit
	XRCC4	+	-	Lig4 structural scaffold
	XLF	+	-	Lig4 structural scaffold
	Lig1	-	+/-	Ligase catalytic subunit
	Lig3	-	+	Ligase catalytic subunit
	XRCC1	-	+/-	Lig3 structural scaffold
MRN	Mre11	+	+	Dimerization, nuclease
	Rad50	+	+	Regulatory ATPase
	Nbs1	+	+	Protein recruitment
PARP	PARP-1	+/-	+	Poly-ADP ribose polymerase
	PARP-3	+		Poly-ADP ribose polymerase
Polymerase	Polμ	+		Gap filling
	Polλ	+		Gap filling, dRP lyase
Nuclease	Artemis	+		Endo/5' exonuclease
	Tdp1	+		3' phosphoesterase
	APLF	+		Endo/3' exonuclease
	CtIP		+	Supports 5' resection
Other	PNKP	+		5' kinase, 3' phosphatase
	Aprataxin	+		5'-AMP intermediate removal

Table 1-2: Host factors involving in delivery and integration of *Agrobacterium* T-DNA.

Transformation process	Host factor	Description	Related Interaction	Phenotype			
				Inhibition			OE
				Transt.	Stable	Other	
Cytoplasmic trafficking	Microtubules/ Dynein	Cargo transport/ MT motor protein	VirE2-ssDNA	N/A	N/A	Disrupted transport	
	Kinesin	MT motor protein		N/A	- - -		
	Actin	Cargo transport		- - -	-		
	Cyclophilins	Cyclosporine binding protein	VirD2	- - -	- - -		
Nuclear targeting	Importin α	NLS receptor/ Nuclear import	VirD2, VirE2, VirE3	- - -	- - -		Mutant rescue
	Importin β	Nuclear import	Importin α	-	-		
	CAK2Ms	CAK*-kinase/ Phosphorylate VirD2	VirD2	N/A	N/A		
	VIP1	VirE2-interacting protein 1	VirE2, MPK3	- - -	- - -		+++
	MPK3	Mitogen-activated protein kinase 3/ Phosphorylate VIP1	VIP1	N/A	N/A	Disrupted VIP1 localization	
	PP2C	Protein phosphatase 2C	VirD2	+	+		
	Caspase	Cysteine proteases/ Apoptosis	VirD2	+++	N/A		
	pBrP	Member of TFIIB	VirE3	N/A	N/A		
Chromatin targeting & T-DNA integration	VIP1	VirE2-interacting protein 1	VirE2, histone, VIP2	- - -	- - -		
	VIP2	VirE2-interacting protein 2	VirE2, VIP1	None	- - -		
	ASK/cullin	Components of SCF complex/ Protein proteolysis	VirF	N/A	N/A		
	VBF	VIP1-binding F-box protein/ Complement VirF	VIP1	N/A	- - -		
	Histones	Components of chromatin	VIP1	- - -	- - -		+++
	TBP	TATA-Binding Protein	VirD2	N/A	N/A		
	CAF-1	Chromatin assembly factor 1	Histones	N/A	+++		
	SGA1	Histone H3 chaperone	Histone H3	- - -	- - -		
	HDT1, HDT2, HDA19	Histone deacetylases	Histones	- - -	- - -		
DNA repair	DNA ligase IV	Component of NHEJ for DSB repair	XRCC4	N/A	None*	Repeat integration	
	Ku70, Rad50, Mre11, Xrs2, Sir4	Component of NHEJ for DSB repair		N/A	- - -	Integrate in (sub)telomere	
	Ku80	Component of NHEJ for DSB repair	Ku70	None	- - -*		
	XRCC4	X-Ray cross complementing protein 4	VirE2, DNA ligase IV				
Unknown	NLP	Nodulin-like protein		None	- - -		
	pCsn5-1	Component of the COP9 signalosome	VirE3	N/A	N/A		

*Controversial results; OE: Overexpression

Reference

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat Rev Genet *12*, 363-376.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., *et al.* (2003). Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science *301*, 653-657.

Anand, A., Krichevsky, A., Schornack, S., Lahaye, T., Tzfira, T., Tang, Y., Citovsky, V., and Mysore, K.S. (2007). Arabidopsis VIRE2 INTERACTING PROTEIN2 is required for Agrobacterium T-DNA integration in plants. Plant Cell *19*, 1695-1708.

Andres, S.N., and Junop, M.S. (2011). Crystallization and preliminary X-ray diffraction analysis of the human XRCC4-XLF complex. Acta crystallographica Section F, Structural biology and crystallization communications *67*, 1399-1402.

Andres, S.N., Vergnes, A., Ristic, D., Wyman, C., Modesti, M., and Junop, M. (2012). A human XRCC4-XLF complex bridges DNA. Nucleic acids research *40*, 1868-1878.

Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796-815.

Audebert, M., Salles, B., and 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. J Biol Chem *279*, 55117-55126.

Audebert, M., Salles, B., and Calsou, P. (2008). Effect of double-strand break DNA sequence on the PARP-1 NHEJ pathway. Biochemical and biophysical research communications *369*, 982-988.

Audebert, M., Salles, B., Weinfeld, M., and 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*, 257-265.

Aylon, Y., and Kupiec, M. (2005). Cell cycle-dependent regulation of double-strand break repair: a role for the CDK. Cell Cycle *4*, 259-261.

Bahmed, K., Nitiss, K.C., and Nitiss, J.L. (2010). Yeast Tdp1 regulates the fidelity of nonhomologous end joining. Proc Natl Acad Sci U S A *107*, 4057-4062.

Banta, L., and Montenegro, M. (2008). Agrobacterium and Plant Biotechnology. In Agrobacterium: From Biology to Biotechnology, T. Tzfira, and V. Citovsky, eds. (Springer New York), pp. 73-147.

Bebenek, K., Garcia-Diaz, M., Zhou, R.Z., Povirk, L.F., and Kunkel, T.A. (2010). Loop 1 modulates the fidelity of DNA polymerase lambda. Nucleic Acids Res *38*, 5419-5431.

Beucher, A., Birraux, J., Tchouandong, L., Barton, O., Shibata, A., Conrad, S., Goodarzi, A.A., Krempler, A., Jeggo, P.A., and Lobrich, M. (2009). ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2. EMBO J *28*, 3413-3427.

Bezzubova, O., Silbergleit, A., Yamaguchi-Iwai, Y., Takeda, S., and Buerstedde, J.M. (1997). Reduced X-ray resistance and homologous recombination frequencies in a RAD54-/- mutant of the chicken DT40 cell line. Cell *89*, 185-193.

Boboila, C., Oksenych, V., Gostissa, M., Wang, J.H., Zha, S., Zhang, Y., Chai, H., Lee, C.S., Jankovic, M., Saez, L.M., *et al.* (2012). Robust chromosomal DNA repair via alternative end-joining in the absence of X-ray repair cross-complementing protein 1 (XRCC1). Proc Natl Acad Sci U S A *109*, 2473-2478.

Brunaud, V., Balzergue, S., Dubreucq, B., Aubourg, S., Samson, F., Chauvin, S., Bechtold, N., Cruaud, C., DeRose, R., Pelletier, G., *et al.* (2002). T-DNA integration into the Arabidopsis genome depends on sequences of pre-insertion sites. EMBO Rep *3*, 1152-1157.

Bundock, P., den Dulk-Ras, A., Beijersbergen, A., and Hooykaas, P.J. (1995). Trans-kingdom T-DNA transfer from Agrobacterium tumefaciens to Saccharomyces cerevisiae. EMBO J *14*, 3206-3214.

Capecchi, M.R. (1989). Altering the genome by homologous recombination. Science *244*, 1288-1292.

Capp, J.P., Boudsocq, F., Bertrand, P., Laroche-Clary, A., Pourquier, P., Lopez, B.S., Cazaux, C., Hoffmann, J.S., and Canitrot, Y. (2006). The DNA polymerase lambda is required for the repair of non-compatible DNA double strand breaks by NHEJ in mammalian cells. Nucleic acids research *34*, 2998-3007.

Capp, J.P., Boudsocq, F., Besnard, A.G., Lopez, B.S., Cazaux, C., Hoffmann, J.S., and Canitrot, Y. (2007). Involvement of DNA polymerase mu in the repair of a specific subset of DNA double-strand breaks in mammalian cells. Nucleic acids research *35*, 3551-3560.

Chan, C.Y., Galli, A., and Schiestl, R.H. (2008). Pol3 is involved in nonhomologous end-joining in Saccharomyces cerevisiae. DNA repair *7*, 1531-1541.

Chilton, M.D., and Que, Q. (2003). Targeted integration of T-DNA into the tobacco genome at double-stranded breaks: new insights on the mechanism of T-DNA integration. Plant Physiol *133*, 956-965.

Chiruvella, K.K., Sebastian, R., Sharma, S., Karande, A.A., Choudhary, B., and Raghavan, S.C. (2012). Time-dependent predominance of nonhomologous DNA end-joining pathways during embryonic development in mice. Journal of molecular biology *417*, 197-211.

Clerici, M., Mantiero, D., Guerini, I., Lucchini, G., and Longhese, M.P. (2008). The Yku70-Yku80 complex contributes to regulate double-strand break processing and checkpoint activation during the cell cycle. EMBO reports *9*, 810-818.

Covo, S., de Villartay, J.P., Jeggo, P.A., and Livneh, Z. (2009). Translesion DNA synthesis-assisted non-homologous end-joining of complex double-strand breaks prevents loss of DNA sequences in mammalian cells. Nucleic acids research *37*, 6737-6745.

Dafny-Yelin, M., Levy, A., Dafny, R., and Tzfira, T. (2015). Blocking Single-Stranded Transferred DNA Conversion to Double-Stranded Intermediates by Overexpression of Yeast DNA REPLICATION FACTOR A. Plant Physiol *167*, 153-163.

Daley, J.M., Laan, R.L., Suresh, A., and Wilson, T.E. (2005). DNA joint dependence of pol X family polymerase action in nonhomologous end joining. The Journal of biological chemistry *280*, 29030-29037.

Daley, J.M., and Wilson, T.E. (2005). Rejoining of DNA double-strand breaks as a function of overhang length. Molecular and cellular biology *25*, 896-906.

de Groot, M.J., Bundock, P., Hooykaas, P.J., and Beijersbergen, A.G. (1998). Agrobacterium tumefaciens-mediated transformation of filamentous fungi. Nat Biotechnol *16*, 839-842.

Delarue, M., Boule, J.B., Lescar, J., Expert-Bezancon, N., Jourdan, N., Sukumar, N., Rougeon, F., and Papanicolaou, C. (2002). Crystal structures of a template-independent DNA polymerase: murine terminal deoxynucleotidyltransferase. EMBO J *21*, 427-439.

Della-Maria, J., Zhou, Y., Tsai, M.S., Kuhnlein, J., Carney, J.P., Paull, T.T., and Tomkinson, A.E. (2011). Human Mre11/human Rad50/Nbs1 and DNA ligase IIIalpha/XRCC1 protein complexes act together in an alternative nonhomologous end joining pathway. The Journal of biological chemistry *286*, 33845-33853.

Deriano, L., Stracker, T.H., Baker, A., Petrini, J.H., and Roth, D.B. (2009). Roles for NBS1 in alternative nonhomologous end-joining of V(D)J recombination intermediates. Mol Cell *34*, 13-25.

DeRose, E.F., Clarkson, M.W., Gilmore, S.A., Galban, C.J., Tripathy, A., Havener, J.M., Mueller, G.A., Ramsden, D.A., London, R.E., and Lee, A.L. (2007). Solution structure of polymerase mu's BRCT Domain reveals an element essential for its role in nonhomologous end joining. Biochemistry *46*, 12100-12110.

Di Virgilio, M., and Gautier, J. (2005). Repair of double-strand breaks by nonhomologous end joining in the absence of Mre11. The Journal of cell biology *171*, 765-771.

Dinkelmann, M., Spehalski, E., Stoneham, T., Buis, J., Wu, Y., Sekiguchi, J.M., and Ferguson, D.O. (2009). Multiple functions of MRN in end-joining pathways during isotype class switching. Nature structural & molecular biology *16*, 808-813.

Dore, A.S., Furnham, N., Davies, O.R., Sibanda, B.L., Chirgadze, D.Y., Jackson, S.P., Pellegrini, L., and Blundell, T.L. (2006). Structure of an Xrcc4-DNA ligase IV yeast ortholog complex reveals a novel BRCT interaction mode. DNA repair *5*, 362-368.

Douglas, P., Gupta, S., Morrice, N., Meek, K., and Lees-Miller, S.P. (2005). DNA-PK-dependent phosphorylation of Ku70/80 is not required for non-homologous end joining. DNA repair *4*, 1006-1018.

Ellenberger, T., and Tomkinson, A.E. (2008). Eukaryotic DNA ligases: structural and functional insights. Annu Rev Biochem *77*, 313-338.

Endo, M., Ishikawa, Y., Osakabe, K., Nakayama, S., Kaya, H., Araki, T., Shibahara, K., Abe, K., Ichikawa, H., Valentine, L., *et al.* (2006). Increased frequency of homologous recombination and T-DNA integration in Arabidopsis CAF-1 mutants. EMBO J *25*, 5579-5590.

Feng, L., and Chen, J. (2012). The E3 ligase RNF8 regulates KU80 removal and NHEJ repair. Nature structural & molecular biology *19*, 201-206.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., *et al.* (2006). Copy number variation: new insights in genome diversity. Genome Res *16*, 949-961.

Friesner, J., and Britt, A.B. (2003). Ku80- and DNA ligase IV-deficient plants are sensitive to ionizing radiation and defective in T-DNA integration. Plant J *34*, 427-440.

Gapud, E.J., and Sleckman, B.P. (2011). Unique and redundant functions of ATM and DNA-PKcs during V(D)J recombination. Cell Cycle *10*, 1928-1935.

Garcia-Diaz, M., Bebenek, K., Krahn, J.M., Kunkel, T.A., and Pedersen, L.C. (2005). A closed conformation for the Pol lambda catalytic cycle. Nat Struct Mol Biol *12*, 97-98.

Gellert, M. (2002). V(D)J recombination: RAG proteins, repair factors, and regulation. Annu Rev Biochem *71*, 101-132.

Gelvin, S.B., and Kim, S.I. (2007). Effect of chromatin upon Agrobacterium T-DNA integration and transgene expression. Biochim Biophys Acta *1769*, 410-421.

Gheysen, G., Villarroel, R., and Van Montagu, M. (1991). Illegitimate recombination in plants: a model for T-DNA integration. Genes Dev *5*, 287-297.

Godio, R.P., Fouces, R., Gudina, E.J., and Martin, J.F. (2004). Agrobacterium tumefaciens-mediated transformation of the antitumor clavaric acid-producing basidiomycete Hypholoma sublateritium. Curr Genet *46*, 287-294.

Grabarz, A., Barascu, A., Guirouilh-Barbat, J., and Lopez, B.S. (2012). Initiation of DNA double strand break repair: signaling and single-stranded resection dictate the choice between homologous recombination, non-homologous end-joining and alternative end-joining. American journal of cancer research *2*, 249-268.

Gu, J., Lu, H., Tippin, B., Shimazaki, N., Goodman, M.F., and Lieber, M.R. (2007). XRCC4:DNA ligase IV can ligate incompatible DNA ends and can ligate across gaps. EMBO J *26*, 1010-1023.

Gu, Y., Seidl, K.J., Rathbun, G.A., Zhu, C., Manis, J.P., van der Stoep, N., Davidson, L., Cheng, H.L., Sekiguchi, J.M., Frank, K., *et al.* (1997). Growth retardation and leaky SCID phenotype of Ku70-deficient mice. Immunity *7*, 653-665.

Gu, Y., Sekiguchi, J., Gao, Y., Dikkes, P., Frank, K., Ferguson, D., Hasty, P., Chun, J., and Alt, F.W. (2000). Defective embryonic neurogenesis in Ku-deficient but not DNA-dependent protein kinase catalytic subunit-deficient mice. Proc Natl Acad Sci U S A *97*, 2668-2673.

Haber, J.E. (1998). Mating-type gene switching in Saccharomyces cerevisiae. Annu Rev Genet *32*, 561-599.

Haber, J.E. (2000). Recombination: a frank view of exchanges and vice versa. Curr Opin Cell Biol *12*, 286-292.

Hammel, M., Rey, M., Yu, Y., Mani, R.S., Classen, S., Liu, M., Pique, M.E., Fang, S., Mahaney, B.L., Weinfeld, M., *et al.* (2011). XRCC4 protein interactions with XRCC4-like factor (XLF) create an extended grooved scaffold for DNA ligation and double strand break repair. The Journal of biological chemistry *286*, 32638-32650.

Hammel, M., Yu, Y., Fang, S., Lees-Miller, S.P., and Tainer, J.A. (2010). XLF regulates filament architecture of the XRCC4.ligase IV complex. Structure *18*, 1431-1442.

Han, L., Mao, W., and Yu, K. (2012). X-ray repair cross-complementing protein 1 (XRCC1) deficiency enhances class switch recombination and is permissive for alternative end joining. Proc Natl Acad Sci U S A *109*, 4604-4608.

Hanin, M., and Paszkowski, J. (2003). Plant genome modification by homologous recombination. Curr Opin Plant Biol *6*, 157-162.

Helmink, B.A., Bredemeyer, A.L., Lee, B.S., Huang, C.Y., Sharma, G.G., Walker, L.M., Bednarski, J.J., Lee, W.L., Pandita, T.K., Bassing, C.H., *et al.* (2009). MRN complex function in the repair of chromosomal Rag-mediated DNA double-strand breaks. J Exp Med *206*, 669-679.

Hopfner, K.P., Craig, L., Moncalian, G., Zinkel, R.A., Usui, T., Owen, B.A., Karcher, A., Henderson, B., Bodmer, J.L., McMurray, C.T., *et al.* (2002). The Rad50 zinc-hook is a structure joining Mre11 complexes in DNA recombination and repair. Nature *418*, 562-566.

Huang, J., and Dynan, W.S. (2002). Reconstitution of the mammalian DNA double-strand break end-joining reaction reveals a requirement for an Mre11/Rad50/NBS1-containing fraction. Nucleic acids research *30*, 667-674.

Huertas, P., and Jackson, S.P. (2009). Human CtIP mediates cell cycle control of DNA end resection and double strand break repair. The Journal of biological chemistry *284*, 9558-9565.

Ira, G., Pellicioli, A., Balijja, A., Wang, X., Fiorani, S., Carotenuto, W., Liberi, G., Bressan, D., Wan, L., Hollingsworth, N.M., *et al.* (2004). DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. Nature *431*, 1011-1017.

Juarez, R., Ruiz, J.F., Nick McElhinny, S.A., Ramsden, D., and Blanco, L. (2006). A specific loop in human DNA polymerase mu allows switching between creative and DNA-instructed synthesis. Nucleic acids research *34*, 4572-4582.

Karanam, K., Kafri, R., Loewer, A., and Lahav, G. (2012). Quantitative Live Cell Imaging Reveals a Gradual Shift between DNA Repair Mechanisms and a Maximal Use of HR in Mid S Phase. Mol Cell *47*, 320-329.

Karathanasis, E., and Wilson, T.E. (2002). Enhancement of Saccharomyces cerevisiae end-joining efficiency by cell growth stage but not by impairment of recombination. Genetics *161*, 1015-1027.

Kim, S.I., Veena, and Gelvin, S.B. (2007). Genome-wide analysis of Agrobacterium T-DNA integration sites in the Arabidopsis genome generated under non-selective conditions. Plant J *51*, 779-791.

Koch, C.A., Agyei, R., Galicia, S., Metalnikov, P., O'Donnell, P., Starostine, A., Weinfeld, M., and Durocher, D. (2004). Xrcc4 physically links DNA end processing by polynucleotide kinase to DNA ligation by DNA ligase IV. EMBO J *23*, 3874-3885.

Kunik, T., Tzfira, T., Kapulnik, Y., Gafni, Y., Dingwall, C., and Citovsky, V. (2001). Genetic transformation of HeLa cells by Agrobacterium. Proceedings of the National Academy of Sciences of the United States of America *98*, 1871-1876.

Lacroix, B., Loyter, A., and Citovsky, V. (2008). Association of the Agrobacterium T-DNA-protein complex with plant nucleosomes. Proceedings of the National Academy of Sciences of the United States of America *105*, 15429-15434.

Lammens, K., Bemeleit, D.J., Mockel, C., Clausing, E., Schele, A., Hartung, S., Schiller, C.B., Lucas, M., Angermuller, C., Soding, J., *et al.* (2011). The Mre11:Rad50 structure shows an ATP-dependent molecular clamp in DNA double-strand break repair. Cell *145*, 54-66.

Langerak, P., Mejia-Ramirez, E., Limbo, O., and Russell, P. (2011). Release of Ku and MRN from DNA ends by Mre11 nuclease activity and Ctp1 is required for homologous recombination repair of double-strand breaks. PLoS genetics *7*, e1002271.

Lee-Theilen, M., Matthews, A.J., Kelly, D., Zheng, S., and Chaudhuri, J. (2011). CtIP promotes microhomology-mediated alternative end joining during class-switch recombination. Nature structural & molecular biology *18*, 75-79.

Lempiainen, H., and Halazonetis, T.D. (2009). Emerging common themes in regulation of PIKKs and PI3Ks. EMBO J *28*, 3067-3073.

Li, J., Vaidya, M., White, C., Vainstein, A., Citovsky, V., and Tzfira, T. (2005). Involvement of KU80 in T-DNA integration in plant cells. Proceedings of the National Academy of Sciences of the United States of America *102*, 19231-19236.

Li, Y., Chirgadze, D.Y., Bolanos-Garcia, V.M., Sibanda, B.L., Davies, O.R., Ahnesorg, P., Jackson, S.P., and Blundell, T.L. (2008). Crystal structure of human XLF/Cernunnos reveals unexpected differences from XRCC4 with implications for NHEJ. EMBO J *27*, 290-300.

Li, Y., Rosso, M.G., Ulker, B., and Weisshaar, B. (2006). Analysis of T-DNA insertion site distribution patterns in Arabidopsis thaliana reveals special features of genes without insertions. Genomics *87*, 645-652.

Liang, L., Deng, L., Nguyen, S.C., Zhao, X., Maulion, C.D., Shao, C., and Tischfield, J.A. (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*, 3297-3310.

Liang, Z., and Tzfira, T. (2013). In vivo formation of double-stranded T-DNA molecules by T-strand priming. Nature communications *4*, 2253.

Lorence, A., and Verpoorte, R. (2004). Gene transfer and expression in plants. Methods Mol Biol *267*, 329-350.

Ma, Y., Lu, H., Tippin, B., Goodman, M.F., Shimazaki, N., Koiwai, O., Hsieh, C.L., Schwarz, K., and Lieber, M.R. (2004). A biochemically defined system for mammalian nonhomologous DNA end joining. Mol Cell *16*, 701-713.

Malik, I., Garrido, M., Bahr, M., Kugler, S., and Michel, U. (2006). Comparison of test systems for RNAinterference. Biochem Biophys Res Commun *341*, 245-253.

Manolis, K.G., Nimmo, E.R., Hartsuiker, E., Carr, A.M., Jeggo, P.A., and Allshire, R.C. (2001). Novel functional requirements for non-homologous DNA end joining in Schizosaccharomyces pombe. EMBO J *20*, 210-221.

Mansour, W.Y., Rhein, T., and Dahm-Daphi, J. (2010). The alternative end-joining pathway for repair of DNA double-strand breaks requires PARP1 but is not dependent upon microhomologies. Nucleic acids research *38*, 6065-6077.

Mao, Z., Hine, C., Tian, X., Van Meter, M., Au, M., Vaidya, A., Seluanov, A., and Gorbunova, V. (2011). SIRT6 promotes DNA repair under stress by activating PARP1. Science (New York, NY) *332*, 1443-1446.

Mari, P.O., Florea, B.I., Persengiev, S.P., Verkaik, N.S., Bruggenwirth, H.T., Modesti, M., Giglia-Mari, G., Bezstarosti, K., Demmers, J.A., Luider, T.M., *et al.* (2006). Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. Proc Natl Acad Sci U S A *103*, 18597-18602.

Martin, M.J., Juarez, R., and Blanco, L. (2012). DNA-binding determinants promoting NHEJ by human Polmu. Nucleic acids research.

Matzke, A.J., and Matzke, M.A. (1998). Position effects and epigenetic silencing of plant transgenes. Curr Opin Plant Biol *1*, 142-148.

Mayerhofer, R., Koncz-Kalman, Z., Nawrath, C., Bakkeren, G., Crameri, A., Angelis, K., Redei, G.P., Schell, J., Hohn, B., and Koncz, C. (1991). T-DNA integration: a mode of illegitimate recombination in plants. EMBO J *10*, 697-704.

Michielse, C.B., Arentshorst, M., Ram, A.F., and van den Hondel, C.A. (2005). Agrobacterium-mediated transformation leads to improved gene replacement efficiency in Aspergillus awamori. Fungal Genet Biol *42*, 9-19.

Milne, G.T., Jin, S., Shannon, K.B., and Weaver, D.T. (1996). Mutations in two Ku homologs define a DNA end-joining repair pathway in Saccharomyces cerevisiae. Molecular and cellular biology *16*, 4189-4198.

Mockel, C., Lammens, K., Schele, A., and Hopfner, K.P. (2012). ATP driven structural changes of the bacterial Mre11:Rad50 catalytic head complex. Nucleic Acids Res *40*, 914-927.

Moon, A.F., Garcia-Diaz, M., Bebenek, K., Davis, B.J., Zhong, X., Ramsden, D.A., Kunkel, T.A., and Pedersen, L.C. (2007). Structural insight into the substrate specificity of DNA Polymerase mu. Nat Struct Mol Biol *14*, 45-53.

Mysore, K.S., Nam, J., and Gelvin, S.B. (2000). An Arabidopsis histone H2A mutant is deficient in Agrobacterium T-DNA integration. Proceedings of the National Academy of Sciences of the United States of America *97*, 948-953.

Nick McElhinny, S.A., Havener, J.M., Garcia-Diaz, M., Juarez, R., Bebenek, K., Kee, B.L., Blanco, L., Kunkel, T.A., and Ramsden, D.A. (2005). A gradient of template dependence defines distinct biological roles for family X polymerases in nonhomologous end joining. Mol Cell *19*, 357-366.

Ochi, T., Wu, Q., Chirgadze, D.Y., Grossmann, J.G., Bolanos-Garcia, V.M., and Blundell, T.L. (2012). Structural Insights into the Role of Domain Flexibility in Human DNA Ligase IV. Structure *20*, 1212-1222.

Palmbos, P.L., Daley, J.M., and Wilson, T.E. (2005). Mutations of the Yku80 C terminus and Xrs2 FHA domain specifically block yeast nonhomologous end joining. Molecular and cellular biology *25*, 10782-10790.

Pascal, J.M., O'Brien, P.J., Tomkinson, A.E., and Ellenberger, T. (2004). Human DNA ligase I completely encircles and partially unwinds nicked DNA. Nature *432*, 473-478.

Perry, J.J., Cotner-Gohara, E., Ellenberger, T., and Tainer, J.A. (2010). Structural dynamics in DNA damage signaling and repair. Curr Opin Struct Biol *20*, 283-294.

Piers, K.L., Heath, J.D., Liang, X., Stephens, K.M., and Nester, E.W. (1996). Agrobacterium tumefaciens-mediated transformation of yeast. Proceedings of the National Academy of Sciences of the United States of America *93*, 1613-1618.

Postel-Vinay, S., Vanhecke, E., Olaussen, K.A., Lord, C.J., Ashworth, A., and Soria, J.C. (2012). The potential of exploiting DNA-repair defects for optimizing lung cancer treatment. Nat Rev Clin Oncol *9*, 144-155.

Postow, L., Ghenoiu, C., Woo, E.M., Krutchinsky, A.N., Chait, B.T., and Funabiki, H. (2008). Ku80 removal from DNA through double strand break-induced ubiquitylation. The Journal of cell biology *182*, 467-479.

Puchta, H. (2005). The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J Exp Bot *56*, 1-14.

Ramsden, D.A., and Asagoshi, K. (2012). DNA polymerases in nonhomologous end joining: Are there any benefits to standing out from the crowd? Environ Mol Mutagen.

Rass, E., Grabarz, A., Plo, I., Gautier, J., Bertrand, P., and Lopez, B.S. (2009). Role of Mre11 in chromosomal nonhomologous end joining in mammalian cells. Nature structural & molecular biology *16*, 819-824.

Robert, I., Dantzer, F., and Reina-San-Martin, B. (2009). Parp1 facilitates alternative NHEJ, whereas Parp2 suppresses IgH/c-myc translocations during immunoglobulin class switch recombination. J Exp Med *206*, 1047-1056.

Rong, Y.S., and Golic, K.G. (2000). Gene targeting by homologous recombination in Drosophila. Science *288*, 2013-2018.

Ropars, V., Drevet, P., Legrand, P., Baconnais, S., Amram, J., Faure, G., Marquez, J.A., Pietrement, O., Guerois, R., Callebaut, I., *et al.* (2011). Structural characterization of filaments formed by human Xrcc4-Cernunnos/XLF complex involved in nonhomologous DNA end-joining. Proc Natl Acad Sci U S A *108*, 12663-12668.

Sallmyr, A., Tomkinson, A.E., and Rassool, F.V. (2008). Up-regulation of WRN and DNA ligase IIIalpha in chronic myeloid leukemia: consequences for the repair of DNA double-strand breaks. Blood *112*, 1413-1423.

Salomon, S., and Puchta, H. (1998). Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. EMBO J *17*, 6086-6095.

Saribasak, H., Maul, R.W., Cao, Z., McClure, R.L., Yang, W., McNeill, D.R., Wilson, D.M., 3rd, and Gearhart, P.J. (2011). XRCC1 suppresses somatic hypermutation and promotes alternative nonhomologous end joining in Igh genes. J Exp Med *208*, 2209-2216.

Schaefer, D.G., and Zryd, J.P. (1997). Efficient gene targeting in the moss Physcomitrella patens. Plant J *11*, 1195-1206.

Schiller, C.B., Lammens, K., Guerini, I., Coordes, B., Feldmann, H., Schlauderer, F., Mockel, C., Schele, A., Strasser, K., Jackson, S.P., *et al.* (2012). Structure of Mre11-Nbs1 complex yields insights into ataxia-telangiectasia-like disease mutations and DNA damage signaling. Nat Struct Mol Biol *19*, 693-700.

Schneeberger, R.G., Zhang, K., Tatarinova, T., Troukhan, M., Kwok, S.F., Drais, J., Klinger, K., Orejudos, F., Macy, K., Bhakta, A., *et al.* (2005). Agrobacterium T-DNA integration in Arabidopsis is correlated with DNA sequence compositions that occur frequently in gene promoter regions. Funct Integr Genomics *5*, 240-253.

Shao, Z., Davis, A.J., Fattah, K.R., So, S., Sun, J., Lee, K.J., Harrison, L., Yang, J., and Chen, D.J. (2012). Persistently bound Ku at DNA ends attenuates DNA end resection and homologous recombination. DNA repair *11*, 310-316.

Sheludko, Y.V. (2008). Agrobacterium-mediated transient expression as an approach to production of recombinant proteins in plants. Recent Pat Biotechnol *2*, 198-208.

Shrivastav, M., De Haro, L.P., and Nickoloff, J.A. (2008). Regulation of DNA double-strand break repair pathway choice. Cell Res *18*, 134-147.

Shultz, J.L., Kurunam, D., Shopinski, K., Iqbal, M.J., Kazi, S., Zobrist, K., Bashir, R., Yaegashi, S., Lavu, N., Afzal, A.J., *et al.* (2006). The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of Glycine max. Nucleic Acids Res *34*, D758-765.

Sibanda, B.L., Chirgadze, D.Y., and Blundell, T.L. (2010). Crystal structure of DNA-PKcs reveals a large open-ring cradle comprised of HEAT repeats. Nature *463*, 118-121.

Simsek, D., Brunet, E., Wong, S.Y., Katyal, S., Gao, Y., McKinnon, P.J., Lou, J., Zhang, L., Li, J., Rebar, E.J., *et al.* (2011). DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation. PLoS genetics *7*, e1002080.

Singer, K., Shiboleth, Y.M., Li, J., and Tzfira, T. (2012). Formation of complex extrachromosomal T-DNA structures in Agrobacterium tumefaciens-infected plants. Plant Physiol *160*, 511-522.

Stewart, C.N., Jr., Richards, H.A.t., and Halfhill, M.D. (2000). Transgenic plants and biosafety: science, misconceptions and public perceptions. Biotechniques *29*, 832-836, 838-843.

Symington, L.S., and Gautier, J. (2011). Double-strand break end resection and repair pathway choice. Annual review of genetics *45*, 247-271.

Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J., and Stahl, F.W. (1983). The double-strand-break repair model for recombination. Cell *33*, 25-35.

Taylor, N.J., and Fauquet, C.M. (2002). Microparticle bombardment as a tool in plant science and agricultural biotechnology. DNA Cell Biol *21*, 963-977.

Tinland, B. (1996). The integration of T-DNA into plant genomes. Trends in Plant Science *1*, 178-184.

Tseng, S.F., Gabriel, A., and Teng, S.C. (2008). Proofreading activity of DNA polymerase Pol2 mediates 3'-end processing during nonhomologous end joining in yeast. PLoS genetics *4*, e1000060.

Tzfira, T., Frankman, L.R., Vaidya, M., and Citovsky, V. (2003). Site-specific integration of Agrobacterium tumefaciens T-DNA via double-stranded intermediates. Plant Physiol *133*, 1011-1023.

Tzfira, T., Li, J., Lacroix, B., and Citovsky, V. (2004). Agrobacterium T-DNA integration: molecules and models. Trends Genet *20*, 375-383.

Walker, J.R., Corpina, R.A., and Goldberg, J. (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature *412*, 607-614.

Wang, H., Rosidi, B., Perrault, R., Wang, M., Zhang, L., Windhofer, F., and Iliakis, G. (2005). DNA ligase III as a candidate component of backup pathways of nonhomologous end joining. Cancer research *65*, 4020-4030.

Wang, M., Wu, W., Wu, W., Rosidi, B., Zhang, L., Wang, H., and Iliakis, G. (2006). PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. Nucleic acids research *34*, 6170-6182.

Wang, Y., Yau, Y.Y., Perkins-Balding, D., and Thomson, J.G. (2011). Recombinase technology: applications and possibilities. Plant Cell Rep *30*, 267-285.

Wang, Y.G., Nnakwe, C., Lane, W.S., Modesti, M., and Frank, K.M. (2004). Phosphorylation and regulation of DNA ligase IV stability by DNA-dependent protein kinase. The Journal of biological chemistry *279*, 37282-37290.

Weinthal, D., Tovkach, A., Zeevi, V., and Tzfira, T. (2010). Genome editing in plant cells by zinc finger nucleases. Trends Plant Sci *15*, 308-321.

Williams, R.S., Dodson, G.E., Limbo, O., Yamada, Y., Williams, J.S., Guenther, G., Classen, S., Glover, J.N., Iwasaki, H., Russell, P., *et al.* (2009). Nbs1 flexibly tethers Ctp1 and Mre11-Rad50 to coordinate DNA double-strand break processing and repair. Cell *139*, 87-99.

Williams, R.S., Moncalian, G., Williams, J.S., Yamada, Y., Limbo, O., Shin, D.S., Groocock, L.M., Cahill, D., Hitomi, C., Guenther, G., *et al.* (2008). Mre11 dimers coordinate DNA end bridging and nuclease processing in double-strand-break repair. Cell *135*, 97-109.

Wu, D., Topper, L.M., and Wilson, T.E. (2008). Recruitment and dissociation of nonhomologous end joining proteins at a DNA double-strand break in Saccharomyces cerevisiae. Genetics *178*, 1237-1249.

Wu, P.Y., Frit, P., Meesala, S., Dauvillier, S., Modesti, M., Andres, S.N., Huang, Y., Sekiguchi, J., Calsou, P., Salles, B., *et al.* (2009). Structural and functional interaction between the human DNA repair proteins DNA ligase IV and XRCC4. Molecular and cellular biology *29*, 3163-3172.

Wu, Q., Ochi, T., Matak-Vinkovic, D., Robinson, C.V., Chirgadze, D.Y., and Blundell, T.L. (2011). Non-homologous end-joining partners in a helical dance: structural studies of XLF-

XRCC4 interactions. Biochemical Society transactions *39*, 1387-1392, suppl 1382 p following 1392.

Xie, A., Kwok, A., and Scully, R. (2009). Role of mammalian Mre11 in classical and alternative nonhomologous end joining. Nature structural & molecular biology *16*, 814-818.

Yamaguchi-Iwai, Y., Sonoda, E., Sasaki, M.S., Morrison, C., Haraguchi, T., Hiraoka, Y., Yamashita, Y.M., Yagi, T., Takata, M., Price, C., *et al.* (1999). Mre11 is essential for the maintenance of chromosomal DNA in vertebrate cells. EMBO J *18*, 6619-6629.

Yu, A.M., and McVey, M. (2010). Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. Nucleic acids research *38*, 5706-5717.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science *296*, 79-92.

Yu, Y., Mahaney, B.L., Yano, K., Ye, R., Fang, S., Douglas, P., Chen, D.J., and Lees-Miller, S.P. (2008). DNA-PK and ATM phosphorylation sites in XLF/Cernunnos are not required for repair of DNA double strand breaks. DNA repair *7*, 1680-1692.

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., Lee, D., Quackenbush, J., and Buell, C.R. (2003). The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res *31*, 229-233.

Yuan, Q., Ouyang, S., Wang, A., Zhu, W., Maiti, R., Lin, H., Hamilton, J., Haas, B., Sultana, R., Cheung, F., *et al.* (2005). The institute for genomic research Osa1 rice genome annotation database. Plant Physiol *138*, 18-26.

Zhang, X., and Paull, T.T. (2005). The Mre11/Rad50/Xrs2 complex and non-homologous end-joining of incompatible ends in S. cerevisiae. DNA repair *4*, 1281-1294.

Zhang, Y., and Jasin, M. (2011). An essential role for CtIP in chromosomal translocation formation through an alternative end-joining pathway. Nature structural & molecular biology *18*, 80-84.

Zhang, Y., Shim, E.Y., Davis, M., and Lee, S.E. (2009). Regulation of repair choice: Cdk1 suppresses recruitment of end joining factors at DNA breaks. DNA repair *8*, 1235-1241.

Zhang, Z., Hu, W., Cano, L., Lee, T.D., Chen, D.J., and Chen, Y. (2004). Solution structure of the C-terminal domain of Ku80 suggests important sites for protein-protein interactions. Structure *12*, 495-502.

Zhang, Z., Zhu, L., Lin, D., Chen, F., Chen, D.J., and Chen, Y. (2001). The three-dimensional structure of the C-terminal DNA-binding domain of human Ku70. J Biol Chem *276*, 38231-38236.

Zhou, T., Akopiants, K., Mohapatra, S., Lin, P.S., Valerie, K., Ramsden, D.A., Lees-Miller, S.P., and Povirk, L.F. (2009). Tyrosyl-DNA phosphodiesterase and the repair of 3'-phosphoglycolate-terminated DNA double-strand breaks. DNA repair *8*, 901-911.

Ziemienowicz, A., Tzfira, T., and Hohn, B. (2008). Mechanisms of T-DNA integration. In Agrobacterium: From Biology to Biotechnology, T. Tzfira, and V. Citovsky, eds. (Springer New York), pp. 395-440.

CHAPTER 2

In Vivo Formation of Double-Stranded T-DNA Molecules by T-Strand Priming

Abstract

During plant genetic transformation, *Agrobacterium* transfers a single-stranded DNA (T-strand) into the host cell. Increasing evidence suggests that double-stranded (ds) T-DNA, converted from T-strands, are potent substrates for integration. Nevertheless, the molecular mechanism governing T-strand conversion to dsT-DNA is unknown. Integrated T-DNA molecules typically exhibit deletions at 3'end as compared with their 5'end. We hypothesize that this may result from asymmetric polymerization of T-DNA's ends. Here we show that β-glucuronidase (GUS) expression from sense T-strands is more efficient than from antisense T-strands, supporting asymmetric conversion. Co-transfection with two partially complementary, truncated GUS-encoding T-strands results in GUS expression, which suggests functional hybridization of the T-strands via complementary annealing and supports the notion that T-strands can anneal with primers. Indeed, red fluorescent protein (RFP) expression from mutated T-strand can be restored by delivery of synthetic DNA and RNA oligonucleotides with partial wild-type *RFP* sequence, implying the involvement of plant DNA repair machinery.

Introduction

Agrobacterium genetically transforms its host by delivering its transferred DNA (T-DNA) into the host genome, where it is stably integrated and expressed. The T-DNA, defined by two 25-bp direct-repeat border sequences, travels as a single-stranded (ss) DNA molecule (T-strand), with a single *Agrobacterium* VirD2 protein attached to its 5' end (Howard *et al.*, 1989; Tinland *et al.*, 1994; Young and Nester, 1988). Once inside the plant cell, the VirD2-capped T-strand is thought to be packed by numerous *Agrobacterium* VirE2 proteins (Citovsky *et al.*, 1989; Abu-Arish *et al.*, 2004). Both VirD2 and VirE2 are thought to function during T-strand nuclear import and intra-nuclear transport (Grange *et al.*, 2008; Lacroix *et al.*, 2008; Ziemienowicz *et al.*, 2001). VirD2 may also function during the genomic integration of the T-DNA through unknown mechanisms (Mysore *et al.*, 1998; Tinland *et al.*, 1995). The T-DNA sequence does not code for the machinery needed for its transport or integration into the plant genome and it can therefore be replaced by virtually any other sequence without affecting the transformation process. T-DNA molecules integrate at random locations across the plant genome possibly via non-homologous end joining, and are only rarely directed to and integrated in specific genomic locations by homologous recombination (Salomon and Puchta, 1998; Tzfira *et al.*, 2004a; Kim *et al.*, 2007). Sequence analysis of integration events has revealed that integrated T-DNAs often lose part of their 3' end, while the 5' end can remain intact (Kim *et al.*, 2007; Alonso *et al.*, 2003; Chen *et al.*, 2003; Mayerhofer *et al.*, 1991; Gheysen *et al.*, 1991). Various models have been proposed to explain the structure of integration events and the random nature of T-DNA integration into the plant genome; of these, the microhomology-based T-DNA integration models are predominant (Tinland *et al.*, 1995; Meza *et al.*, 2002). According to these

models, integration is initiated by annealing of the 3’ or 5’ end of the T-strand to unwound plant DNA at pre-integration genomic sites, followed by nicking of the target DNA, ligation of the 5’ end to one of the target DNA strands and trimming of the T-strand's unprotected 3' end (Tinland *et al.*, 1995; Meza *et al.*, 2002). Alternatively, the T-strand might be converted into double-stranded (ds) T-DNA molecules in the plant cell, which can then integrate into genomic double-strand breaks (DSBs) (Salomon and Puchta, 1998; Chilton and Que, 2003; Tzfira *et al.*, 2003). While the direct integration of T-strand molecules cannot be ruled out, the rapid and efficient conversion of T-strands to dsT-DNA suggests that the latter may serve as an important substrate for the integration machinery. Notably, dsT-DNA molecules are important substrates for many recombinase-mediated gene-replacement and site-specific-integration systems (Tzfira *et al.*, 2012; Puchta, 2003). However, the mechanism by which T-strands are converted into dsT-DNAs is still unknown.

Here we use several functional assays to investigate the process of T-strand conversion into dsT-DNA molecules. We show that expression of the β-glucuronidase (GUS) encoding gene from sense T-strands is more efficient than from antisense molecules and that co-transfection of tobacco leaves with two partially complementary, truncated GUS-encoding T-strands, but not with individual or partially overlapping truncated GUS-encoding T-strands, results in GUS expression. Using functional mutated RFP-repair assay, we demonstrate that T-strand molecules can anneal with nucleic acid primers in vivo. A model of dsT-DNA formation in plant cells is suggested.

Materials and Methods

DNA constructs. We first produced the pSAT1.35S.intGUS.35ST plasmid, in which the intron-containing GUS-encoding sequence (intGUS) was cloned between the dual 35S CaMV promoter and the 35S CaMV terminator and the entire plant expression cassette was flanked by *Asc*I sites. We also produced a binary vector, pRCS2-adapter, by replacing the multicloning site of pRCS2 (Tzfira *et al.*, 2005; Dafny-Yelin and Tzfira, 2007) with a single *Asc*I site. We next transferred the intGUS expression cassette into the *Asc*I site of the pRCS2-adapter, producing the *Agrobacterium*'s binary vectors with sense (S) and antisense (AS) intGUS expression cassettes. We then produced 5'- and 3'-truncated intGUS expression cassettes by PCR on pSAT1.35S.intGUS.35ST template using different primer pairs. S5': 5'-GGAATTCTGTGCCAGGCAGTTTTAACGA and 5'-TCCCGAGGGGAACCCTGTGG; AS3' and AS5': 5'-AGGCGCGCCTCGTCCACCCAGGTGTT and 5'-TCCCGAGGGGAACCCTGTGG. The truncated intGUS expression cassettes were next transferred as *EcoR*I and *Asc*I (S5') or *Asc*I (AS3' and AS5') into the same sites of the pRCS2-adapter, producing the *Agrobacterium* binary vectors S5', AS3' and AS5'. The T-RFP binary vector was constructed by cloning the EYFP-CHS (i.e. yellow fluorescent protein [EYFP] tagged with endoplasmic reticulum [ER]-bound chalcone synthase [CHS]) and the RFP-NLS (i.e. the red fluorescent protein [RFP] tagged with a nuclear localization signal [NLS]) expression cassettes (Tzfira *et al.*, 2005) into a single binary vector. More specifically, the EYFP-CHS and RFP-NLS expression cassettes driven by dual 35S promoter were subsequently cloned as PI-*Psp*I and I-*Ceu*I fragments into the I-*Ceu*I and I-*Sce*I sites of pRCS2 (Tzfira *et al.*, 2005; Dafny-Yelin and Tzfira, 2007), respectively. To construct T-mutRFP-1 and T-mutRFP-2, we first performed

PCR on T-RFP template using primers with extra 1 or 11 nt. T-mutRFP-1: 5'-AACTGCAGGCGAGGAGTCCTGGGTCACGGTCACCA and 5'-AACTGCAGGCATCTTGGCTATTGGCACT; T-mutRFP-2: 5'-AACTGCAGGAACCAATTCATGAGGAGTCCTGGGTCACGGTCACCA and 5'-AACTGCAGGCATCTTGGCTATTGGCACT. The PstI-digested PCR products were then cloned into the *Sbf*I sites of the T-RFP plasmid, replacing the wild-type RFP-coding sequence. To produce cRFP and anti-cRFP, a small portion of the wild-type RFP CDS was amplified by PCR using different primer pairs. cRFP: 5'-GGCGCGCCCTTGGCCATGTAGGTGGTCT and 5'-GGAATTCCGAGGACGTCATCAAGGAGT; anti-cRFP: 5'-GGCGCGCCCTTGGCCATGTAGGTGGTCT and 5'-GGCGCGCCCGAGGACGTCATCAAGGAGT-3'. PCR fragments were then transferred as *EcoR*I and *Asc*I (cRFP) or *Asc*I (anti-cRFP) into the same sites of the pRCS2-adapter. mut-RFP was produced by PstI digestion, Klenow overhang removal and re-ligation of cRFP. pSAT6.35SP.ECFP-C1.35ST, in which a ECFP-encoding sequence was cloned between the dual 35S CaMV promoter and the 35S CaMV terminator, was used to monitor biolistic transfection efficiency (Dafny-Yelin and Tzfira, 2007).

Transient expression. Binary vectors were transferred into *A. tumefaciens* strain EHA105. For the T-DNA repair assay, *Agrobacterium* cells were grown overnight at 28 ℃ in YEP medium supplemented with 80 μg ml^{-1} spectinomycin and 200 μg ml^{-1} streptomycin, collected by centrifugation for 10 min at 3000*g* and resuspended in induction medium [10.5 g L^{-1} K_2HPO_4, 4.5 g L^{-1} KH_2PO_4,1 g L^{-1} $(NH_4)_2PO_4$, 0.5 g L^{-1} sodium citrate, 1 g L^{-1} glucose, 1 g L^{-1} fructose, 4 g L^{-1} glycerol, 120 mg L^{-1} $MgSO_4$, 2 g L^{-1} 2-(N-morpholine)-ethanesulfonic acid

(MES), pH 5.6] supplemented with 100 μM acetosyringone and antibiotics. They were then grown to an OD_{600} of 0.6, again collected by centrifugation for 10 min at 3000g and resuspended in infiltration medium (10 mM $MgSO_4$,10 mM MES, pH 5.6) supplemented with 200 μM acetosyringone to an OD_{600} of 0.2. The *Agrobacterium* infiltration of *N. tabacum* adopted protocol reported by Van der Hoorn *et al*. (2000). More specifically, *Agrobacterium* cultures were infiltrated using a 2-ml sterile syringe into fresh leaves of 4- to 5-week-old tobacco plants. Equal volumes of the bacterial cultures with the same cell concentration were used in all infiltration experiments within the same panel in **Figure 2-1**. Cell-number ratio of *Agrobacterium* was 1:1 between T-RFP and T-mutRFP-1 and 1:9 between T-mutRFP-1 and cRFP/anti-cRFP in **Figure 2-3.** Tobacco explants were collected at the specified time or 48 h (if not specified) after infiltration. Explants were then subjected to GUS-staining analysis or confocal fluorescence microscopy. GUS-staining analysis was performed according to Jefferson *et al.* (1987). More specially, infected explants were vacuum infiltrated with the chromatogenic substrate X-Gluc and incubated at 37 ℃ overnight. Explants were then de-stained with 70% ethanol at room temperature for several days.

Molecular analysis of infected plants. Total plant DNA was extracted from leaves of infected tobacco 48 h after *Agrobacterium* infiltration using the phenol-chloroform method. The region surrounding T-RFP and mut-cRFP hybridization was amplified by PCR using primers 5'-CCAGCTTGATGTCGGTCT-3' and 5'-GATAGCCATGGCCTCCT-3'. The resulting PCR fragment was digested with PstI for 4 h. DNA digestion was visualized by standard agarose gel electrophoresis.

Biolistic transfection. Biolistic transfection was carried out using the Bio-Rad Helios Gene Gun System as instructed by the manufacturer (http://www.bio-rad.com). The transfection control plasmid pSAT6.35SP.ECFP-C1.35ST was prepared using the Qiagen Plasmid MidiPrep Kit. DNA and RNA oligonucleotides used in the biolistic transfection were purchased from IDT (Coralville, IA). Plasmid (5 μg) and oligonucleotides (25 μg) were fully mixed before precipitation on the gold bullets. Fresh leaves from 4- to 5-week-old tobacco plants were used in this experiment.

Confocal microscopy. Plant tissue was viewed directly with a confocal laser scanning microscope (TCS SP5; Leica). EYFP was excited with an argon laser at 514 nm, and fluorescence was monitored between 520 and 542 nm. RFP was excited with a helium-neon laser at 561 nm, and fluorescence was monitored between 607 and 642 nm. ECFP was excited with an argon laser at 458 nm, and fluorescence was monitored between 465 and 514 nm. Chlorophyll fluorescence was monitored above 700 nm. Sequential scanning was implemented to detect different fluorescence.

Results

Detecting in vivo formation of dsT-DNAs by functional assay. Given that the nature of T-strand conversion to a double-stranded form is a process of DNA polymerization, it may be initiated by T-strand priming by short plant primers. Such primers can be de-novo-synthesized RNA primers or endogenous pre-existing ssDNA or RNA fragments. Because priming may occur at random locations along the T-strand and DNA polymerization follows the 5' to 3' direction, the resulting dsT-DNA intermediates may remain single-stranded at the ends, and in particular at their 3' ends. These single-stranded regions are susceptible to nuclease degradation, potentially leading to truncations of dsT-DNA molecules, in particular at their 3' ends. Indeed, sequencing data of T-DNA integration sites have revealed that the 3' end of a T-DNA is more vulnerable to deletion than its 5' end (Alonso *et al.*, 2003; Chen *et al.*, 2003; Mayerhofer *et al.*, 1991; Gheysen *et al.*, 1991).We hypothesized that truncation of T-DNA ends (in particular at the 3' end) may arise, at least partially, from incomplete T-strand to double-stranded conversion, which occurs prior to the integration step. Notably, truncation of T-DNA ends (especially the 3' end) may also occur during T-DNA integration, as suggested by Tinland *et al*. (1995) in their ssT-DNA integration model.

We tested our hypothesis by monitoring the formation of dsT-DNA intermediates using a functional assay in which transient expression of the GUS reporter gene occurs independently of T-DNA integration, in a manner similar to this assay's previous use to determine that T-DNAs are transferred as single-stranded, but not double-stranded molecules into the plant cell nucleus (Tinland *et al.*, 1994). In our assay, two binary vectors were engineered to produce either sense (S, **Figure 2-1A**) or antisense (AS, **Figure 2-1A**) T-strand molecules of the same functional

GUS expression cassette. The structure of our GUS expression cassettes included a long 886-bp dual 35S CaMV promoter and a short 211-bp 35S CaMV terminator. In the design of our vectors, even moderate deletions at the 3' end are likely to interfere with GUS expression in sense T-strand molecules because of the relatively short terminator region (ca 350 bp from the beginning of the terminator to the end of the T-DNA's border region); however, because a single 35S CaMV promoter, and even just a minimal promoter sequence (Odell *et al.*, 1985), is sufficient to produce GUS expression, larger deletions (i.e. longer than 750 bp, in our constructs) at the 3' end will be required to interfere with GUS expression in antisense T-strand molecules. This may lead to a difference in GUS expression between transfected sense and antisense T-strands. In fact, infection of tobacco leaves with an identical concentration of *Agrobacterium* cells carrying sense or antisense T-strand molecules showed that GUS expression from antisense T-strands is more efficient than from sense T-strands (compare AS with S, **Figure 2-1B**). This observation is compatible with the hypothesized 3' deletion of pre-integrated T-DNA, implicating an alternative mechanism for the frequently observed 3' truncation of integrated T-DNA. This possible mechanism emphasizes the notion that the double-stranded conversion stabilizes T-DNA by inhibiting nuclease degradation of its single-stranded form; whereas, the random annealing nature of the primers along the T-strand may not be able to initiate its complete conversion, leaving the non-complementary portion, more likely to be the 3' end, susceptible to degradation. Note that aside from the possible asymmetrical double-stranded conversion of the 5' and 3' ends, other factors, such as suggested VirD2 protection of the 5' end (Durrenberger *et al.*, 1989), might also contribute to the unequal stability of these two ends. Interestingly, GUS expression in leaves which were co-infected with both sense and antisense

Agrobacterium strains, each at half concentration, was more efficient than in leaves infected by individual strains at full concentration (compare S+AS with S or AS, **Figure 2-1B**). It has been shown that GUS expression can be restored by hybridization of two putative T-strand molecules[2] and we thus suggested that in these co-infection experiments, the sense and antisense T-strand molecules, which may have been stripped of their escorting VirE2 molecules (Tzfira *et al.*, 2004b; Zaltsman *et al.*, 2010), hybridized with each other and produced an artificial dsT-DNA, most likely with higher efficiency and stability than the natural conversion of individual T-strands.

In vivo hybridization of single stranded T-DNAs. In vivo annealing of two partially complementary T-strands or a single T-strand carrying inverted repeats was reported, in support of the notion that T-DNA is transferred in single-stranded form from *Agrobacterium* to it host[2]. We thus took advantage of such T-strand annealing to unequivocally test whether annealed T-strands can serve as primers (and by implication as DNA and RNA oligonucleotides) to initiate the formation of dsT-DNA. Note that in our design (S5' and AS3', **Figure 2-1A**), we intentionally infected cells with two T-strands, each with polymerase-extendable 3' ends. This design allowed us to investigate the elongation of the priming T-strand by plant DNA polymerases using the annealed T-strand as template. We infected tobacco leaves with T-DNA molecules which were engineered to carry non-functional portions of the GUS expression cassette (S5' and AS3', **Figure 2-1A**). As expected, no GUS expression was observed in these leaf samples (S5' and AS3', **Figure 2-1C**). However, GUS expression was clearly observed in leaf samples co-infected with two independent *Agrobacterium* strains, each carrying a different truncated T-DNA molecule (compare the combination of S5'+AS3' with S5' or AS3', **Figure 2-**

1C). The restoration of GUS expression was likely caused by in vivo annealing and subsequent DNA polymerization of two truncated T-strands. Alternatively, it might have derived from homologous recombination between dsS5' and dsAS3' intermediates. To rule out the latter possibility, we produced an antisense version of S5', designated AS5' (**Figure 2-1A**), in which a truncated S5' molecule was delivered in an antisense orientation, and we used it in co-infection experiments with AS3' (**Figure 2-1A**). **Figure 2-1D** shows that whereas clear GUS expression was observed in leaf samples co-infected with strains S5' and AS3', no GUS expression was observed in leaf samples co-infected with strains AS5' and AS3'. These observations clearly support the notion that the detected reconstitution of a functional GUS-expressing T-DNA from truncated T-strand molecules is derived from annealing of the sense and antisense T-strands and not from recombination between two homologous dsT-DNA molecules. Importantly, they also suggest that the 3' end of a complementary ssDNA fragment can serve as an initiation point for DNA polymerization, using the primary strand (i.e. T-strand) as template. Note that while we did not detect homologous recombination between dsS5' and dsAS3' intermediates, we cannot exclude the occurrence, albeit at low frequency, of such events.

T-strand-repair assay unveils priming of T-strand molecules. To further investigate the notion that short nucleic acid fragments are capable of priming T-strands, and the function of these fragments as initiation points for polymerization of the nascent strand, we developed a T-strand-repair assay which allowed monitoring of the annealing and elongation of an artificial T-strand primer (cRFP, **Figure 2-2A,B**) and at least one more, endogenous, plant primer on the target T-strand (T-RFP, **Figure 2-2A,B**). The T-strand molecule, T-RFP, was engineered to carry two independent and functional constitutive plant expression cassettes. The first cassette, which

was cloned near the T-DNA's right border, expressed the EYFP protein tagged with endoplasmic reticulum (ER)-bound chalcone synthase (CHS), while the second cassette expressed the RFP protein tagged with nuclear localization signal (NLS). We used confocal fluorescence microscopy to monitor T-strand conversion to dsT-DNA in *Agrobacterium*-infected tobacco leaves. As expected, T-RFP-infected cells exhibited both ER-bound EYFP and nuclear RFP signals (T-RFP, **Figure 2-3**). We next produced two mutated variants of T-RFP, designated T-mutRFP-1 and T-mutRFP-2, in which the RFP-coding sequence was interrupted with a 1-nt or 11-nt long insertion, respectively (**Figure 2-2B**). As expected, these mutations interfered with RFP, but not with EYFP expression in infected leaves (T-mutRFP-1 and T-mutRFP-2, **Figure 2-3**). We also constructed a T-strand molecule, cRFP (**Figure 2-2A,B**), which was engineered to carry a partial wild-type *RFP* sequence (ca 500 nucleotides long) that is complementary to the mutated *RFP* region of T-mutRFP-1 and T-mutRFP-2. An antisense version of the cRFP molecule, anti-cRFP, was also constructed (**Figure 2-2B**). Annealing of cRFP to T-mutRFP-1 and T-mutRFP-2 creates a bulging or mismatched DNA region (Lyons and O'Brien, 2010) of which DNA repair may restore the wild-type *RFP*-coding sequence in T-mutRFP-1 and T-mutRFP-2 if cRFP was recognized as the repair template. Co-transformation of T-mutRFP-1 or T-mutRFP-2 with cRFP resulted in restoration of RFP expression in some of the infected cells (**Figure 2-2**). However, RFP expression was not restored when T-mutRFP-1 was co-transformed with anti-cRFP (**Figure 2-2**). These observations indicated that restoration of RFP expression is derived from annealing between complementary T-strand molecules (e.g. between T-mutRFP-1 and cRFP) but not by putative recombination between homologous dsT-DNA intermediates. Note that in addition to the short wild-type *RFP* sequence, cRFP also carries parts of the binary

vector sequence (ca 100 nucleotides long) on both ends, including the remaining border sequence and the junction sequence between the border and the *RFP* sequence (**Figure 2-2A**). Therefore, restoration of RFP expression could only occur if the cRFP molecule's left and right border tails had been trimmed and the DNA bulge between cRFP and T-mutRFP-1 or T-mutRFP-2 had been repaired. Efficient restoration of RFP expression in our assay strongly suggested that T-strand hybrids were substrates for the plant DNA repair machinery, thus supporting the notion that plant DNA repair machinery is actively involved in T-strand to dsT-DNA conversion. Finally, because the restoration of RFP-NLS expression also depended on the complete double-stranded conversion of the RFP expression cassette, our data indicate that at least one other endogenous plant primer must have annealed to the 3' end region of the T-mutRFP-1 and T-mutRFP-2 (**Figure 2-2A**), initiating another continuous T-strand double-stranded conversion upstream of the cRFP annealing region. Our data also implicate the possible involvement of plant DNA ligase(s) in ligating the gaps of newly synthesized DNA fragments, similar to ligation of Okazaki fragments in lagging-strand DNA synthesis, forming consecutive dsT-DNA molecules.

To confirm that DNA repair had indeed occurred in the bulging DNA region between the two T-strands, we engineered a mut-cRFP T-strand molecule carrying a 4-nt deletion which, upon annealing to T-RFP, abolishes the *Pst*I site on the wild-type RFP coding sequence (**Figure 2-2B**). To detect the mutation by DNA repair, we designed a pair of primers that anneal to the *RFP* sequence surrounding, but outside of, the cRFP or mut-cRFP carrying sequence and are thus only able to amplify T-RFP, but not cRFP or mut-cRFP T-DNAs. Using this pair of primers, PCR amplification of mixtures of T-RFP with cRFP or mut-cRFP binary plasmids, as well as mixtures of *Agrobacterium* strains carrying T-RFP and cRFP or mut-cRFP binary plasmids,

resulted in PstI-digestible products (**Figure 2-2C**). In addition, PCR amplification of plants co-infected with T-RFP and cRFP resulted in PstI-digestible products (**Figure 2-2C**). However, PCR amplification of plants co-infected with T-RFP and mut-cRFP yielded a PstI-indigestible band which was similar in length to the undigested T-RFP PCR product (**Figure 2-2C**). Sequencing analysis of the DNA recovered from this band confirmed the bona fide deletion in the *PstI* site. We should note that PstI-indigestible PCR products could have been amplified from either integrated or non-integrated T-DNAs. Given the relative amounts of PstI-indigestible and PstI-digestible products in the lane (plant: RFP & mut-cRFP), it is likely that most, if not all of the PstI-indigestible product stemmed from non-integrated T-DNAs.

In vivo priming of T-strand molecules by synthetic primers. To further substantiate our hypothesis that DNA polymerization of the nascent strand is initiated by priming of short DNA and RNA fragments, we mimicked the putative natural process of T-strand conversion by testing whether short DNA and RNA oligonucleotides can serve as primers for this conversion. We used our T-strand-repair assay (**Figure 2-2A**), except that this time we supplied T-strand-transformed cells with short complementary DNA or RNA oligonucleotides rather than a second T-strand molecule. We first infected tobacco leaves with *Agrobacterium* strain T-mutRFP-1 (**Figure 2-4A,B**) and incubated the leaves overnight. Then we bombarded the infected leaves with different short DNA oligonucleotides (e.g. primer I, **Figure 2-4B**) and ECFP-expressing plasmid, and left the leaves to incubate for 24 h. DNA oligos can anneal to T-mutRFP-1 in the region where the 1-nt insertion located (**Figure 2-4A,B**). We then used confocal fluorescence microscopy to detect the priming of the DNA oligonucleotides to the T-strand by monitoring the restoration of nuclear RFP expression. EYFP expression was an indicator of successful

Agrobacterium-mediated transformation, implying the presence of T-strand T-mutRFP-1 molecules in the EYFP-expressing cells. ECFP expression was used to determine cells in which successful biolistic transformation had occurred, implying that short DNA oligonucleotides had been delivered into those cells. No RFP-NLS-expressing cells were observed in T-mutRFP-1-infected leaves bombarded with only the ECFP plasmid (**Figure 2-4C**). RFP-NLS-expressing cells were observed in T-mutRFP-1-infected leaves co-bombarded with ECFP and DNA primer I. The percentage of RFP-NLS-expressing cells in ECFP and EYFP-CHS co-expressing cells was low and we suggest that this can be attributed to the high sensitivity of DNA oligonucleotides to endo- and exonucleases within the transfected cells. We therefore used modified DNA oligonucleotides (i.e. primer II, **Figure 2-4B**) in which several nucleotides were modified with phosphorothioate, which confers higher resistance to nuclease attack. As expected, chemical modification of the DNA oligonucleotides resulted in a higher frequency of RFP-NLS recovery, and 86.1% of the ECFP and EYFP co-expressing cells also expressed RFP-NLS (**Figure 2-4C,D**). Next, we produced primer III oligonucleotide which included, in addition to the phosphorothioates, an inverted dT nucleotide at its 3' end (**Figure 2-4B**). This modification is expected to inhibit elongation of the 3' end of the oligonucleotide by DNA polymerases and by implication, RFP-NLS recovery. Indeed, we observed a lower frequency (59.9%) of RFP-NLS-expressing cells among the ECFP and EYFP co-expressing cells (**Figure 2-4C,D**). The difference in RFP-NLS restoration between DNA primer II and primer III was significant ($P < 0.01$), suggesting the provided oligonucleotides can act as substrates of the yet-to-be identified plant DNA polymerase(s) in our T-strand-repair assay. The fact that T-strand conversion in T-mutRFP-1-infected and primer III-bombarded cells was less efficient than in their primer II-

bombarded counterparts, yet not completely blocked, further indicated involvement of the plant DNA repair machinery in detecting and resolving primer-template mismatch during the process of dsT-DNA formation.

We next tested whether RNA primers can also anneal with T-strand molecules and initiate their double-stranded conversion. We bombarded tobacco leaves infected with *Agrobacterium* as above with RNA primers that had the same sequence and chemical modulation as primer II (**Figure 2-4B**). Since RNA primers are highly susceptible to contaminating nucleases, we omitted ECFP-expressing plasmid from the bombardment to avoid potential RNA degradation by residual RNase from the plasmid preparation. Confocal fluorescence microscopy analysis of the infected cells revealed that while most of them exhibited only EYFP-CHS expression, 20 individual cells, as shown, for example, in **Figure 2-4E**, also exhibited RFP-NLS expression. No RFP-NLS-expressing cells were detected in leaf samples when an empty bullet was used in the bombardment. Because of the instability of the RNA primer and the extremely low efficiency of putative RNA-mediated DNA repair (Storici *et al.*, 2007), we could only detect about 1 or 2 events per leaf sample when using RNA primer II. Whereas, our observations indicated that the RNA primers like DNA primers can also anneal to the incoming T-strand during its conversion to dsT-DNA. Collectively, results of our T-strand-repair assays indicated that T-strand molecules are subjected to DNA repair by the plant machinery, which is active during the conversion to dsT-DNA molecules.

Discussion

Taken together, our data show that T-strand molecules, which are likely to be stripped of their coating proteins prior to integration (Tzfira *et al.*, 2004b; Zaltsman *et al.*, 2010), can interact with DNA and RNA molecules in living cells. Our observations thus suggest a mechanism for T-strand conversion to a dsT-DNA molecule (**Figure 2-5**) in which RNA primers are synthesized de novo in random positions along the T-strand template, and/or endogenous pre-existing ssDNA/RNA fragments randomly anneal to the T-strand. These nucleic acid molecules then serve as primers in the polymerization of the T-DNA nascent strand. Short DNA fragments may derive from DNA fragmentation during the occurrence of genomic DSBs, while RNA primers may be synthesized by primase or derive from various cellular pathways (e.g. the plant silencing and DNA polymerization machineries). Such naturally occurring primers may not accurately match the T-strands. It would be interesting to test whether mutations within the T-DNA can indeed arise from natural dsT-DNA formation via incorporation of these putative mismatched DNA/RNA primers, or some unknown mechanism that exists to prevent mutagenesis of T-strand molecules. Interestingly, the reconstruction of RFP-NLS-encoding sequences in mutated T-strands by various types of primers indicates that dsT-DNA molecules are likely subjected to various types of DNA repair by the plant machinery. Our observations thus also provide an alternative, mechanistic explanation for the frequently seen degeneration of the integrated T-DNA molecule's termini, especially the 3’ end. Combining our results with previously suggested T-DNA integration models (Tzfira *et al.*, 2004a), truncation of T-DNA likely occurs during its conversion to double stranded form, as shown here, and/or during integration, at time when plant DNA processing factors are activated. Studies have shown that

two T-DNA molecules deriving from different *Agrobacterium* cells often integrate at the same genomic location (De Buck *et al.*, 2009), and that various host proteins may mediate the interaction of T-complexes with plant nucleosomes (Lacroix *et al.*, 2008) and aid in stripping T-strands of their coating proteins. It would be intriguing to test whether T-strands can be actively directed to the integration sites where active DNA synthesis and repair are occurring and reagents (i.e. DNA or RNA fragments and dNTPs) are abundant for the conversion process.

The question remains whether T-strand complementation occurs in close proximity to integration sites and during the integration steps. It was well established that dsT-DNA molecules are potent substrates for genomic integration via DSB repair (Salomon and Puchta, 1998; Chilton and Que, 2003; Tzfira *et al.*, 2003). It is thus possible that T-strand conversion to dsT-DNA may be, at least in part, a limiting factor in T-DNA integration. Nevertheless, recent evidence suggests that dsT-DNA intermediates can undergo inter/intra-molecular cyclization and form circular molecules which may become incapable for integration (Singer *et al.*, 2012). Therefore, dsT-DNA formation and its accessibility to integration may be regulated in a more comprehensive manner.

We report here that oligonucleotides can efficiently interact with T-strand molecules in living cells. Thus, while the specific set of proteins involved in the process of T-strand conversion still needs to be identified, our observations suggest a possible strategy to enhance the rate of T-strand conversion to double-stranded form and by implication, to improve the efficiency of gene-replacement and site-specific-integration systems in crops and model plants.

Acknowledgments

All the contents in Chapter 2 have been published in 2013 in *Nat Commun* 4:2253 doi: 10.1038/ncomms3253 with the title: In vivo formation of double-stranded T-DNA molecules by T-strand priming. I am the sole first author in this project.

I would like to thank Tzvi Tzfira for his mentorship and help in design of the experiments and analysis of the data and Yoel Shiboleth, Vardit Zeevi, Dan Weinthal and Kamy Singer for their helpful discussion. I would also like to thank Dr. Lyle Simmons and Dr. Thomas E Wilson for their suggestions and support in the manuscript writing phase.

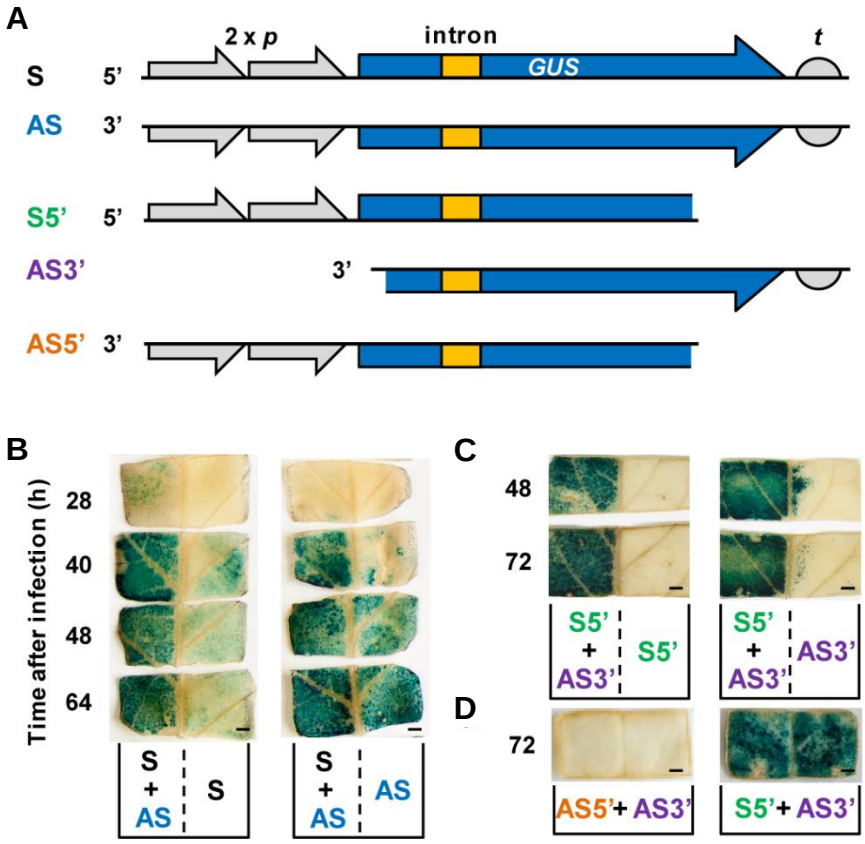


Figure 2-1. Detection of T-strand conversion to dsT-DNA by in-vivo annealing of complementary T-strands. (**A**) Schematic of several constructed T-strand molecules. S and AS molecules carry the sense or antisense strand of a functional GUS expression cassette, respectively. S5', AS3' and AS5' molecules carry the sense (S5') or antisense (AS3' and AS5') strand of two partial, non-functional, yet overlapping GUS expression cassettes, respectively. The GUS expression cassette contains an intron to prevent its expression in *Agrobacterium*. (**B**–**D**) Analysis of GUS expression in infected tobacco leaves at different time points after *Agrobacterium* infiltration. The *Agrobacterium* strains used for the infiltration are indicated in each sample. Bars are 2 mm in length. 2 x *p*, dual 35S promoter; *t*, 35S terminator; *GUS*, β-glucuronidase-encoding gene.

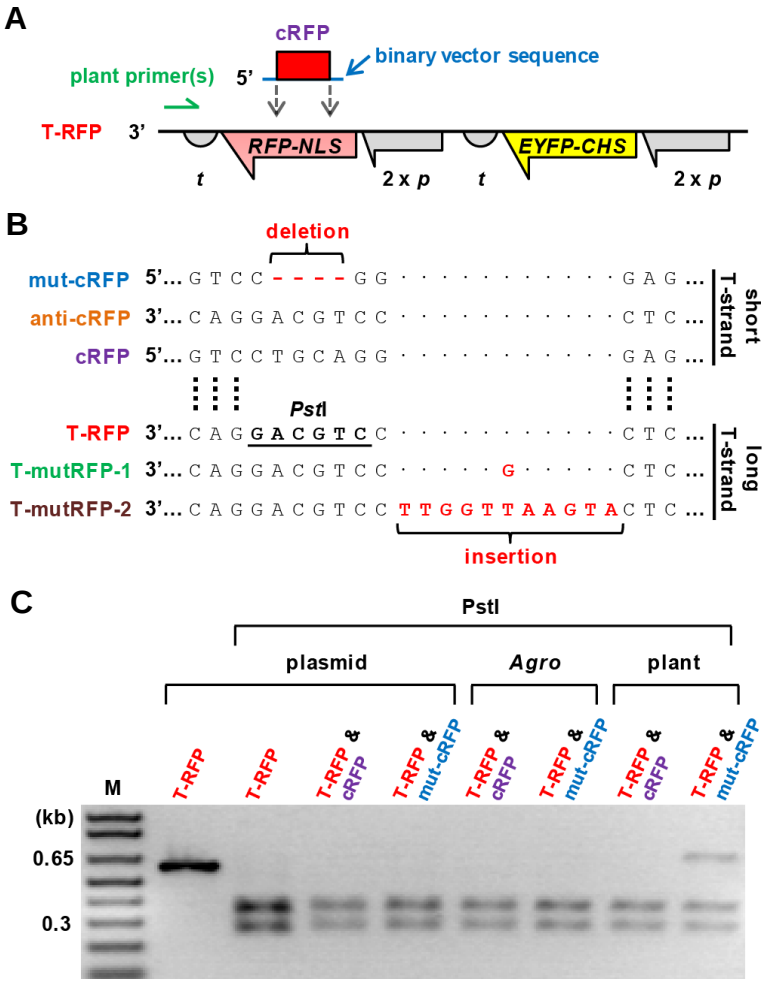


Figure 2-2. Detection of T-strand conversion to dsT-DNA by site-specific mutagenesis. (**A**) Schematic of in-vivo annealing of T-RFP and cRFP molecules. Putative endogenous primer(s) (i.e. plant primers) to initiate the double stranded conversion of T-RFP's 3' end are presented. (**B**) Partial sequences of several T-strand constructs in the *RFP* region bearing site-specific mutagenesis. T-RFP carries functional RFP-NLS and EYFP-CHS expression cassettes. T-mutRFP-1 and T-mutRFP-2 molecules carry a 1- or 11-nt insertion as indicated in the RFP-encoding sequence, respectively. cRFP carries a partial RFP-encoding sequence. anti-cRFP is the complementary version of cRFP. mut-cRFP carries the same sequence as cRFP except for a 4-nt deletion in the *Pst*I site, as indicated. cRFP, anti-cRFP and mut-cRFP serve as short T-strands which can potentially anneal with long T-strands in the plant nucleus. (**C**) PstI digestion analysis of PCR products from T-RFP-specific PCR amplification. Extracted total DNA from *Agrobacterium*-infected plant tissues and corresponding *Agrobacterium* strains and binary vector plasmids was used as templates for individual PCRs. 2 x *p*, dual 35S promoter; *t*, 35S terminator; M, DNA ladder.

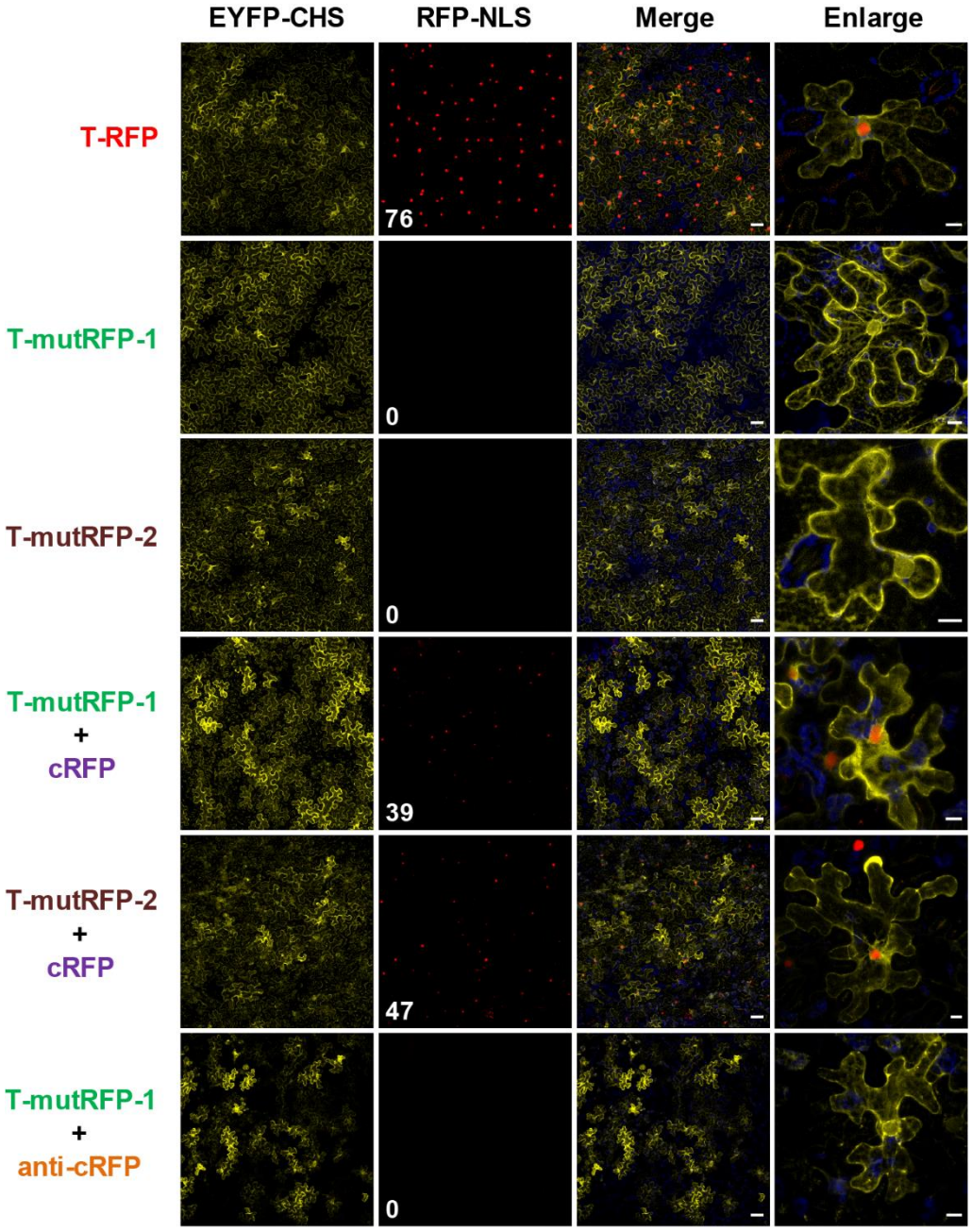


Figure 2-3. Confocal fluorescence microscopy analysis of in vivo formation of dsT-DNAs. *Agrobacterium* strains used for each infiltration are indicated on the left. From left to right: EYFP (yellow), RFP (red), merged image of EYFP, RFP and chlorophyll (dark blue) signals and enlargement of representative cell. Number of RFP-NLS-expressing cells is shown in the bottom left corner of each RFP-NLS image. Scale bars of merged images are 50 μm in length. Scale bars of enlarged images are 10 μm in length.

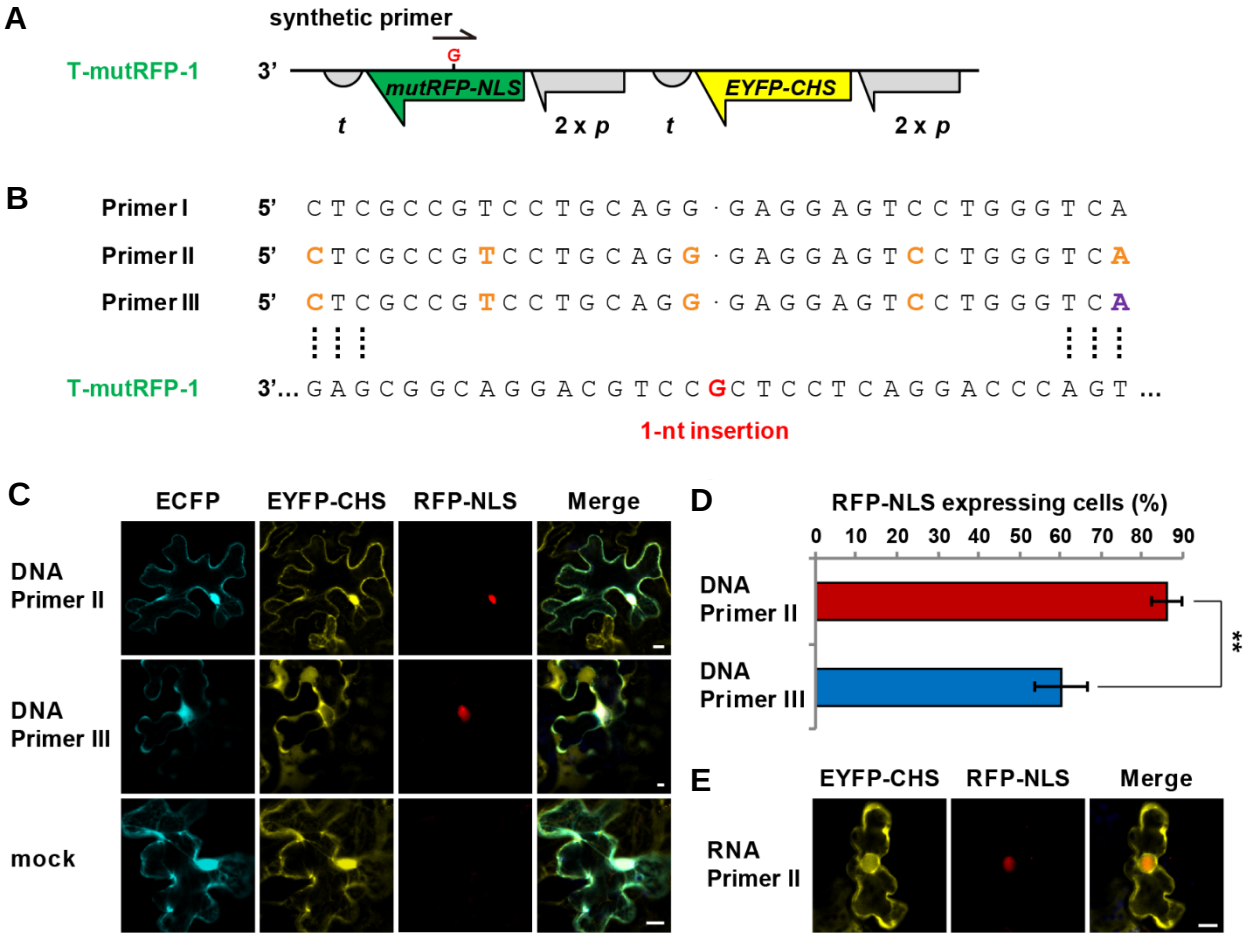


Figure 2-4. In vivo priming of T-strand molecules by synthetic primers. (**A**) Schematic in vivo annealing of T-mutRFP-1 with synthetic oligonucleotides. (**B**) Sequences and chemical modifications of DNA primers for biolistic transformation of tobacco. Primers carry wild-type RFP-encoding sequence in the region where T-mutRFP-1 has a 1-nt insertion (red). Orange bases have phosphorothioates modification. Purple base has an inverted dT nucleotide at its 3' end. (**C**) Representative images of *Agrobacterium* T-mutRFP-1-infected cells bombarded with a combination of primers and ECFP-expressing plasmid. Scale bars are 10 μm in length. (**D**) Frequency of RFP-NLS-expressing cells among ECFP and EYFP-CHS co-expressing cells. Data are presented as mean ± SD. (Significant difference at $P < 0.01$ in a two-tailed t-test between primer II and primer III assuming unequal variances). (**E**) Representative images of *Agrobacterium* T-mutRFP-1-infected cells bombarded with an RNA primer sharing the same sequence and chemical modifications as primer II. Scale bars are 10 μm in length.

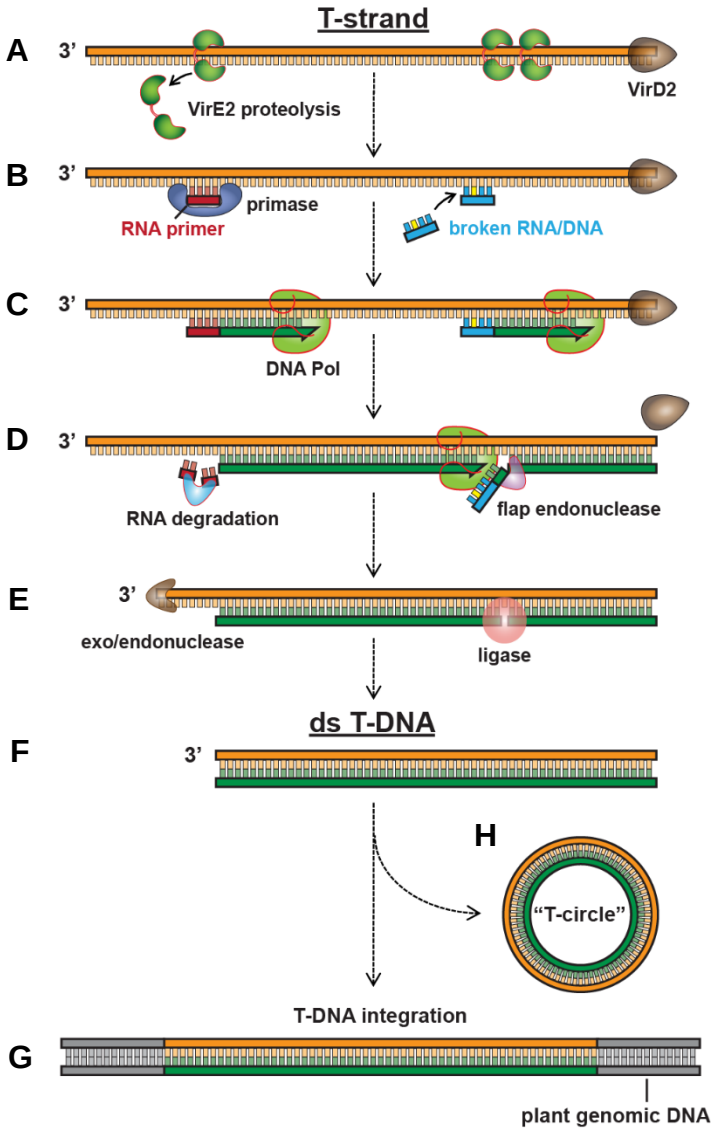


Figure 2-5. Hypothetical model of dsT-DNA formation in plant cells. (**A**) VirE2 proteins are removed from T-strand molecules by targeted proteolysis. (**B**) Primase synthesizes de-novo RNA primers using T-strand as the template, and endogenous pre-existing DNA/RNA fragments anneal to the T-strand. Base pair mismatch (shown in yellow) between the primer and the T-strand may lead to mutagenesis in the T-DNA sequence. (**C**) DNA/RNA primers initiate the nascent strand polymerization which is carried out by plant DNA polymerase(s). (**D**) DNA/RNA primer degradation. (**E**) The unconverted 3' end of the T-strand is sensitive to nuclease degradation. (**F-H**) 3'-truncated dsT-DNA molecules integrate into plant genomic double-strand breaks (**G**), or ligate intra/inter-molecularly forming the recently identified "T-circle" molecules (Singer *et al.*, 2012) (**H**).

Reference

Abu-Arish, A., Frenkiel-Krispin, D., Fricke, T., Tzfira, T., Citovsky, V., Wolf, S.G., and Elbaum, M. (2004). Three-dimensional reconstruction of Agrobacterium VirE2 protein with single-stranded DNA. J Biol Chem *279*, 25359-25363.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., *et al.* (2003). Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science *301*, 653-657.

Chen, S., Jin, W., Wang, M., Zhang, F., Zhou, J., Jia, Q., Wu, Y., Liu, F., and Wu, P. (2003). Distribution and characterization of over 1000 T-DNA tags in rice genome. Plant J *36*, 105-113.

Chilton, M.D., and Que, Q. (2003). Targeted integration of T-DNA into the tobacco genome at double-stranded breaks: new insights on the mechanism of T-DNA integration. Plant Physiol *133*, 956-965.

Citovsky, V., Wong, M.L., and Zambryski, P. (1989). Cooperative interaction of Agrobacterium VirE2 protein with single-stranded DNA: implications for the T-DNA transfer process. Proc Natl Acad Sci U S A *86*, 1193-1197.

Dafny-Yelin, M., and Tzfira, T. (2007). Delivery of multiple transgenes to plant cells. Plant Physiol *145*, 1118-1128.

De Buck, S., Podevin, N., Nolf, J., Jacobs, A., and Depicker, A. (2009). The T-DNA integration pattern in Arabidopsis transformants is highly determined by the transformed target cell. Plant J *60*, 134-145.

Durrenberger, F., Crameri, A., Hohn, B., and Koukolikova-Nicola, Z. (1989). Covalently bound VirD2 protein of Agrobacterium tumefaciens protects the T-DNA from exonucleolytic degradation. Proc Natl Acad Sci U S A *86*, 9154-9158.

Gheysen, G., Villarroel, R., and Van Montagu, M. (1991). Illegitimate recombination in plants: a model for T-DNA integration. Genes Dev *5*, 287-297.

Grange, W., Duckely, M., Husale, S., Jacob, S., Engel, A., and Hegner, M. (2008). VirE2: a unique ssDNA-compacting molecular machine. PLoS Biol *6*, e44.

Howard, E.A., Winsor, B.A., De Vos, G., and Zambryski, P. (1989). Activation of the T-DNA transfer process in Agrobacterium results in the generation of a T-strand-protein complex: Tight association of VirD2 with the 5' ends of T-strands. Proc Natl Acad Sci U S A *86*, 4017-4021.

Jefferson, R.A., Bevan, M., and Kavanagh, T. (1987). The use of the Escherichia coli beta-glucuronidase as a gene fusion marker for studies of gene expression in higher plants. Biochem Soc Trans *15*, 17-18.

Kim, S.I., Veena, and Gelvin, S.B. (2007). Genome-wide analysis of Agrobacterium T-DNA integration sites in the Arabidopsis genome generated under non-selective conditions. Plant J *51*, 779-791.

Lacroix, B., Loyter, A., and Citovsky, V. (2008). Association of the Agrobacterium T-DNA-protein complex with plant nucleosomes. Proc Natl Acad Sci U S A *105*, 15429-15434.

Lyons, D.M., and O'Brien, P.J. (2010). Human base excision repair creates a bias toward -1 frameshift mutations. J Biol Chem *285*, 25203-25212.

Mayerhofer, R., Koncz-Kalman, Z., Nawrath, C., Bakkeren, G., Crameri, A., Angelis, K., Redei, G.P., Schell, J., Hohn, B., and Koncz, C. (1991). T-DNA integration: a mode of illegitimate recombination in plants. EMBO J *10*, 697-704.

Meza, T.J., Stangeland, B., Mercy, I.S., Skarn, M., Nymoen, D.A., Berg, A., Butenko, M.A., Hakelien, A.M., Haslekas, C., Meza-Zepeda, L.A., *et al.* (2002). Analyses of single-copy Arabidopsis T-DNA-transformed lines show that the presence of vector backbone sequences, short inverted repeats and DNA methylation is not sufficient or necessary for the induction of transgene silencing. Nucleic Acids Res *30*, 4556-4566.

Mysore, K.S., Bassuner, B., Deng, X.B., Darbinian, N.S., Motchoulski, A., Ream, W., and Gelvin, S.B. (1998). Role of the Agrobacterium tumefaciens VirD2 protein in T-DNA transfer and integration. Mol Plant Microbe Interact *11*, 668-683.

Odell, J.T., Nagy, F., and Chua, N.H. (1985). Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. Nature *313*, 810-812.

Puchta, H. (2003). Towards the ideal GMP: homologous recombination and marker gene excision. J Plant Physiol *160*, 743-754.

Salomon, S., and Puchta, H. (1998). Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. EMBO J *17*, 6086-6095.

Singer, K., Shiboleth, Y.M., Li, J., and Tzfira, T. (2012). Formation of complex extrachromosomal T-DNA structures in Agrobacterium tumefaciens-infected plants. Plant Physiol *160*, 511-522.

Storici, F., Bebenek, K., Kunkel, T.A., Gordenin, D.A., and Resnick, M.A. (2007). RNA-templated DNA repair. Nature *447*, 338-341.

Tinland, B., Hohn, B., and Puchta, H. (1994). Agrobacterium tumefaciens transfers single-stranded transferred DNA (T-DNA) into the plant cell nucleus. Proc Natl Acad Sci U S A *91*, 8000-8004.

Tinland, B., Schoumacher, F., Gloeckler, V., Bravo-Angel, A.M., and Hohn, B. (1995). The Agrobacterium tumefaciens virulence D2 protein is responsible for precise integration of T-DNA into the plant genome. EMBO J *14*, 3585-3595.

Tzfira, T., Frankman, L.R., Vaidya, M., and Citovsky, V. (2003). Site-specific integration of Agrobacterium tumefaciens T-DNA via double-stranded intermediates. Plant Physiol *133*, 1011-1023.

Tzfira, T., Li, J., Lacroix, B., and Citovsky, V. (2004a). Agrobacterium T-DNA integration: molecules and models. Trends Genet *20*, 375-383.

Tzfira, T., Tian, G.W., Lacroix, B., Vyas, S., Li, J., Leitner-Dagan, Y., Krichevsky, A., Taylor, T., Vainstein, A., and Citovsky, V. (2005). pSAT vectors: a modular series of plasmids for autofluorescent protein tagging and expression of multiple genes in plants. Plant Mol Biol *57*, 503-516.

Tzfira, T., Vaidya, M., and Citovsky, V. (2004b). Involvement of targeted proteolysis in plant genetic transformation by Agrobacterium. Nature *431*, 87-92.

Tzfira, T., Weinthal, D., Marton, I., Zeevi, V., Zuker, A., and Vainstein, A. (2012). Genome modifications in plant cells by custom-made restriction enzymes. Plant Biotechnol J *10*, 373-389.

Van der Hoorn, R.A., Laurent, F., Roth, R., and De Wit, P.J. (2000). Agroinfiltration is a versatile tool that facilitates comparative analyses of Avr9/Cf-9-induced and Avr4/Cf-4-induced necrosis. Mol Plant Microbe Interact *13*, 439-446.

Young, C., and Nester, E.W. (1988). Association of the virD2 protein with the 5' end of T strands in Agrobacterium tumefaciens. J Bacteriol *170*, 3367-3374.

Zaltsman, A., Krichevsky, A., Loyter, A., and Citovsky, V. (2010). Agrobacterium induces expression of a host F-box protein required for tumorigenicity. Cell Host Microbe *7*, 197-209.

Ziemienowicz, A., Merkle, T., Schoumacher, F., Hohn, B., and Rossi, L. (2001). Import of Agrobacterium T-DNA into plant nuclei: two distinct functions of VirD2 and VirE2 proteins. Plant Cell *13*, 369-383.

CHAPTER 3

Assembly of Multigene Vectors for *Agrobacterium*-Mediated Plant Transformation and Generation of the Potent Anti-Malarial Drug Artemisinin in Tobacco

Abstract

Binary vectors are an indispensable component of modern *Agrobacterium*-mediated plant genetic transformation systems. A remarkable variety of binary plasmids has been developed to support the cloning and transfer of foreign genes into plant cells. The majority of these systems is, however, limited to the cloning and transfer of just a single gene of interest. Thus, plant biologists and biotechnologists face a major obstacle when planning the introduction of multigene traits into transgenic plants. Here we describe the assembly of multi-transgene binary vectors by using a combination of engineered zinc finger nucleases (ZFNs) and homing endonucleases. Our system is composed of a modified binary vector that has been engineered to carry an array of unique recognition sites for ZFNs and homing endonucleases and a family of modular satellite (SAT) vectors. By combining the use of designed ZFNs and commercial restriction enzymes, multiple plant expression cassettes were sequentially cloned into the acceptor binary vector. Using this system, we produced binary vectors that carried up to nine genes. *Arabidopsis* protoplasts and plants were transiently and stably transformed, respectively, by several multigene constructs and the expression of the transformed genes was monitored across several generations.

To further demonstrate the advantages of our multigene transformation system, we assembly binary vectors that deliver five transgenes which encode catalytic enzymes in the metabolic

biosynthesis pathway of the potent anti-malarial drug artemisinin. We confirm the artemisinin production from transgenic tobacco plants with the integration of the delivered multigenes.

Because ZFNs can potentially be engineered to digest a wide variety of target sequences, our system allows overcoming the problem of the very limited number of commercial homing endonucleases. Thus, users of our system can enjoy a rich resource of plasmids that can be easily adapted to their various needs and since our cloning system is based on ZFN and homing endonucleases, it may be possible to reconstruct other type of binary vectors and adapt our vectors for cloning on multigene vectors systems in various binary plasmids.

Introduction

To date, biotechnological improvement of plant species has largely been limited to the introduction of single novel traits into the genomes of target plants. However, many agronomic traits may depend on complex interactions between several proteins, and biotechnological improvement of a particular species may thus require the delivery and expression of whole, complex metabolic pathways (Halpin *et al*., 2001; Daniell and Dhingra, 2002; Lyznik and Dress, 2008; Naqvi *et al*., 2010).

For example, the emergence of multidrug-resistant strains of *Plasmodium* spp., the etiological agent of malaria, constitutes a major threat to controlling the disease (Ro *et al.*, 2006). Artemisinin, a natural compound from *Artemisia annua* (sweet wormwood) plants, is highly effective against drug-resistant malaria. Even so, low cost artemisinin-based drugs are lacking because of the high cost of obtaining natural or chemically synthesized artemisinin (Ro *et al.*, 2006). Martin *et al.* (2003) were the first to report the generation of an artemisinin precursor in a microbial system. They engineered *Escherichia coli* with a synthetic mevalonate pathway from *Saccharomyces cerevisiae*. Expression of amorphadiene synthase (*ADS*) from *A. annua*in this strain allowed production of amorpha-4,11-diene, the sesquiterpene olefin precursor to artemisinin. However, despite extensive effort invested in the past decade in metabolic engineering of artemisinin and its precursors in both microbial and heterologous plant systems (Ro *et al.*, 2006; Martin *et al.*, 2003; van Herpen *et al.*, 2010; Zhang *et al.*, 2011; Covello, 2008), production of artemisinin itself has never been achieved.

In addition, transgenic modification of commercially important plant species also calls for the development of novel tools for the removal, addition and replacement of existing transgenes

in plant cells (Hanin and Paszkowski, 2003; Porteus, 2009; Moon *et al*., 2010; Weinthal *et al*., 2010).

Thus, two of the major challenges that are still to be addressed in plant biotechnology are the development of a technology to combine several transgenic traits in a single plant by stacking a number of genes in the same chromosomal locus into a single multigene array and the successive manipulation of this array by genome editing (Dafny-Yelin and Tzfira, 2007; Lyznik and Dress, 2008; Taverniers *et al*., 2008; Naqvi *et al*., 2010). The main strategies for the introduction of multiple genes into plant cells include re-transformation, co- transformation, sexual crossing and transformation of multigene constructs (Reviewed by Dafny-Yelin and Tzfira, 2007; Naqvi *et al*., 2010). While proven useful for the production of transgenic plants with novel traits, re-transformation, co-transformation and sexual crossing approaches all suffer from several flaws (Dafny-Yelin and Tzfira, 2007; Naqvi *et al*., 2010). Re-transformation and sexual crossing, for example, are time consuming and rely on the use of different selectable marker genes for each transformation/crossing cycle. In co-transformation, it is virtually impossible to predict the number of insertions and the distribution of the inserted genes across the plant genome. In addition, co-transformation may result in complex integration patterns which may hinder the use of such plants for commercial purposes for which single and well-characterized integration events are required. While sexual crosses may be simplified by using marker free transgenic plants, the delivery of multiple genes as a single, well-defined, multigene array would perhaps be the simplest and most reliable method for the production of multigene transgenic plants. Furthermore, multigene arrays may also offer the advantage of simplifying successive manipulations of multigene assays in transgenic plants by genome editing technologies (Porteus,

2009; Weinthal *et al*., 2010). Yet, while multigene constructs have been successfully used in several studies (e.g. Bohmert *et al*., 2000; Bohmert *et al*., 2002; Wu *et al*., 2005; Zhong *et al*., 2007; Fujisawa *et al*., 2009), the assembly of multigene constructs remains challenging, being nearly impossible to achieve by traditional cloning methods (Dafny-Yelin and Tzfira, 2007; Naqvi *et al*., 2010).

Only a handful of dedicated vector assembly systems has been developed in the past several years for the assembly of multigene transformation vectors (e.g. Cheo *et al*., 2004; Sasaki *et al*., 2004; Karimi *et al*., 2005; Chen *et al*., 2006; Wakasa *et al*., 2006; Chen *et al*., 2010). Most of these vector assembly systems have a rather limited capacity and have hence been utilized for the delivery of only a small number (i.e., up to five) of transgenes in a single array. Noteworthy here is the assembly system of Lin *et al*. (2003), in which a combination of the Cre/loxP recombination system and two homing endonucleases is used for successive cloning of a potentially infinite number of genes onto a transformation-competent artificial chromosome (TAC)-based vector. Lin *et al*. (2003) used their system for delivery of eight different genes from two independent transferred DNA (T-DNA) molecules that were launched from a single binary vector into the rice genome. However, a crucial limitation of the ingenious approach of Lin *et al*. (2003) was that in contradiction to classical cloning by restriction enzymes, once assembled into the binary vector, new DNA fragments can no longer be removed or replaced by others.

Another important contribution to the field is the unique multiple-round in vivo site-specific assembly (MISSA) system of Chen (2010) that enables in vivo recombination-based assembly of multigene vectors. The system was used to construct several multi-transgene binary vectors and led to the production of transgenic plants with eight transgenes (Chen *et al*., 2010). A

number of other recombination-based cloning strategies has also been developed for the construction of multigene plant transformation vectors (Cheo *et al*., 2004; Sasaki *et al*., 2004; Karimi *et al*., 2005; Chen *et al*., 2006; Wakasa *et al*., 2006; Chen *et al*., 2010). However, here too, the irreversible nature of the recombination-based reactions does not enable the modification of existing binary vectors, and users of such systems may be required to completely rebuild their transformation vectors to give different gene combinations. Multigene transformation vectors have also been assembled by using both traditional cloning methods. Wu *et al*. (2005), for example, combined the use of type-II restriction enzymes and Gateway®- mediated cloning to construct several binary vectors with up to ten different expression cassettes and used the combination to reconstitute the docosahexaenoic acid biosynthetic pathway in Indian mustard. Fujisawa *et al*. (2009) combined the use of type-II restriction enzymes and homing endonucleases to construct a seven-transgene-long T-DNA molecule, which they applied to genetically modify the carotenoid biosynthesis pathway in *Brassicanapus*. Other examples include the engineering of soybean, potato, rice, *Arabidopsis* and several other plant species by using four- to five-transgene-long transformation vectors (reviewed in Naqvi *et al*., 2010; Peremarti *et al*., 2010). While proven successful for the metabolic engineering of various plant species, such vectors were mostly custom designed to their specific tasks, and their modification for other metabolic pathways and multigene traits may prove to be difficult or even impossible to achieve.

A versatile and modular system for the assembly of multigene binary vectors has been developed by Goderis *et al*. (2002) who exploited a set of homing endonucleases to construct a vector system facilitating the successive cloning of independent plant expression cassettes. The principles of the assembly system and successor pSAT vector system was previously reported by

Tzfira *et al*. (2005). An important advantage of this method, over the above described approaches, is its modularity. Plant expression cassettes can easily be removed or replaced from existing binary vectors. Nevertheless, the capacity of the vector system and its successors is limited by the very small number of commercially available homing endonucleases.

We have recently shown that zinc finger nucleases (ZFNs), engineered restriction enzymes that can be designed to bind and cleave long stretches of DNA sequences (Mani *et al*., 2005), can be used for molecular cloning, and we have used such enzymes for the construction of dual gene binary vectors (Zeevi *et al*., 2008). In the present paper, we describe a modular binary vector assembly system that represents a fundamental improvement over the original design of Goderis *et al*. (2002) in that it supports the construction of multigene transformation vectors not only by homing endonucleases but also by designed ZFNs. We describe the design of our system and demonstrate its use by cloning nine different DNA fragments onto a modified binary transformation vector. We also show that such vectors can be used for the production of multigene transgenic plants and for transient expression of multigenes in plant protoplasts. The advantage of using ZFNs for the construction of multigene transformation systems is discussed.

Here we also report the metabolic engineering of tobacco to produce artemisinin, generating transgenic plants that express five plant- and yeast-derived genes involved in the mevalonate and artemisinin pathways, all expressed from a single vector. Our experiments demonstrate that artemisinin can be fully biosynthesized in a heterologous (that is, other than *A. annua*) plant system, such as tobacco. Although the artemisinin levels we have generated in transgenic tobacco are currently lower than those in *A. annua*, our experimental platform should

lead to the design of new routes for the drug's commercial production in heterologous plant systems.

Materials and Methods

Construction of pSAT vectors. To construct pSAT12.MCS, we PCR-amplified the pSAT6.MCS (Tzfira *et al*., 2005) backbone by using 5'ATAAGAATGCGGCCGCGTAAGTGTTGGTGCTGTAAGTATGGATGCAGTAATCATGG TCAT AGCTGTTTCC and 5'GACGCACCGGTAGCACCAACACTTACGTTGGTGCTGGCACTGGCCGTCGTTTTACA ACG and ligated the *Age*I-*Not*I digested PCR product to the *Age*I-*Not*I 1.2-kb fragment from pSAT6.MCS (Tzfira *et al*., 2005). The construction of pSAT10.MCS and pSAT11.MCS was previously described (Zeevi *et al*., 2008). pRCS11.1 was constructed by cloning the *Sma*I fragment of self-annealed pairs of primers 5' TCCCCCGGGTTCCCACAAACTTACTTGTGGGAAAGCACCAACACTTACGTTGGTGCT CCCGGGGGA and 5' TCCCCCGGGAGCACCAACGTAAGTGTTGGTGCTTTCCCACAAGTAAGTTTGTGGGAA CCCGGGGGA which encode the ZFN11 and ZFN12 sites and *Kpn*I fragment of self-annealed pairs of primers 5' GGGGTACCTGCATCCATGTAAGTATGGATGCAGGTACCCC and 5' GGGGTACCTGCATCCATACTTACATGGATGCAGGTACCCC, which encodes the ZFN10 into the same sites of pRCS2. The ChrD-RFP expression cassette was transferred from pSAT6A.ChrD-RFP (Citovsky *et al*., 2006) as an *Age*I-*Not*I fragment into pSAT12.MCS to produce pSAT12.ChrD-RFP. The construction of pSAT10.YFP-CHS and pSAT11.DsRed2-P has been previously described (Zeevi *et al*., 2008). To produce pSAT2.N, the nucleocapsid protein (N) of the sonchus yellow net virus expression cassette was transferred as an *Age*I-*Not*I fragment from pSAT3.N (Tzfira *et al*., 2005) into pSAT2.MCS (Tzfira *et al*., 2005). pSAT5A.RbcsP.BAR

was constructed by transferring the BAR encoding sequence as an *Xma*I-*Xho*I fragment from pSAT1A.ocsP.BAR (Chung *et al*., 2005) into the same sites of pSAT5A.RbcsP.MCS (Chung *et al*., 2005). pSAT4.hsP.GUS was constructed by transferring the *Nco*I-*Bam*HI GUS coding sequence from pRTL2-GUS into pSAT4.hsP.MCS (Tovkach *et al*., 2009). pSAT6A.PAP1 was constructed by transferring the *Kpn*I-*Bam*HI PAP1 coding sequence from pCHS3-PAP1 (Borevitz *et al*., 2000) into pSAT6A.MCS (Chung *et al*., 2005).

Assembly, expression and purification of ZFNs. The coding sequences of ZFN10, ZFN11 and ZFN12 were assembled from overlapping oligonucleotides and cloned into the bacterial expression vector pET28-XH to produce pET28-ZFN10, pET28-ZFN11 and pET28-ZFN12. The protocol for molecular assembly of ZFNs from overlapping oligonucleotides, and their expression and purification have been previously described in detail (Zeevi *et al*., 2010). Briefly, the ZFN DNA binding coding sequences were assembled by a single PCR reaction from a mixture of ZFN backbone primers and ZFN finger specific primers. Each PCR product was then cloned into pET28-XH, where it was fused with the *Fok*I endonuclease domain and a 6xHis-tag. For expression in *E. coli* cells, the ZFN expression vectors were transferred into BL21 GOLD (DE3) PlysS cells (Stratagene). The cells were cultured, harvested and lysed, and the extracted proteins were purified on Ni-NTA agarose beads (Qiagen) as previously described (Zeevi *et al*., 2010). Eluted ZFNs were stored at -20 C in 50% glycerol. Alternatively, we used the Expressway in vitro protein synthesis system (Invitrogen) for in vitro expression of ZFNs.

Construction of binary vectors. For ZFN-mediated digestion of pSAT expression cassettes and binary plasmids, about 200 ng, each, of acceptor and donor plasmids were digested with 0.05–1 µl of purified enzyme in 10 mM Tris [pH8.8], 50 mM NaCl, 1 mM DTT, 100 µM

$ZnCl_2$, 50 μg/ml BSA and 100 μg/ml tRNA in a total reaction volume of 20–30 μl. The reaction was preincubated for 30 min at room temperature, followed by addition of $MgCl_2$ to a final concentration of 5 mM. The digestion reaction mixture was further incubated for 2–40 min at room temperature. For homing-endonuclease-mediated digestion of pSAT expression cassettes and binary plasmids, we followed the recommended reaction conditions for each enzyme. Cleaved fragments were separated by gel electrophoresis, purified with a GFX Gel Band Purification Kit (Amersham), dephosphorylated with shrimp alkaline phosphatase (Fermentas), ligated with T4 ligase (NEB) and transferred to chemically competent DH5α *E. coli* cells by using standard molecular biology protocols. The order in which the different pSAT expression cassettes were assembled into pRCS11.1 is described in the Results section. Binary vectors were transformed into chemically competent EHA105 *Agrobacterium* cells as previously described (Tzfira *et al*., 1997). pET28 ZFN-expression plasmids and the pSAT and binary plasmids described in this study are available upon request.

Protoplast transfection and production of transgenic plants. The Tape-*Arabidopsis* Sandwich (Wu *et al*., 2009) method was used for protoplast isolation and transfection by multigene binary vectors. Transfected protoplasts were cultured in W5 solution in 1% BSA-coated 6-well plates for 16-24 h at 24 ℃ to allow expression of the transfected DNA. Transgenic *Arabidopsis* plants were produced by using the standard flower-dip transformation method (Clough and Bent, 1998). Transgenic plants were selected on a hygromycin-selection medium. For analysis of resistance to Basta, hygromycin-resistant seedlings were transferred to soil, allowed to grow, and then sprayed with commercial Basta.

Confocal microscopy. Protoplasts and plant tissues were viewed directly under a confocal

laser-scanning microscope (TCS SP5; Leica). EYFP was excited with an argon laser at 514 nm, and fluorescence was monitored between 525 and 540 nm. DsRed2 and mRFP were excited with a helium-neon laser at 561 nm, and fluorescence was monitored between 570 and 630 nm. Chlorophyll fluorescence was monitored above 660 nm.

Expression vectors and plant transformation for artemisinin biosynthesis.

Artemisinin pathway genes *ADS, CYP71AV1* and *DBR2,* as well as *CPR* (GenBank accession nos. Q9AR04, DQ315671, EU704257, and DQ984181, respectively) were PCR-amplified from *Artemisia annua* cDNA and cloned into pGEMT vector (Promega). Generation of mitochondrial-targeted amorpha-4,11-diene synthase (*mtADS*) and a mutated form of yeast 3-hydroxy-3-methylglutaryl-coenzyme A reductase (*tHMG*) was as reported previously (Farhi *et al.*, 2011). The open reading frames were cloned as PCR fragments into pSAT series vectors (Tovkach *et al.*, 2009; Tzfira *et al.*, 2005; Takahashi and Komeda, 1989; Chung *et al.*, 2005; Ni *et al.*, 1995), producing pSAT2A.nosP.tHMG, pSAT5.1.hspP.CYP, pSAT6A.supP.DBR2, pSAT4.1A.rbcP.CPR, pSAT5A.35SP.ADS and pSAT5A.35SP.mtADS. The expression cassettes from pSAT were cloned into a pRCS16F vector alongside the kanamycin-resistance cassette from pSAT1A.ocsAocsP.nptII.ocsT (Takahashi and Komeda, 1989). *Agrobacterium tumefaciens* strain AGL0 carrying these vectors was used for stable transformation of *Nicotiana tabacum* cv. Samsun by the standard leaf-disc transformation method (Gallois and Marinho, 1995). Briefly, tobacco leaf disks were inoculated with the bacteria and maintained on a regeneration medium containing Murashige and Skoog (MS) salt mixture, 2% sucrose, 1% manitol, 0.8% agar, 2 mg/l of zeatin, 0.1 mg/l of indoleacetic acid (pH 5.8) and supplemented with 300 µg/l kanamycin and 300 µg/l carbenicillin. Developed

shoots were individually transferred to tissue-culture flasks containing MS, 2% sucrose, 1% agar and 300 µg/l kanamycin. Once plantlets rooted, they were transplanted to soil and grown in the greenhouse in random plots of plants at day/night temperatures of 38/25 ℃ under natural photoperiod (1500–1700 µmol/m^2s) in Rehovot, Israel. Transgenic lines (ADS or mtADS lines) did not show any phenotypic differences, i.e. in germination, leaf area, stature, time to flowering or vigor, compared to control plants. Once plants reached the age of 4 months they were sampled for chemical analysis (during July to September, 2010): 5 to 10 leaves from each line were harvested, sun-dried for 3 days and ground.

Chemical analysis of artemisinin. Ground dried leaf samples (100 mg from each independent transgenic line) were supplemented with 10 ng deuterium-labeled artemisinin (Toronto Research Chemicals) and extracted by sonication for 15 min with 2 ml hexane. After partitioning into 1 ml methanol, phases were separated and the methanolic layer was concentrated to ~100 µl under a nitrogen stream. High-resolution LC-MS was performed using an Accela LC system coupled with the LTQ Orbitrap Discovery Hybrid FT (Thermo Fisher Scientific Inc.). For chromatographic separation, two Zorbax Eclipse XDB-C18 (100 × 2.1 mm, 1.8 µm, Agilent Technologies) columns were connected in sequence, followed by Synergy Fusion-RP (100 × 2 mm, 2.5 µm, Phenomenex). Column temperature was maintained at 40 ℃, flow rate was 250 µl/min and injection volume was 10 µl. Chromatographic analysis was performed using a binary gradient of water/acetone/0.1% acetic acid. The mass spectrometer was equipped with an APCI ion source operated in positive ionization mode. The ion source parameters were as follows: corona discharge needle current 5 µA, capillary temperature 250 ℃, sheath gas rate 50 (arb), auxiliary gas rate 10 (arb), vaporizer

temperature 400 °C; 99.5% nitrogen was used as sheath and auxiliary gas. Ion transfer optical parameters were optimized for protonated artemisinin using the automatic tune option. Mass spectra were acquired in *m/z* 200-800 Da range, with a resolution of 30,000. Accurate mass ions of *m/z* 283.1530 and 286.1733 were used to monitor artemisinin and artemisinin-d3, respectively. The system was controlled, and data analyzed, using Xcalibur software (Thermo Fisher Scientific Inc.). UPLC-MS/MS analysis was performed using the Agilent 1200 system coupled to an Agilent 6410 triple-quadrupole mass spectrometer. Chromatographic analysis was performed as for UPLC-HR-MS. The MS was equipped with an electrospray ionization ion source which was operated in positive mode with the following parameters: capillary voltage 4000 V, nebulizer pressure 241 kPa, drying gas 10 l/min, gas temperature 350 °C; 99.5% nitrogen was used as nebulizer and drying gas and 99.999% nitrogen was used as the collision gas. Artemisinin was detected in MRM mode by monitoring three transitions (283 $[MH^+] \rightarrow 247$, $283 \rightarrow 265$ and $283 \rightarrow 219$) under the following optimized parameters: fragmentor voltage 80 V and CID energy 4 eV. The system was controlled, and data were analyzed using MassHunter software (Agilent Technologies Inc.). Artemisinin levels per gram DW of plant material were calculated using the deuterium-labeled artemisinin internal standard.

GC-MS analysis of amorpha-4,11-diene. Leaf samples of tobacco or *Artemisia annua* (500 mg from each independent line) were ground in liquid nitrogen and extracted twice by sonication for 30 min with 2 ml hexane and 600 ng of the sesquiterpene valencene as internal standard. The extract was partially purified on a silica gel column (100 mesh) and washed with hexane. The eluate was concentrated under nitrogen stream before analyzing a 1-µl aliquot by GC-MS. The system was composed of a TRACE GC 2000 gas chromatograph and a

TRACE DSQ quadrupole mass spectrometer (ThermoFinnigan). GC was run in a 30 m Rtx-5Sil MS column with 0.25-µm film thickness (Restek). The injection temperature was set at 250 ℃, with an initial oven temperature of 100 ℃ for 1 min, followed by a 5 ℃/min ramp to 270 ℃. MS was operated in EI mode (70 eV) in both scanning mode (40-325 *m/z*) and selected ion monitoring of the molecular and fragment ions (204 and 119, 161 and 189 m/z). Amorpha-4,11-diene was identified and quantified using the selected ions' responses compared to that of the valencene internal standard as described previously (Farhi *et al.*, 2011).

Localization of ADS and mtADS. *ADS* and *mtADS* were cloned in frame upstream of *EGFP* in pSAT6-EGFP-C1 vector (Chung *et al.*, 2005). The resulting plasmids were used for transformation of protoplasts isolated from *Arabidopsis thaliana* leaf mesophyll using the TAPE-Arabidopsis Sandwich protoplast isolation method (Wu *et al.*, 2009) and cellular localization of EGFP signal was analyzed as reported previously (Spitzer-Rimon *et al.*, 2010).

Results

Design of the binary vector and novel pSATs. We followed the basic design of the pAUX (Goderis *et al*., 2002) and its successor pSAT (Chung *et al*., 2005; Tzfira *et al*., 2005; Dafny-Yelin *et al*., 2007; Dafny-Yelin and Tzfira, 2007) family of plasmids to facilitate the assembly of multigenes into a single binary plasmid. In the pSAT system, functional plant expression cassettes are individually cloned into different pSAT plasmids (e.g., geneB in pSAT2, **Figure 3-1A**), and an expression cassette from each type of pSAT plasmid can be cloned onto pPZP-RCS1- or pPZP-RCS2-based (Goderis *et al*., 2002) *Agrobacterium* binary plasmids by using matching homing endonucleases (**Figure 3-1A**). As was the case for the original set of pAUX plasmids, the original set of pSAT plasmids was composed of seven different versions, in which the expression cassettes were flanked by *Asc*I (pSAT1), *Asc*I and I-*Ppo*I (pSAT2), I-*Ppo*I (pSAT3), I-*Sce*I (pSAT4), I-*Ceu*I (pSAT5), PI-*Psp*I (pSAT6) and PI-*Tli*I (pSAT7). Here we subsequently expanded this set to include pSAT10, pSAT11 and pSAT12, in which the expression cassettes were flanked by ZFN10, ZFN11 and ZFN12 (3x3 finger ZFNs), respectively (e.g., geneA in pSAT12, **Figure 3-1A**). We also constructed pRCS11.1, which is a modification of pPZP-RCS1 that was engineered to include, in addition to the original recognition sites of *Asc*I, I-*Ppo*I, I-*Sce*I, I-*Ceu*I, PI-*Psp*I and PI-*Tli*I, recognition sites to the ZFN10, ZFN11 and ZFN12 zinc finger nucleases (**Figure 3-1A**). This system was subsequently used for the construction of several multigene binary plasmids, as described below.

Construction of multigene vectors by exploiting ZFNs. We constructed a set of pSAT plasmids carrying a wide variety of genes. **Table 3-1** lists the different pSAT plasmids constructed, the genes they carry, and the enzymes used for their cloning into pRCS11.1. We

started by cloning the constitutive expression cassette of the endoplasmic-reticulum (ER)-bound chalcone synthase (CHS) gene (Pelletier and Shirley, 1996), tagged with EYFP from pSAT10.EYFP-CHS by using ZFN10 to produce pRCS11[10.EYFP- CHS]. Next we added the constitutive expression cassette of the P protein of Sonchus Yellow Net Rhabdovirus (SYNV) (Goodin *et al*., 2001; Goodin *et al*., 2002) tagged with DsRed2 from pSAT11.DsRed2-P by using ZFN11, to produce pRCS11.1[10.EYFP-CHS][11.DsRed2-P].

The key to the versatility of the original pSAT/pRCS2 multigene assembly system lies in the ability it offers not only to add but also to remove and replace DNA fragments by homing endonucleases during the construction of the various binary plasmids (Tzfira *et al*., 2005).Thereafter, we demonstrated that ZFNs too can also be used to remove and replace DNA fragments from existing binary constructs. We started by adding the plasmid backbone from pSAT12.MCS into pRCS11.1[10.EYFP-CHS][11.DsRed2-P] and produced pRCS11.1[10.EYFP-CHS][11.DsRed2-P][12.AMP]. As expected, the resultant plasmid conferred resistance to ampicillin in *Escherichia coli* cells. We also observed that DNA preparation of this plasmid from an overnight *E. coli* culture resulted in much higher DNA yield than that obtained with its progenitor or with other pRCS2-based vectors, most probably due to the presence of the ColE1 origin of replication on the pSAT12.MCS backbone. We then used ZFN12 to remove the pSAT12.MCS backbone and replaced it with a constitutive expression cassette of the chromoplast-specific carotenoid-associated protein ChrD from *Cucumis sativus* (Vishnevetsky *et al*., 1996; Vishnevetsky *et al*., 1999) tagged with RFP from pSAT12.ChrD-RFP to produce pRCS11.1[10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP]. As expected, this plasmid no longer conferred resistance to ampicillin in *E. coli* cells. We tested the expression of the three reporter

genes from pRCS11.1[10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP] in *Arabidopsis* protoplasts. As expected, the YFP-tagged CHS associated with the rough ER throughout the cell, around the protoplasts and around the nucleus (**Figure 3-2A**); DsRed2-P was observed in the cytoplasm clustered around the chloroplasts; and ChrD-RFP, which has been previously characterized as a chloroplast associated protein (Ben Zvi *et al*., 2008), indeed localized in the chloroplasts (**Figure 3-2B**). The clustering of DsRed2-P was probably due to the tendency of DsRed2 to aggregate in living cells. We next added a fourth plant expression cassette for the SYNV N protein and produced pRCS11.1[2.N][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP]. Interaction of the SYNV N protein with SYNV P is required to translocate it to into sub-nuclear compartments in *Nicotiana bentamiana* cells (Goodin *et al*., 2001; Goodin *et al*., 2002). Indeed, nuclear localization and sub-nuclear compartmenting of the DsRed2-P signal was observed in pRCS11.1[2.N][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP]-infected protoplasts (**Figure 3-2F** and **2K**) but not in pRCS11.1[10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP]-infected protoplasts (**Figure 3-2B** and **2I**). Our data thus show that designed ZFNs can be used for the construction of multigene vectors and that such vectors can be used to drive the simultaneous expression of several genes in plant cells.

Assembly of multigene binary vectors for stable transformation.

pRCS11.1[2.N][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP] carried four pairs of repetitive elements (i.e., CaMV dual 35S promoters and CaMV 35S terminators). We next tested whether the four transgenes and two additional genes (i.e., a hygromycin resistance gene driven by the control of the octopine synthase promoter and terminator and the Basta resistance gene driven by the control of the Rubisco small subunit promoter and terminator) can be stably transformed and

expressed in transgenic plants. Many commonly used binary vectors have been designed with the selection marker cassette cloned next to the T-DNA's left border (van Engelen *et al*., 1995; Hellens *et al*., 2000). We elected to clone the hygromycin-resistant expression cassette near the T-DNA's right border. This design, while having the potential to produce transgenic plants with truncated T-DNAs, allowed us to examine whether repetitive elements have a negative impact on the T-DNA structure during transformation. To achieve the above design, we first produced pRCS11.1[1.HYG][2.N][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP], in which the hygromycin-resistance gene expression cassette was cloned into the *Asc*I site of pRCS11.1[3.N][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP]. We then added the plasmid backbone from pSAT3.MCS into the I-*Ppo*I site to facilitate future cloning and subsequent isolation of T-DNA-plant junction sequences by plasmid rescue, and the Basta hygromycin-resistant expression cassette into the I-*Ceu*I site of pRCS11.1[3.N][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP]. Finally, we used the seven-transgene-long pRCS11.1[1.HYG][2.N][3.AMP][5.BAR][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP] binary plasmid (**Figure 3-3A**) to produce transgenic *Arabidopsis* plants. Confocal microscopy analysis revealed that 9 of our 27 transgenic lines exhibited yellow and red fluorescence expression patterns similar to the patterns of protoplasts infected with pRCS11.1[2.N][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP] (data not shown). Molecular analysis of two randomly selected transgenic lines revealed that PCR amplification of their T-DNA regions was similar to that of the seven-transgene-long binary vector (**Figure 3-3A**); these findings indicate that our transgenic lines carried all the T-DNA molecule-encoded transgenes within their genome.

We then allowed several randomly selected transgenic lines to mature and set seed. Shown

in **Figure 3-4** are examples of hygromycin- and Basta-resistant seedlings. Collectively, our data indicate that the T-DNA-encoded seven transgenes were stability expressed and inherited.

Production of nine-transgene-long transgenic plants. We constructed a nine-transgene-long binary vector by adding the GUS and PAP1 (*Arabidopsis* transcription factor production of Anthocyanin pigment) (Borevitz *et al*., 2000; Ben Zvi *et al*., 2008) expression cassettes into the I-*Sce*I and PI-*Psp*I sites of pRCS11.1[1.HYG][2.N][3.AMP][5.BAR][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP]. In the final binary vector, pRCS11.1[1.HYG][2.N][3.AMP][4.GUS][5.BAR][6.PAP1][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP], GUS expression is controlled by the hsp18.1 heat shock promoter (Takahashi and Komeda, 1989), while PAP1 expression is controlled by CaMV dual 35S promoter. We used this vector to produce over 30 hygromycin-resistant transgenic *Arabidopsis* plants, of which about half exhibited PAP1-related reddish hypocotyls and cotyledons during germination. We further characterized these lines and determined whether they are likely to harbor the full T-DNA molecule by monitoring DsRed2-P expression, which was located next to the left border. Twelve T_0 plants of the PAP1 transgenic plants also expressed DsRed2-P, which indicated that they are likely to harbor at least one full T-DNA copy. Molecular analysis of several lines revealed that PCR amplification of several regions on their T-DNA regions was similar to that of the nine-transgene-long binary vector, as demonstrated, for example, in **Figure 3-3B**. These observations indicate that the plants carried the entire T-DNA molecule-encoded genes.

We allowed seven of the transgenic lines to mature and set seed and determined the inheritance and stability of the T-DNA in the next generations. T_0 plants of lines L6, L7 and L10 exhibit a plausible single insert segregation pattern when grown on a hygromycin-containing

medium (**Table 3-2**). T_1 seedlings of lines L2, L6, L7 and L10, T_2 seedlings that derived from homozygous (i.e., L6-1) and heterozygous (i.e. L6-2 and L10-2) parents, exhibited resistance to Basta, as did T_2 seedlings derived from lines L2 and L3 (**Table 3-2** and **Figure 3-5A**). Confocal microscopy analysis revealed that all T_0 lines and their progeny (**Table 3-2**) exhibited red and yellow fluorescence, as shown, for example, for line L6-1 (**Figure 3-5C-F**). Analysis of several T_2 lines revealed that they all expressed heat-shock-induced GUS expression (**Figure 3-5G**). GUS expression was not observed in untreated plants. In addition, all T_1, T_2 and T_3 hygromycin- resistant plants exhibited PAP1 expression (**Table 3-1**), as shown, for example, for L3-3-1 (**Figure 3-5B**) and L6-2-1-1 (**Figure 3-5F**) plants. The segregation analyses suggest that a single transgene locus was obtained in some of the analyzed lines (i.e. L6, L7 and L10).

We further analyzed four transgenic lines to demonstrate the stability and integrity of the T-DNA inserts in T_3 (lines L2 and L10) and T_4 (lines L6 and L7) hygromycin-resistant plants, by PCR amplification of overlapping fragments of their T-DNA region (**Figure 3-6**). When combined with segregation analysis, our molecular data further supported the notion that our transgenic plants carry full T-DNA inserts. No less important, our data clearly show that multigene- expressing plants can be produced using our multigene transformation vectors.

Generation of the potent anti-malarial drug artemisinin in tobacco. The World Health Organisation (WHO; Geneva) promotes the use of artemisinin as a first-line treatment for malaria, and it is heavily involved in facilitating the development of artemisinin-based anti-malaria drugs1. Artemisinin is biosynthesized from terpenoid backbones generated by the mevalonate and methyl-erythritol phosphate (MEP) pathways (Brown, 2010; Towler and Weathers, 2007; Schramek *et al.*, 2010) (**Figure 3-7A**). Although detailed knowledge of the artemisinin-biosynthesis pathway is

still lacking, it initiates with the cyclization of farnesyl diphosphate by ADS to form amorpha-4,11-diene, which is then oxidized by the cytochrome P450 CYP71AV1, reduced by artemisinic aldehyde reductase (DBR2) and possibly reoxidized by aldehyde dehydrogenase to yield dihydroartemisinic acid—the presumed precursor of artemisinin in plants (Covello, 2008; Brown, 2010; Zhang *et al.*, 2008). Normally, dihydroartemisinic acid accumulates in *A. annua* and slowly converts to artemisinin, a process that can be stimulated after harvest by drying in the sun. The transformation of dihydroartemisinic acid to artemisinin in *A. annua* has been proposed to be nonenzymatic, requiring only the presence of light and molecular oxygen (Covello, 2008; Brown, 2010); a singlet oxygen formed as a consequence of exposure to UV/visible light may react with dihydroartemisinic acid to form a ketoenol, which can then react with ground-state oxygen to form a second hydroperoxide that spontaneously forms artemisinin. Despite the great interest in enhancing yields of artemisinin in its host *A. annua*, classic breeding and genetic engineering strategies have met with only limited success (Covello, 2008; Graham *et al.*, 2010). Moreover, recent attempts to produce the artemisinin precursors artemisinic and dihydroartemisinic acids in heterologous plants (van Herpen *et al.*, 2010; Zhang *et al.*, 2011) did not lead to their accumulation due to internal glycosylation and insufficient oxidation toward the acids. Harnessing *E. coli* and *S. cerevisiae* has allowed the production of high titers of amorpha-4,11-diene and artemisinic acid (Ro *et al.*, 2006; Martin *et al.*, 2003), but not of the active artemisinin drug itself.

To reconstruct the artemisinin-producing pathway in *Nicotiana tabacum*, we first generated a mega-vector carrying cytochrome P450 reductase (*CPR*) from *A. annua* to prevent the accumulation of inactive oxidized P450, as well as *ADS*, *CYP71AV1* and *DBR2* (**Figure 3-7B** and **Materials and Methods**). The mega-vector also contained a truncated and deregulated 3-

hydroxy-3-methylglutarylcoenzyme A reductase (*tHMG*) from yeast to increase the supply of precursor from the mevalonate pathway for artemisinin production. To ensure genetic stability and expression in plants, all five genes and an antibiotic-resistance selection gene were placed under the control of different promoter and terminator sequences (**Materials and Methods**). To enhance production yields of foreign terpenes in plants, we tested the possibility of targeting ADS to the mitochondria using a mega-vector carrying *ADS* fused to *COX4* signal peptide (Farhi *et al.*, 2011; Kappers *et al.*, 2005) (*mtADS*). Shunting terpenoid biosynthesis via the mitochondria has recently been shown to be greatly advantageous for the metabolic engineering of plant terpenoids (Farhi *et al.*, 2011; Kappers *et al.*, 2005). Targeting of mtADS and ADS to the mitochondria and cytosol, respectively, was confirmed using protoplasts and fluorescent protein tags (**Figure 3-8A**).

Tobacco plants were stably transformed with the mega-vector harboring either *ADS* or *mtADS* (ADS and mtADS plants, respectively), and expression of all five genes in transgenic plants was confirmed by reverse transcriptase PCR analysis (**Figure 3-8B**). Plants transgenic for all of these genes but lacking *ADS* or *mtADS*, plants expressing GFP and wild-type tobacco were used as controls. No phenotypic differences were observed between artemisinin-producing ADS and mtADS plants and control lines. Artemisinin production was monitored in independent transgenic tobacco lines (four mtADS and four ADS lines) using ultrahigh performance liquid chromatography coupled with high-resolution accurate mass spectrometry (UPLC-HR-MS). Artemisinin was identified in extracts from *ADS*- and *mtADS*-carrying transgenic lines, based on both an authentic standard and internal deuterium-labeled artemisinin (**Figure 3-9A**). To further validate artemisinin identification, extracts were analyzed using UPLC coupled to a triple-quadrupole mass spectrometer (UPLC-MS/MS) operated in multiple reaction monitoring (MRM)

mode (**Figure 3-9B**). MRM traces identical to those of an artemisinin standard were detected in transgenic tobacco extracts that had shown artemisinin accumulation in the UPLC-HR-MS analysis. Analyses of artemisinin yield in transgenic tobacco lines using a deuterium-labeled artemisinin standard revealed that harnessing the mitochondria for amorpha-4,11-diene production yields higher levels of artemisinin: independent transgenic tobacco lines with cytosolic ADS produced 0.75, 0.88, 0.94 and 0.48 mg artemisinin/g dry weight, whereas tobacco lines with mtADS generated 5.0, 5.9, 6.3 and 6.8 mg artemisinin/g dry weight. Amorpha-4,11-diene production levels in the mtADS transgenic tobacco plants were also higher compared with ADS plants and similar to levels obtained in previous attempts to generate artemisinin in heterologous hosts (van Herpen *et al.*, 2010; Zhang *et al.*, 2011). In transgenic plants generated with the ADS construct, we observed amorpha-4,11-diene levels of 26–72 ng/g fresh weight, whereas in plants generated with the mtADS construct, amorpha-4,11-diene accumulated to about tenfold higher levels, 137–827 ng/g fresh weight (as measured by gas chromatography-MS analysis; **Materials and Methods**). Amorpha-4,11-diene levels in *A. annua* plants were 5.2–10.8 mg/g fresh weight, which is comparable to previous reports (Ma *et al.*, 2009). The observation that tobacco and *A. annua* plants accumulated lower levels of amorpha-4,11-diene compared with artemisinin raises the possibility that the olefin's levels are a limiting factor in the drug's production.

Discussion

We demonstrated that by using engineered ZFNs we can overcome the limitation imposed by the small number of commercial homing endonucleases to the construction of multigene binary vectors. We used several ZFN- and homing-endonuclease-constructed binary vectors for transient and stable genetic transformation of plant cells. We then showed that transgenic plants that had been stably transformed by nine-transgene-long T-DNA stably expressed the cloned genes across several generations. To the best of our knowledge, our report is one of just a small number of studies in which a very large number of independently expressed genes were delivered using a single T-DNA molecule into plant cells and moreover it is the only report in which a multigene vector has been assembled by a modular, step-by-step construction method (reviewed in Naqvi *et al*., 2010). Thus, our approach represents an important technical leap in the construction of multigene plant transformation vectors.

There are three main advantages of our system over other multigene construction systems, as detailed below. First, we demonstrated that ZFNs can be used in a manner similar to the technique that relies on homing endonucleases and that ZFNs can be exploited not only for adding but also for replacing DNA fragments from existing multigene binary vectors. Thus, in contrast to many other multigene vector assembly systems (e.g. Lin *et al*., 2003; Karimi *et al*., 2005; Wakasa *et al*., 2006; Chen *et al*., 2010), our system has the dual advantage of being modular and of enabling the modification of existing plant expression cassettes at any stage during the construction of the transformation vector. In addition, the cloning capacity of existing pRCS11.1-based binary vectors can be increased by reengineering their multiple cloning sites (MCSs) for additional features (which may include, for example, recognition sites for new ZFNs

and homing endonucleases or even sites for Cre/Lox and Gateway® recombination systems) not only prior to but also during the process of assembling multigene vectors. This can be achieved by cloning an expanded ZFN-recognition-MCS into the growing pRCS11.1-based binary vectors, which can then be used for successive additions of new expression cassettes. This feature further adds to the modularity of our ZFN-based vector assembly system, as compared with the more rigid design of other binary vector assembly systems.

Second, because ZFNs can be designed to target and digest an extremely large number of target sequences (Maeder *et al*., 2008), novel ZFNs can be developed and used to expand our nine-gene-long system beyond its current capacity. Since most pSAT vectors share a similar basic structure (Tzfira *et al*., 2005), it is easy to convert existing pSAT vectors into novel vectors by creating pSAT backbones with new ZFN recognition sites. Furthermore, since we used semi palindromic ZFN recognition target sequences, it may be feasible to adapt existing ZFNs (i.e., those that have been developed for targeting experiments in various non-plant eukaryotic cells [e.g. Urnov *et al*., 2005; Carroll *et al*., 2008; Zimmerman *et al*., 2008; Geurts *et al*., 2009; Takasu *et al*., 2010]) for cloning purposes. Indeed, we have previously showed that ZFNs can be expressed and purified to the level of molecular-biology reagents by using relatively simple expression and purification steps (Zeevi *et al*., 2010; Tovkach *et al*., 2011). While specific modifications may be required to adapt our expression and purification system for novel ZFNs, the simplicity of the process and the availability of dedicated ZFN assembly and expression vectors for bacterial expression and for plant genome editing (Tovkach *et al*., 2010; Zeevi *et al*., 2010) may further facilitate the use of our system for assembly of multigene vector systems.

Third, our system supports the use of a very large family of pSAT- and pAUX-based

plasmids. The pSAT family of plasmids is composed of plasmids that have been designed to facilitate: (i) the overexpression of target genes under the control of various promoters and terminators (Chung *et al*., 2005; Tzfira *et al*., 2005), (ii) analysis of protein-protein interactions by using the bimolecular fluorescence complementation and multi-color bimolecular fluorescence complementation assays (Citovsky *et al*., 2006; Lee *et al*., 2008), (iii) the fusion of target genes to various autofluorescence proteins (Tzfira *et al*., 2005), (iv) RNAi-mediated downregulation (Dafny-Yelin *et al*., 2007), (v) expression of ZFNs (Tovkach *et al*., 2009; Tovkach *et al*., 2010), (vi) Gateway®-mediated gene cloning (Tzfira *et al*., 2005; Chakrabarty *et al*., 2007) and (vii) expression of epitope-tagged proteins (Tzfira T., unpublished data). Thus, users of our system can enjoy a rich resource of plasmids that can be easily adapted to their various needs. Furthermore, since our cloning system is based on ZFN and homing endonucleases, it may be simple to transfer or reconstruct the ZFN and homing endonuclease MCS from pRCS11.1 into any other type of binary vectors and adapt our pSAT set of vectors for cloning on multigene vectors systems in various binary plasmids. Such binaries can, for example, be based on the TAC (Liu *et al*., 1999; Lin *et al*., 2003) and BIBAC (Hamilton, 1997; Frary and Hamilton, 2001) vectors which can facilitate the cloning and transfer of extremely large number of independent plant expression cassettes.

An important feature of our multigene vector assembly system lies in the unique structure of the final array of transgenes, in which each expression cassette or gene is flanked by pairs of unique ZFNs or homing endonucleases. This structure may facilitate genomic editing of multi-transgene arrays in transgenic plants by harnessing the cells' non-homologous end joining (NHEJ) DNA repair pathway. Both homing endonucleases and ZFNs have been used for gene

targeting in plant cells (i.e., site-specific mutagenesis, gene addition deletion and/or replacement) (e.g. Salomon and Puchta, 1998; Chilton and Que, 2003; Tzfira *et al*., 2003; Cai *et al*., 2009; Shukla *et al*., 2009; Townsend *et al*., 2009). Yet, while most ZFN-mediated genome editing systems rely on homologous recombination between the donor DNA and the target genome, the presence of repetitive elements (i.e., promoter and terminator sequences) in multi-transgene arrays may hinder homologous recombination-mediated genome editing of such structures. Thus, genome editing of multi-transgene arrays, i.e., removal, replacement and addition of expression cassettes by NHEJ, can potentially be achieved by transient ZFN-expression in target cells, with or without the addition of a donor DNA molecule (Weinthal *et al*.,2010). Indeed, both I-*Sce*I- and I-*Ceu*I-mediated transgene addition (Salomon and Puchta, 1998; Chilton and Que, 2003; Tzfira *et al*., 2003) and ZFN-mediated transgene deletion (Petolino *et al*., 2010) have been reported in plant cells. In addition, we have recently demonstrated that ZFN- and NHEJ-mediated gene replacement is also feasible in plant species (Weinthal *et al.*, 2013). Worth noting is the fact that by flanking each expression cassette with semi palindromic sequences, the editing process may be technically simplified since, in contrast to most homologous recombination-mediated gene-replacement methods, it will require the expression of just one ZFN monomer and not pairs of ZFNs. Also worth noting is that other types of engineered enzymes [i.e. engineered meganucleases (Gao *et al*., 2010) and TALENs (Cermak *et al*., 2011 Mahfouz *et al*., 2011)] have also been used for genome editing in plant cells. Developing procedures for the expression and purification of engineered meganucleases and TALENs for cloning purposes will further extend the repertoire of enzymes suitable for assembly of multigene vector systems and their successive manipulation in plant cells.

The production of transgenic plants often results in a wide range of transgene-related phenotypes, a phenomenon that is typically attributed to a transgene positional affect (Matzke and Matzke, 1998). Thus, obtaining plants with superior transgenic performances (e.g., high and stable expression levels during the plant life span and across several generations) often calls for the production and screening of a large number of transgenic plants. Increasing the number of transgenes in plants is likely to increase the number of transgenic plants that need to be produced, screened and selected for the desired performances. This procedure may be time consuming and labor intensive, especially if the different transgenes are scattered across the genome. Linking several transgenes to the same genomic location may assist in the screening process and in maintaining superior clones across several generations. The tendency of multiple DNA molecules, derived from co-transformation of several different *Agrobacterium* strains or from co-bombarded of several plasmids, to integrate into the same genomic locus (De Neve *et al*., 1997) has been utilized as a viable method for the production of single-locus multigene transgenic plants (reviewed in Naqvi *et al*., 2010; Peremarti *et al*., 2010). While this approach has been successfully used to produce transgenic plants in which up to 11 transgenes are co-integrated into the same locus (Chen *et al*., 1998; Zhu *et al*., 2008; Naqvi *et al*., 2009), it is impossible to predict the physical organization and the arrangement of co-transformed DNA molecules in the transgenic plants. Indeed, it has been suggested that future progress in multigene transformation may depend not only on driving the transgenes under the specific combinations of promoter and terminator sequences but also on organizing the transgenes in a predetermined pattern (Peremarti *et al*., 2010). Our modular system can thus be efficiently adapted for the analysis of various regulatory elements and gene organization patterns by shuffling between different pSAT vectors.

We observed a variety of phenotypes among our different transgenic lines, as was evident, for example, by differences in hygromycin resistance (**Figure 3-3A**) and GUS expression (**Figure 3-5G**). We also observed that transgenic lines in which all the transgenes were expressed across several generations could be obtained (**Table 3-2**). Yet, we could not identify a clear correlation between the phenotypes of different transgenes on a given T-DNA or between different lines. Thus, for example, lines that exhibited strong resistance to hygromycin did not necessarily exhibit high levels of PAP1 or GUS expression, while plants that exhibited high levels of PAP1 expression, did not yield high intensity of their fluorescence genes. Similarly, Fujisawa *et al*. (Fujisawa *et al*., 2009) reported that they could not associate the expression level of individual transgenes with the performances of three transgenic lines; similarly, they could not determine the efficiency of a specific promoter in a given multigene array due to differences in the gene expression levels, which did not correlate with their promoter types. Similarly, Chen *et al*. (2010) reported phenotypical variation not only between the same trait in different transgenic lines but also between individual traits, driven under the same promoter in a given multigene array, in a selected transgenic line. It thus seems that not only may the expression levels of individual genes vary among different lines, but also that similar promoters, driving different genes, may behave differently in a particular multigene cluster. We are currently applying our multigene vector assembly system to construct a set of multigene binary vectors in which identical expression cassettes will be cloned to different organizations. We will use our vectors to produce collections of independent transgenic plants, which will be subjected to gene expression analysis in an attempt to produce the necessary data that will assist in defining the putative rules for the organization of regulatory elements within a given construct.

Our ability to successfully produce artemisinin in tobacco is most likely due to a combination of factors: expression of *CPR* concomitant with *CYP71AV1*, expression of *tHMG* to enhance precursor availability, the use of intense light conditions during plant cultivation and single-vector-based transformation. Although it is well established that the oxidizing activity of CYP71AV1 is coupled to that of CPR2, the effect of elevated reductase levels on oxygenation rates by cytochrome P450s, including CYP71AV1, in the context of metabolic engineering has never been studied in plants. In this respect, Zhang *et al.* (2011) suggested that their lack of success in producing dihydroartemisinic acid in a heterologous plant system might be due to low oxidation rates. There is evidence that in the presence of light and oxygen, conversion of dihydroartemisinic acid to artemisinin in *A. annua* is spontaneous rather than enzymatically catalyzed (Covello, 2008; Brown, 2010). Thus, the use of intense light conditions during transgenic tobacco cultivation may promote this conversion and lead to accumulation of artemisinin in the metabolically engineered tobacco. The integration of five genes encoding a metabolic pathway into a single vector, while employing a range of promoters, has also been shown to be instrumental in supporting robust metabolic engineering of, for example, carotenoids, vitamin E and polyunsaturated fatty acids in plants (Naqvi *et al.*, 2010; Peremarti *et al.*, 2010). HMG-R is a putative rate-limiting step in the mevalonate pathway; enhancing carbon flux through the mevalonate pathway by expressing a truncated and deregulated HMG-R has been shown to increase the accumulation of both native and foreign terpenes (Farhi *et al.*, 2011; Wu *et al.*, 2006). The use of mitochondria as an alternative or additional cellular compartment recently has been shown to enhance the production of olefin terpenoids (Farhi *et al.*, 2011; Kappers *et al.*, 2005). Indeed, tobacco plants expressing the catalytic domain of HMG-R (tHMG) and producing the

sesquiterpene backbone of artemisinin in mitochondria generated the highest levels of artemisinin. Simultaneous expression of *tHMG* and *mtADS* yielded levels of artemisinin approximately eightfold higher than cytosol-driven drug production.

Although the generated levels of artemisinin in tobacco were lower than in *A. annua*, co-expression of additional enzymes (e.g., aldehyde dehydrogenase 1 and farnesyl diphosphate synthase (van Herpen *et al.*, 2010; Zhang *et al.*, 2011; Wu *et al.*, 2006)), pyramiding different intracellular compartments for drug production (e.g., cytosol with mitochondria and/or chloroplasts (Farhi *et al.*, 2011; Kappers *et al.*, 2005; Wu *et al.*, 2006)), improving the interaction between consecutive enzymatic steps as well as the use of alternative plant systems should further optimize the yield of metabolically engineered artemisinin. The use, for example, of established industrial crops with well-developed large-scale agricultural practice and extraction methods may substantially reduce the cost of drug production.

The nonprofit drug company OneWorld Health (S. San Francisco, CA, USA) declared 2011 a target year for the development of artemisinin precursor in yeast and its subsequent chemical conversion to a semisynthetic drug. Reaching this important goal should lead to a 30–60% cost reduction in artemisinin-based malaria therapy (Hale *et al.*, 2007). Compared with the use of microbial systems for the production of artemisinin-related compounds (Ro *et al.*, 2006; Martin *et al.*, 2003), the plant-based platform presented here enables complete synthesis of artemisinin. Moreover, in light of tobacco's high biomass and rapid growth, coupled with the feasibility of generating commercially viable plant cell cultures, our results pave the way for the development of a sustainable plant-based platform for the production of an anti-malarial drug.

Acknowledgments

All the contents in Chapter 3 have been published in 2012 in *Plant Physiol* 158(1): 132-144, doi: 10.1104/pp.111.184374 with the title: Zinc finger nuclease and homing endonuclease-mediated assembly of multigene plant transformation vectors, and in 2011 in *Nat Biotechnol* 29(12): 1072-1074, doi: 10.1038/nbt.2054 with the title: Generation of the potent anti-malarial drug artemisinin in tobacco. I am the co-first author in the first research article and middle author in the second one.

I would like to thank Tzvi Tzfira for his mentorship and help in design of the experiments and analysis of the data, and Vardit Zeevi (co-author) for our shared work in the construction of the reported multigene binary vectors and generation of the transgenic *Arabidopsis* plants shown in **Figure 3-4** and **Figure 3-5.**

I would also like to thank Uri Arieli for the PCR analysis of T-DNA inserts in offspring of transgenic *Arabidopsis* plants shown in **Figure 3-6,** and Moran Farhi, Elena Marhevka, Julius Ben-Ari and others in the Alexander Vainstein and Betty Schwartz labs for their work in generation and validation of artemisinin production in the transgenic tobacco plants shown in **Figure 3-9**.

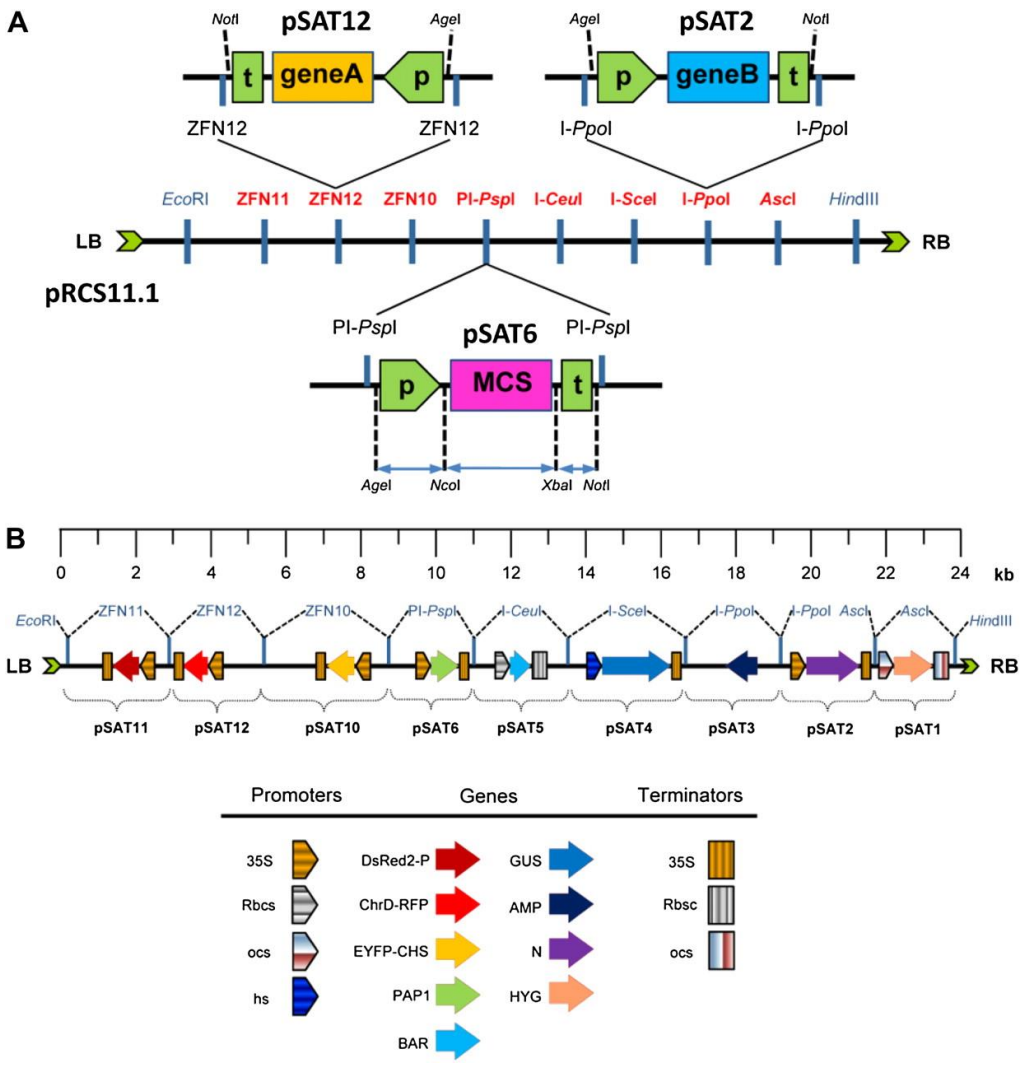


Figure 3-1. General features of ZFN- and homing-endonuclease-mediated multigene binary vector assembly system. (**A**) The general structure of a typical pSAT vector is exemplified by pSAT6, in which the promoter is flanked by the unique *Age*I and *Nco*I sites, the terminator is flanked by the unique *Xba*I and *Not*I sites, the gene of interest is cloned into an extended MCS, and the entire plant expression cassette is flanked by recognition sites for ZFNs (e.g., ZFN12 in pSAT12) or homing endonuclease (e.g., PI-*Psp*I in pSAT6). The ZFN- and homing-endonuclease recognition sites are shown on the binary pRCS11.1. Using ZFNs and homing endonucleases, up to nine expression cassettes can be transferred from the pSAT vectors into the T-DNA region of the pRCS11.1. LB, left border; MCS, multi-cloning site; p, promoter sequence; RB, right border; t, terminator sequence. (**B**) Structure and scale of the pRCS11.1[1.HYG][2.N][3.AMP][4.GUS][5.BAR][6.PAP1][10.EYFP-CHS][11.DsRed2-P][12.ChrD-RFP] nine-transgene-long plant transformation binary vector. The general structure and direction of each expression cassette is shown. Promoters and terminators: 35S, double cauliflower mosaic virus (CaMV) 35S; Rbcs, Rubisco small subunit; ocs, octopine synthase; hs,

hsp18.1 heat shock. Proteins: DsRed2-P, P protein of Sonchus Yellow Net Rhabdovirus fused to DsRed2; ChrD-RFP, chromoplast-specific carotenoid-associated protein ChrD fused to RFP; EYFP-CHS, chalcone synthase fused to EYFP; PAP1, *Arabidopsis* transcription factor production of anthocyanin pigment 1; BAR, Basta resistance encoding gene; GUS, β glucuronidase; AMP, ampicillin resistance; N, N protein of Sonchus Yellow Net Rhabdovirus; HYG, hygromycin resistance.

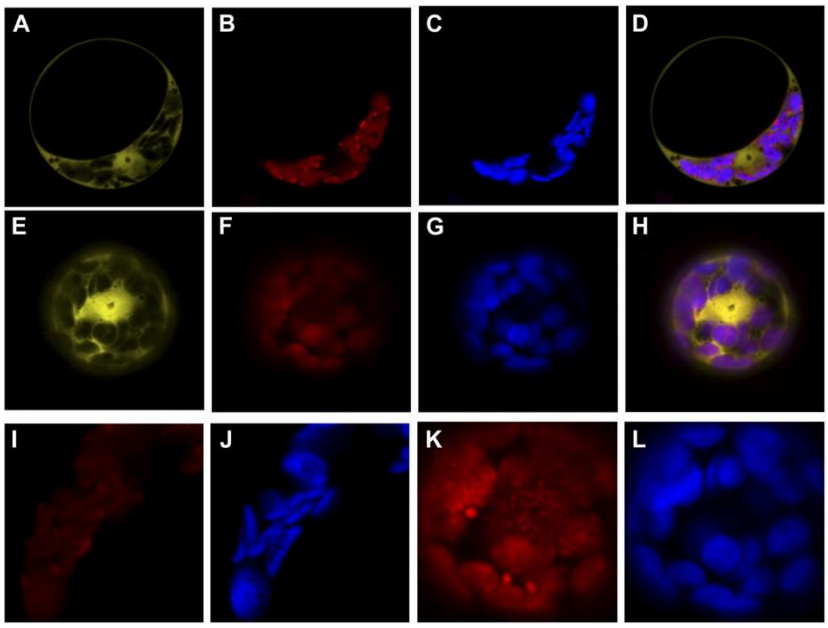


Figure 3-2. Multiple gene expression in protoplasts from triple- and quadruple-gene long binary vectors. EYFP-CHS expression was targeted to the rough ER (shown in yellow in panels **A** and **E**), while ChrD-RFP was targeted to the chloroplasts (shown in red in panels **B**, **F**, **I** and K) and overlapped with the chloroplasts' autofluorescence (shown in blue in panels **C**, **G**, **J** and **L**). DsRed2-P was targeted to the cytoplasm (shown in red in panels **B** and **I**) and was redirected into the nucleus in the presence of free N protein to form sub-nuclear aggregates (shown in red in panels **F** and **K**). Panels **D** and **H** present merged signals of EYFP, DsRed2, RFP and plastid autofluorescence. Panels **I** and **J** are magnifications of panels **B** and **C**; Panels **K** and **L** are magnifications of panels **F** and **G**. All Panels are single confocal sections.

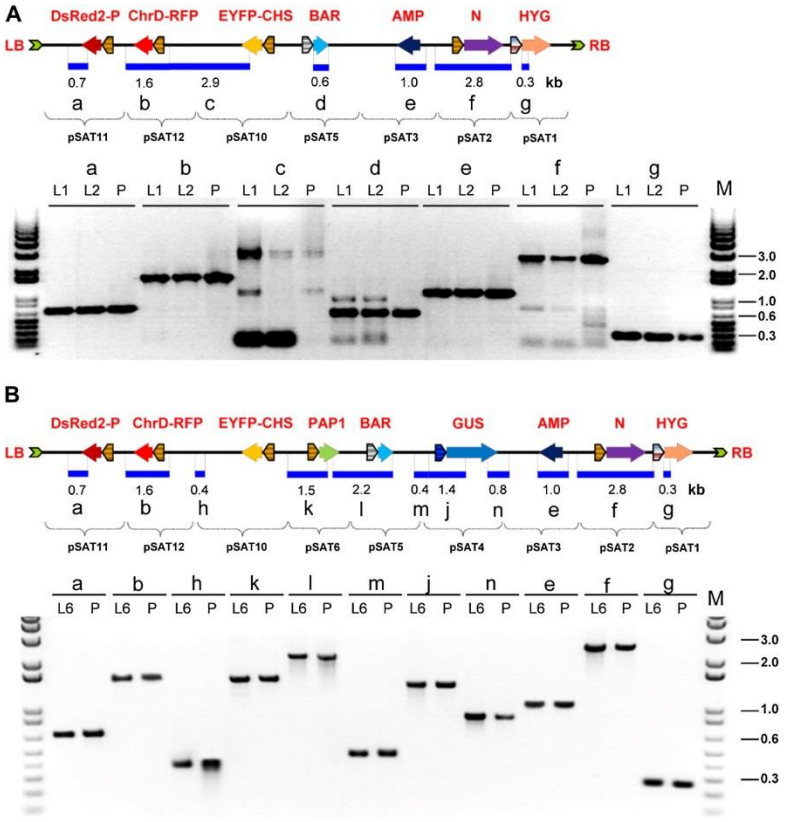


Figure 3-3. PCR analysis of multigene binary T-DNAs in transgenic plants. (**A**) PCR analysis of seven-transgene-long T-DNA inserts in a binary plasmid (P) and two (L1 and L2) transgenic plants. (**B**) PCR analysis of nine-transgene-long T-DNA inserts in a transgenic plant (L6) and in a binary plasmid (P). The general structure and the expected sizes (in kb) of the PCR bands are shown.

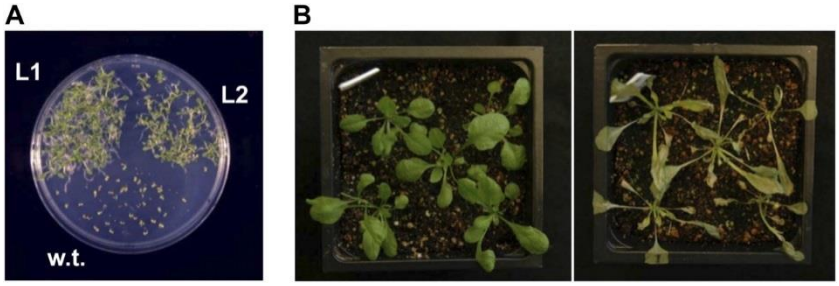


Figure 3-4. Phenotypic analysis seven-transgene-long transgenic *Arabidopsis* plants. (**A**) Hygromycin resistance in two independent transgenic T_1 (L1 and L2) *Arabidopsis* plants. w.t., wild type plants. (**B**) Basta resistance in a hygromycin-resistant transgenic (left) *Arabidopsis* line. On the right, wild type plants.

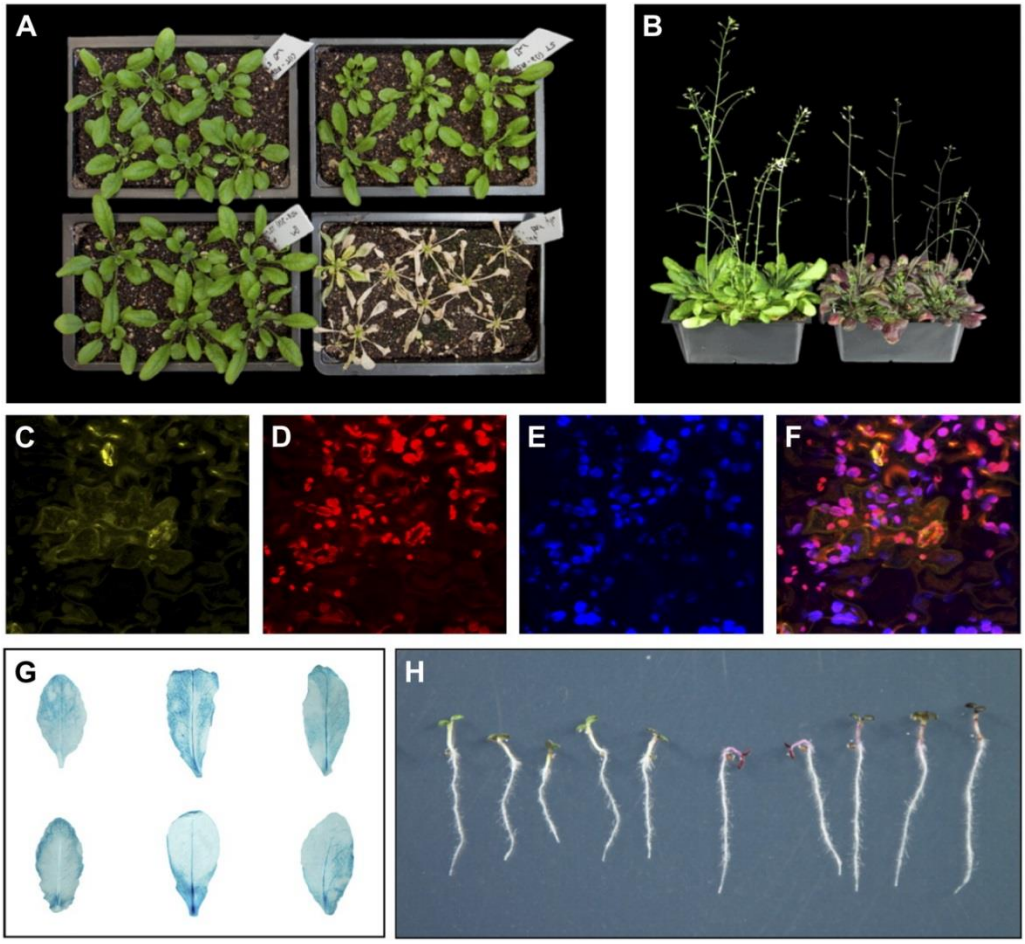


Figure 3-5. Phenotypic analysis of nine-transgene-long transgenic *Arabidopsis* plants. (**A**) Basta resistance in three independent transgenic lines. Clockwise from bottom left: offspring of T_0 L10 plants, offspring of T_1 L6-1 plants, offspring of T_1 L7 plants, wild type plants. (**B**) PAP1 phenotype in offspring of T_2 L3-3-1 plants (right). On the left, wild-type plants. (**C-F**) EYFP-CHS expression (in yellow), ChrD-RFP and DsRed2-P expression (in red), plastid autofluorescence (in blue) and merged signals in leaf of T_1 L6-1 plants. All panels are single confocal sections. (**G**) Heat-shock-induced GUS expression in transgenic leaves. Clockwise from top left: samples from L2-1-1, L2-2-1, L6-1-1, L10-2-1, L10-1-1, L6-2-2. (**H**) PAP1 phenotype in seedlings of offspring of T_2 L6-2-2 plants (left). On the right, seedlings of wild-type plants.

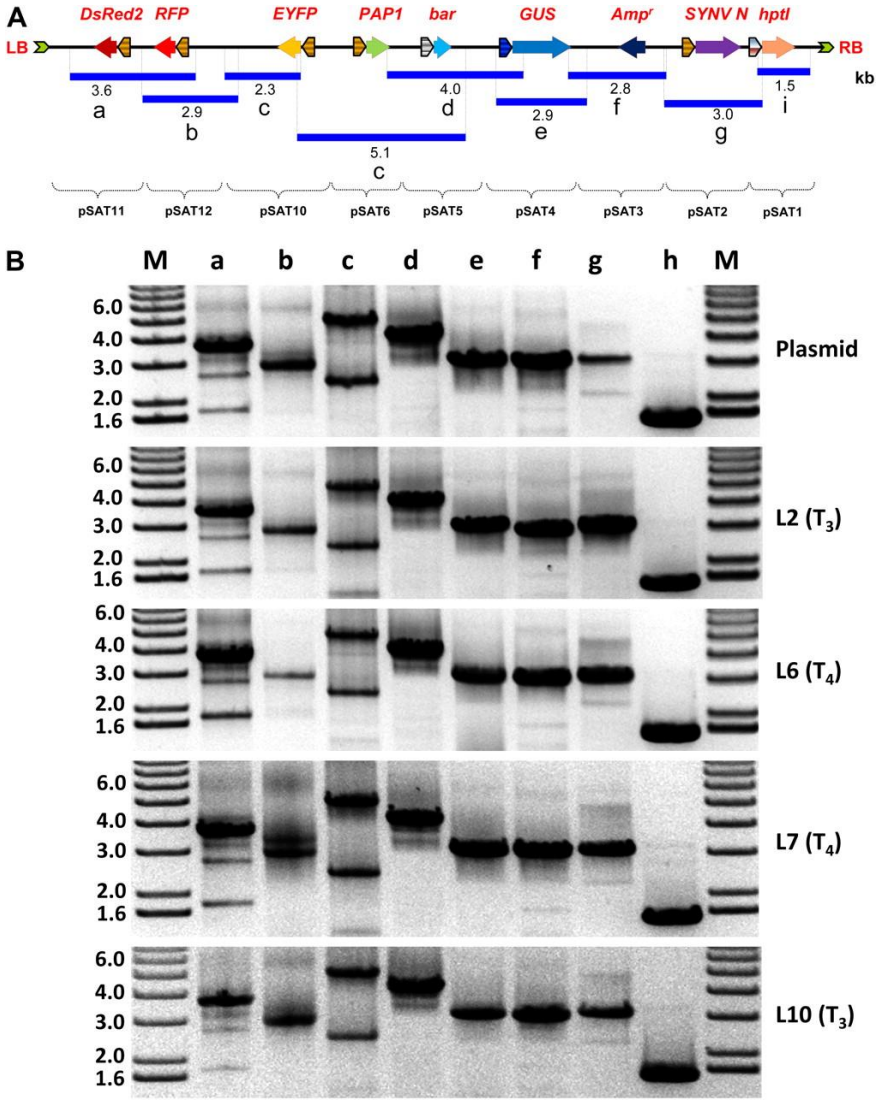


Figure 3-6. PCR analysis of T-DNA inserts in offspring of transgenic *Arabidopsis* plants. (**A**) Structure of the 9-gene-long T-DNA molecule, sizes and locations of the overlapping amplified PCR fragments. (**B**) PCR analysis of binary plasmid, T3 (lines L2 and L10) and T4 (lines L6 and L7) hygromycin-resistant plants. Note the two distinct fragments in lane c, produced using a single pair of primers due to the presence of repetitive elements on the 9-gene-long T-DNA molecule. The molecular marker (M) and expected sizes of the PCR fragments are given in kb.

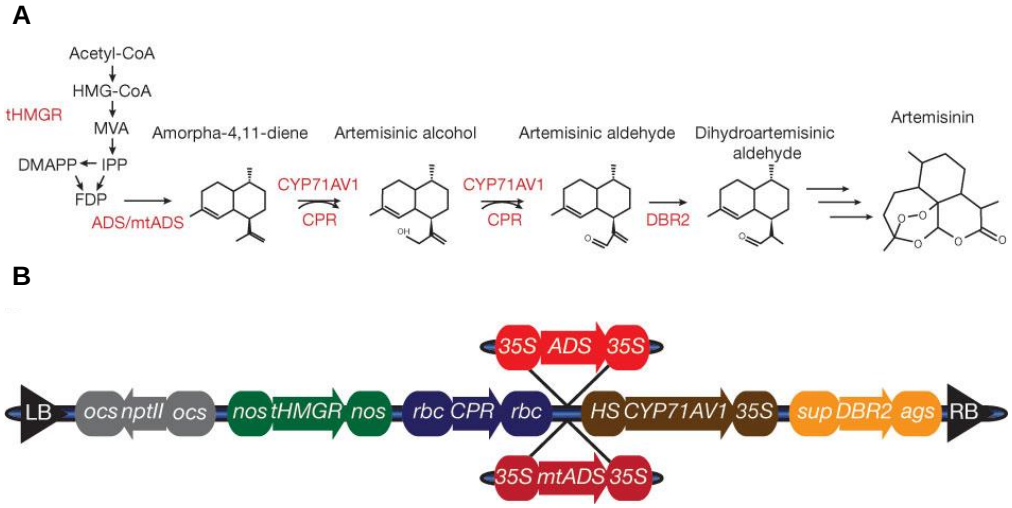


Figure 3-7. The pathway and constructs for engineering artemisinin production in tobacco. (**A**) Schematic outline of the mevalonate and artemisinin pathways (engineered genes are in red). (**B**) Gene constructs assembled to engineer artemisinin production. Arrows indicate genes, boxes indicate promoters and terminators; constructs with either *ADS* or *mtADS* are shown. ocs, octopine synthase; nos, nopaline synthase; rbc, rubisco; 35S, cauliflower mosaic virus (CaMV) 35S; HS, hps18.1 promoter; sup, super-promoter; ags, agrocinopine synthase.

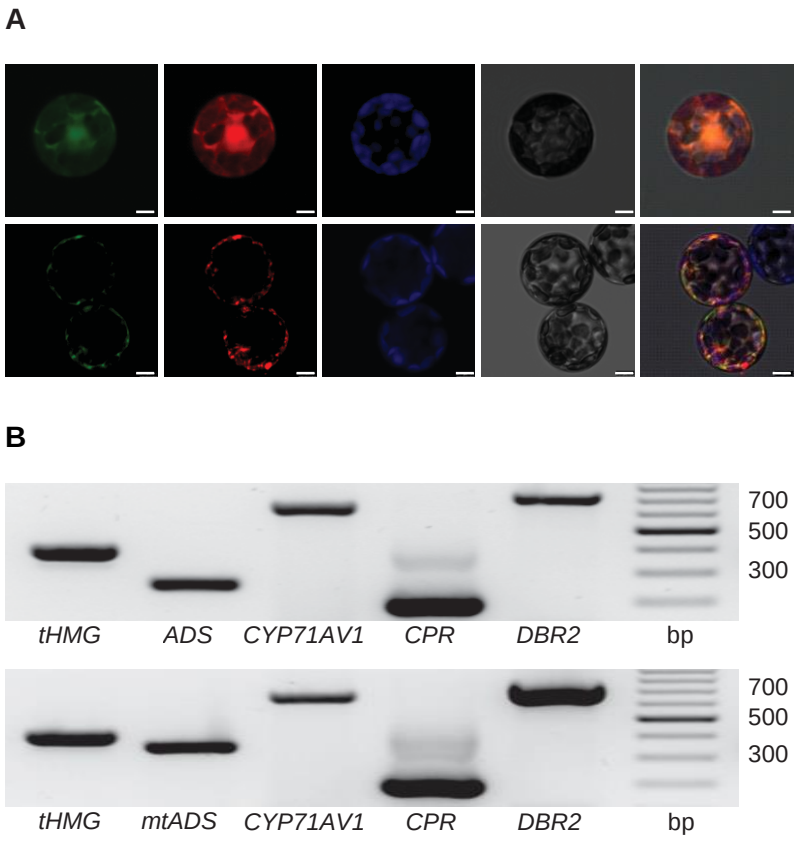


Figure 3-8. Characterization of transgenic tobacco and subcellular localization of ADS and mtADS. (**A**) Transient expression of EGFP fused to ADS (top row) or mtADS (bottom row) in A. thaliana protoplasts. Transient expression of RFP (for cytosol labeling) and MitoTracker-stained mitochondria are in red, EGFP fluorescence is in green, plastid autofluorescence is in blue. Micrographs on far right are merged figures of RFP (top row) or MitoTracker (bottom row) with EGFP, and second to last micrographs are bright-field. (**B**) Expression of tHMG and artemisinin-pathway genes in representative ADS and mtADS N. tabacum plants (upper and lower gel, respectively) determined by RT-PCR.

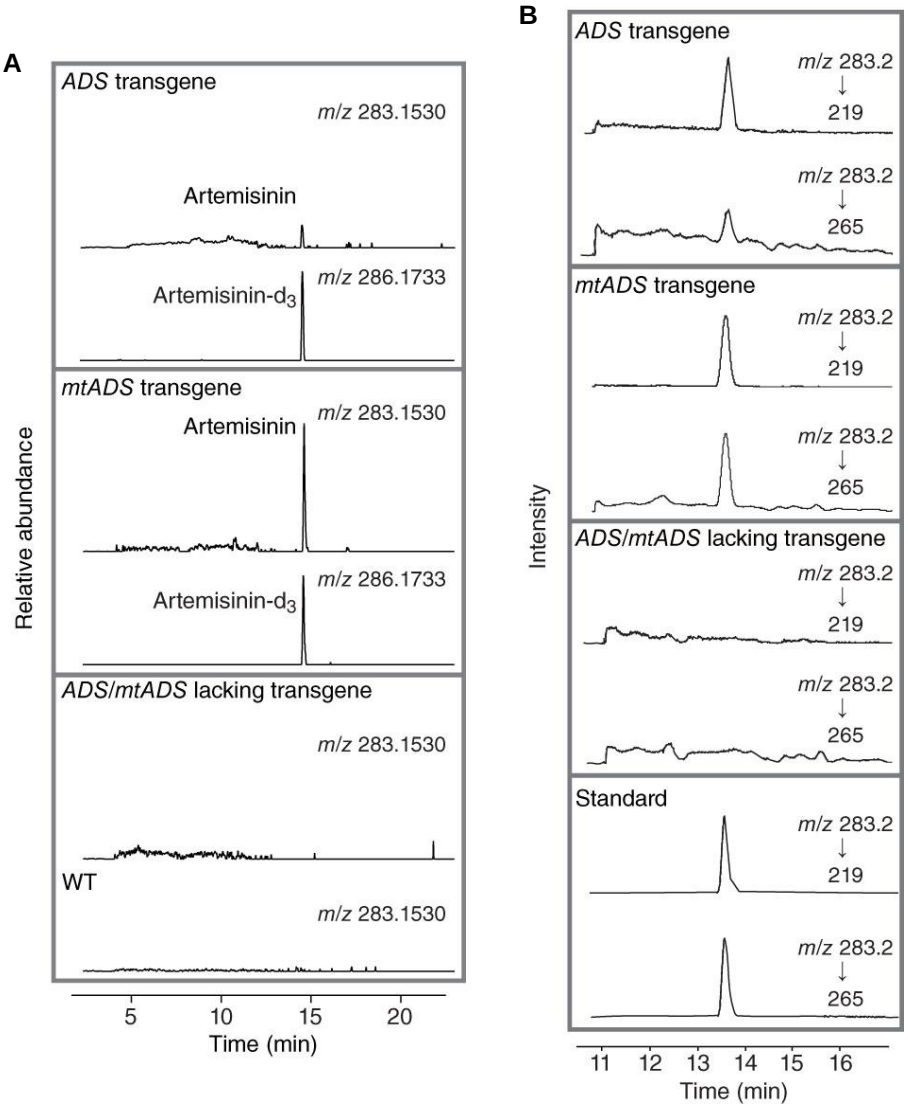


Figure 3-9. Production of artemisinin in engineered tobacco plants. (**A**) Extracted ion accurate mass chromatograms (UPLC-HR-MS) for artemisinin (*m*/*z* 283.1530) and artemisinin-d_3 (*m*/*z* 286.1733) in extracts of plants transformed with the *ADS* (upper panel) or *mtADS* (middle panel) construct. Lower panel shows artemisinin ion in extracts from *ADS*/*mtADS*-lacking transgenic (top chromatogram) and wild-type (bottom chromatogram) control plants. (**B**) Artemisinin production in tobacco extracts analyzed by UPLC-MS/MS operated in MRM mode. Extracts from *ADS*, *mtADS* and *ADS*/*mtADS*-lacking transgenic plants are shown in the upper three panels, respectively. Artemisinin standard is shown in the lower panel. Chromatograms in each panel show artemisinin-specific MRM traces of *m*/*z* 283.2–219 and *m*/*z* 283.2–265.

Table 3-1. Genes, vectors and enzymes used for the construction of multigene vectors, in order of their assembly in pRCS11.1.

pSAT	Protein	Enzyme(s)	Phenotype
10	EYFP-CHS	ZFN10	Decoration of endoplasmic reticulum by yellow fluorescence
11	DsRed2-P	ZFN11	Decoration of the cytoplasm and nucleus by red fluorescence
12	ChrD-RFP	ZFN12	Decoration of the chloroplasts by red fluorescence
2	N	*Asc*I-I-*Ppo*I	Directing DsRed2-P into subnuclear compartments
1	HYG	*Asc*I	Hygromycin resistance
3	AMP	I-*Ppo*I	Bacterial ampicilliinresistance gene and origin of replication
5	BAR	I-*Ceu*I	Basta resistance
4	GUS	I-*Sce*I	Heat shock-induced GUS expression
6	PAP1	PI-*Psp*I	Activation of anthocyanin expression

Table 3-2. Analysis of transgenic Arabidopsis plants across several generations.

T0					T1								T2				T3		
Line	PAP1	DsRed2	HYG[a]	BAR[b]	Line	PAP1	DsRed2	EYFP	RFP	GUS	HYG[a]	BAR[b]	Line	PAP1	GUS	HYG[a]	Line	PAP1	DsRed2
			%								%					%			
L1	+[c]	+	100	ND[d]															
					L1-1	+	+	+	+	+	ND								
L2	+	+	100	+															
					L2-1	+	+	+	+	+	100	+							
													L2-1-1	+	+				
													L2-1-2	+	+				
													L2-1-3	+	+				
					L2-2	+	+	+	+	+	100	+	L2-2-1	+	+				
L3	+	+	100	ND															
					L3-1	+	+	+	+	+	100								
					L3-3	+	+	+	+	+	100	+	L3-3-1	+	ND				
L6	+	+	75	+															
					L6-1	+	+	+	+	+	100	+							
													L6-1-1	+	+	100			
																	L6-1-1-1	+	+
													L6-1-2	+	+	ND			
													L6-1-3	+	+	ND			
				+	L6-2	+	+	+	+	+	75	+							
													L6-2-1	+	+	75			
																	L6-2-1-1	+	+
													L6-2-2	+	+	100			
																	L6-2-2-1	+	+
L7	+	+	75	+															
					L7-1	+	+	+	+	+	75	ND							
													L7-1-1	+	ND	75			
																	L7-1-1-1	+	+
													L7-1-2	+	ND	75			
													L7-1-3	+	ND	100			
				+	L7-2	+	+	+	+	+	75	ND							
													L7-2-1	+	ND				
													L7-2-2	+	ND	75			
L8	+	+	100	ND	L8-1	+	+	+	+	+	ND								
L10	+	+	75	+															
					L10-1	+	+	+	+	+	100	ND							
													L10-1-1	+	ND				
													L10-1-2	+	ND				
					L10-2	+	+	+	+	+										
											75	+							
													L10-2-1	+	+				
													L10-2-2	+	+				
					L10-3	+	+	+	+	+	ND								
					L10-4	+	+	+	+	+	ND								
					L10-5	+	+	+	+	+	ND								

[a] Percentage of hygromycin- resistant plantlets growing on hygromycin-containing germination medium.
[b] Several hygromycin-resistant plantlets were transferred to soil and sprayed with Basta.
[c] '+' denote positive phenotype.
[d] ND, not determined.

Reference

Ben Zvi, M.M., Negre-Zakharov, F., Masci, T., Ovadis, M., Shklarman, E., Ben-Meir, H., Tzfira, T., Dudareva, N., and Vainstein, A. (2008). Interlinking showy traits: co-engineering of scent and colour biosynthesis in flowers. Plant Biotechnol J *6*, 403-415.

Bohmert, K., Balbo, I., Kopka, J., Mittendorf, V., Nawrath, C., Poirier, Y., Tischendorf, G., Trethewey, R.N., and Willmitzer, L. (2000). Transgenic Arabidopsis plants can accumulate polyhydroxybutyrate to up to 4% of their fresh weight. Planta *211*, 841-845.

Bohmert, K., Balbo, I., Steinbuchel, A., Tischendorf, G., and Willmitzer, L. (2002). Constitutive expression of the beta-ketothiolase gene in transgenic plants. A major obstacle for obtaining polyhydroxybutyrate-producing plants. Plant Physiol *128*, 1282-1290.

Borevitz, J.O., Xia, Y., Blount, J., Dixon, R.A., and Lamb, C. (2000). Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. Plant Cell *12*, 2383-2394.

Brown, G.D. (2010). The biosynthesis of artemisinin (Qinghaosu) and the phytochemistry of Artemisia annua L. (Qinghao). Molecules *15*, 7603-7698.

Cai, C.Q., Doyon, Y., Ainley, W.M., Miller, J.C., Dekelver, R.C., Moehle, E.A., Rock, J.M., Lee, Y.L., Garrison, R., Schulenberg, L., *et al.* (2009). Targeted transgene integration in plant cells using designed zinc finger nucleases. Plant Mol Biol *69*, 699-709.

Carroll, D., Beumer, K.J., Morton, J.J., Bozas, A., and Trautman, J.K. (2008). Gene targeting in Drosophila and Caenorhabditis elegans with zinc-finger nucleases. Methods Mol Biol *435*, 63-77.

Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res *39*, e82.

Chakrabarty, R., Banerjee, R., Chung, S.M., Farman, M., Citovsky, V., Hogenhout, S.A., Tzfira, T., and Goodin, M. (2007). PSITE vectors for stable integration or transient expression of autofluorescent protein fusions in plants: probing Nicotiana benthamiana-virus interactions. Mol Plant Microbe Interact *20*, 740-750.

Chen, L., Marmey, P., Taylor, N.J., Brizard, J.P., Espinoza, C., D'Cruz, P., Huet, H., Zhang, S., de Kochko, A., Beachy, R.N., *et al.* (1998). Expression and inheritance of multiple transgenes in rice plants. Nat Biotechnol *16*, 1060-1064.

Chen, Q.J., Xie, M., Ma, X.X., Dong, L., Chen, J., and Wang, X.C. (2010). MISSA is a highly efficient in vivo DNA assembly method for plant multiple-gene transformation. Plant Physiol *153*, 41-51.

Chen, Q.J., Zhou, H.M., Chen, J., and Wang, X.C. (2006). A Gateway-based platform for multigene plant transformation. Plant Mol Biol *62*, 927-936.

Cheo, D.L., Titus, S.A., Byrd, D.R., Hartley, J.L., Temple, G.F., and Brasch, M.A. (2004). Concerted assembly and cloning of multiple DNA segments using in vitro site-specific recombination: functional analysis of multi-segment expression clones. Genome Res *14*, 2111-2120.

Chilton, M.D., and Que, Q. (2003). Targeted integration of T-DNA into the tobacco genome at double-stranded breaks: new insights on the mechanism of T-DNA integration. Plant Physiol *133*, 956-965.

Chung, S.M., Frankman, E.L., and Tzfira, T. (2005). A versatile vector system for multiple gene expression in plants. Trends Plant Sci *10*, 357-361.

Citovsky, V., Lee, L.Y., Vyas, S., Glick, E., Chen, M.H., Vainstein, A., Gafni, Y., Gelvin, S.B., and Tzfira, T. (2006). Subcellular localization of interacting proteins by bimolecular fluorescence complementation in planta. J Mol Biol *362*, 1120-1131.

Clough, S.J., and Bent, A.F. (1998). Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant J *16*, 735-743.

Covello, P.S. (2008). Making artemisinin. Phytochemistry *69*, 2881-2885.

Dafny-Yelin, M., Chung, S.M., Frankman, E.L., and Tzfira, T. (2007). pSAT RNA interference vectors: a modular series for multiple gene down-regulation in plants. Plant Physiol *145*, 1272-1281.

Dafny-Yelin, M., and Tzfira, T. (2007). Delivery of multiple transgenes to plant cells. Plant Physiol *145*, 1118-1128.

Daniell, H., and Dhingra, A. (2002). Multigene engineering: dawn of an exciting new era in biotechnology. Curr Opin Biotechnol *13*, 136-141.

De Neve, M., De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1997). T-DNA integration patterns in co-transformed plant cells suggest that T-DNA repeats originate from co-integration of separate T-DNAs. Plant J *11*, 15-29.

Farhi, M., Marhevka, E., Masci, T., Marcos, E., Eyal, Y., Ovadis, M., Abeliovich, H., and Vainstein, A. (2011). Harnessing yeast subcellular compartments for the production of plant terpenoids. Metab Eng *13*, 474-481.

Frary, A., and Hamilton, C.M. (2001). Efficiency and stability of high molecular weight DNA transformation: an analysis in tomato. Transgenic Res *10*, 121-132.

Fujisawa, M., Takita, E., Harada, H., Sakurai, N., Suzuki, H., Ohyama, K., Shibata, D., and Misawa, N. (2009). Pathway engineering of Brassica napus seeds using multiple key enzyme genes involved in ketocarotenoid formation. J Exp Bot *60*, 1319-1332.

Gallois, P., and Marinho, P. (1995). Leaf Disk Transformation Using Agrobacterium tumefaciens-Expression of Heterologous Genes in Tobacco. In Plant Gene Transfer and Expression Protocols, H. Jones, ed. (Springer New York), pp. 39-48.

Gao, H., Smith, J., Yang, M., Jones, S., Djukanovic, V., Nicholson, M.G., West, A., Bidney, D., Falco, S.C., Jantz, D., *et al.* (2010). Heritable targeted mutagenesis in maize using a designed endonuclease. Plant J *61*, 176-187.

Geurts, A.M., Cost, G.J., Freyvert, Y., Zeitler, B., Miller, J.C., Choi, V.M., Jenkins, S.S., Wood, A., Cui, X., Meng, X., *et al.* (2009). Knockout rats via embryo microinjection of zinc-finger nucleases. Science *325*, 433.

Goderis, I.J., De Bolle, M.F., Francois, I.E., Wouters, P.F., Broekaert, W.F., and Cammue, B.P. (2002). A set of modular plant transformation vectors allowing flexible insertion of up to six expression units. Plant Mol Biol *50*, 17-27.

Goodin, M.M., Austin, J., Tobias, R., Fujita, M., Morales, C., and Jackson, A.O. (2001). Interactions and nuclear import of the N and P proteins of sonchus yellow net virus, a plant nucleorhabdovirus. J Virol *75*, 9393-9406.

Goodin, M.M., Dietzgen, R.G., Schichnes, D., Ruzin, S., and Jackson, A.O. (2002). pGD vectors: versatile tools for the expression of green and red fluorescent protein fusions in agroinfiltrated plant leaves. Plant J *31*, 375-383.

Graham, I.A., Besser, K., Blumer, S., Branigan, C.A., Czechowski, T., Elias, L., Guterman, I., Harvey, D., Isaac, P.G., Khan, A.M., *et al.* (2010). The genetic map of Artemisia annua L. identifies loci affecting yield of the antimalarial drug artemisinin. Science *327*, 328-331.

Hale, V., Keasling, J.D., Renninger, N., and Diagana, T.T. (2007). Microbially derived artemisinin: a biotechnology solution to the global problem of access to affordable antimalarial drugs. Am J Trop Med Hyg *77*, 198-202.

Halpin, C., Barakate, A., Askari, B.M., Abbott, J.C., and Ryan, M.D. (2001). Enabling technologies for manipulating multiple genes on complex pathways. Plant Mol Biol *47*, 295-310.

Hamilton, C.M. (1997). A binary-BAC system for plant transformation with high-molecular-weight DNA. Gene *200*, 107-116.

Hanin, M., and Paszkowski, J. (2003). Plant genome modification by homologous recombination. Curr Opin Plant Biol *6*, 157-162.

Hellens, R., Mullineaux, P., and Klee, H. (2000). Technical Focus:a guide to Agrobacterium binary Ti vectors. Trends Plant Sci *5*, 446-451.

Kappers, I.F., Aharoni, A., van Herpen, T.W., Luckerhoff, L.L., Dicke, M., and Bouwmeester, H.J. (2005). Genetic engineering of terpenoid metabolism attracts bodyguards to Arabidopsis. Science *309*, 2070-2072.

Karimi, M., De Meyer, B., and Hilson, P. (2005). Modular cloning in plant cells. Trends Plant Sci *10*, 103-105.

Lee, L.Y., Fang, M.J., Kuang, L.Y., and Gelvin, S.B. (2008). Vectors for multi-color bimolecular fluorescence complementation to investigate protein-protein interactions in living plant cells. Plant Methods *4*, 24.

Lin, L., Liu, Y.G., Xu, X., and Li, B. (2003). Efficient linking and transfer of multiple genes by a multigene assembly and transformation vector system. Proceedings of the National Academy of Sciences of the United States of America *100*, 5962-5967.

Liu, Y.G., Shirano, Y., Fukaki, H., Yanai, Y., Tasaka, M., Tabata, S., and Shibata, D. (1999). Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. Proceedings of the National Academy of Sciences of the United States of America *96*, 6535-6540.

Lyznik, L.A., and Dress, V. (2008). Gene targeting for chromosome engineering applications in eukaryotic cells. Recent Pat Biotechnol *2*, 94-106.

Ma, C., Wang, H., Lu, X., Wang, H., Xu, G., and Liu, B. (2009). Terpenoid metabolic profiling analysis of transgenic Artemisia annua L. by comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. Metabolomics *5*, 497-506.

Maeder, M.L., Thibodeau-Beganny, S., Osiak, A., Wright, D.A., Anthony, R.M., Eichtinger, M., Jiang, T., Foley, J.E., Winfrey, R.J., Townsend, J.A., *et al.* (2008). Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. Mol Cell *31*, 294-301.

Mahfouz, M.M., Li, L., Shamimuzzaman, M., Wibowo, A., Fang, X., and Zhu, J.K. (2011). De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. Proceedings of the National Academy of Sciences of the United States of America *108*, 2623-2628.

Mani, M., Kandavelou, K., Dy, F.J., Durai, S., and Chandrasegaran, S. (2005). Design, engineering, and characterization of zinc finger nucleases. Biochem Biophys Res Commun *335*, 447-457.

Martin, V.J., Pitera, D.J., Withers, S.T., Newman, J.D., and Keasling, J.D. (2003). Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. Nat Biotechnol *21*, 796-802.

Matzke, A.J., and Matzke, M.A. (1998). Position effects and epigenetic silencing of plant transgenes. Curr Opin Plant Biol *1*, 142-148.

Moon, H.S., Li, Y., and Stewart, C.N., Jr. (2010). Keeping the genie in the bottle: transgene biocontainment by excision in pollen. Trends Biotechnol *28*, 3-8.

Naqvi, S., Farre, G., Sanahuja, G., Capell, T., Zhu, C., and Christou, P. (2010). When more is better: multigene engineering in plants. Trends Plant Sci *15*, 48-56.

Naqvi, S., Zhu, C., Farre, G., Ramessar, K., Bassie, L., Breitenbach, J., Perez Conesa, D., Ros, G., Sandmann, G., Capell, T., *et al.* (2009). Transgenic multivitamin corn through biofortification of endosperm with three vitamins representing three distinct metabolic pathways. Proceedings of the National Academy of Sciences of the United States of America *106*, 7762-7767.

Ni, M., Cui, D., Einstein, J., Narasimhulu, S., Vergara, C.E., and Gelvin, S.B. (1995). Strength and tissue specificity of chimeric promoters derived from the octopine and mannopine synthase genes. The Plant Journal *7*, 661-676.

Pelletier, M.K., and Shirley, B.W. (1996). Analysis of flavanone 3-hydroxylase in Arabidopsis seedlings. Coordinate regulation with chalcone synthase and chalcone isomerase. Plant Physiol *111*, 339-345.

Peremarti, A., Twyman, R.M., Gomez-Galera, S., Naqvi, S., Farre, G., Sabalza, M., Miralpeix, B., Dashevskaya, S., Yuan, D., Ramessar, K., *et al.* (2010). Promoter diversity in multigene transformation. Plant Mol Biol *73*, 363-378.

Petolino, J.F., Worden, A., Curlee, K., Connell, J., Strange Moynahan, T.L., Larsen, C., and Russell, S. (2010). Zinc finger nuclease-mediated transgene deletion. Plant Mol Biol *73*, 617-628.

Porteus, M.H. (2009). Plant biotechnology: Zinc fingers on target. Nature *459*, 337-338.

Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., *et al.* (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature *440*, 940-943.

Salomon, S., and Puchta, H. (1998). Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. EMBO J *17*, 6086-6095.

Sasaki, Y., Sone, T., Yoshida, S., Yahata, K., Hotta, J., Chesnut, J.D., Honda, T., and Imamoto, F. (2004). Evidence for high specificity and efficiency of multiple recombination signals in mixed DNA cloning by the Multisite Gateway system. J Biotechnol *107*, 233-243.

Schramek, N., Wang, H., Romisch-Margl, W., Keil, B., Radykewicz, T., Winzenhorlein, B., Beerhues, L., Bacher, A., Rohdich, F., Gershenzon, J., *et al.* (2010). Artemisinin biosynthesis in growing plants of Artemisia annua. A 13CO2 study. Phytochemistry *71*, 179-187.

Shukla, V.K., Doyon, Y., Miller, J.C., DeKelver, R.C., Moehle, E.A., Worden, S.E., Mitchell, J.C., Arnold, N.L., Gopalan, S., Meng, X., *et al.* (2009). Precise genome modification in the crop species Zea mays using zinc-finger nucleases. Nature *459*, 437-441.

Spitzer-Rimon, B., Marhevka, E., Barkai, O., Marton, I., Edelbaum, O., Masci, T., Prathapani, N.K., Shklarman, E., Ovadis, M., and Vainstein, A. (2010). EOBII, a gene encoding a flower-specific regulator of phenylpropanoid volatiles' biosynthesis in petunia. Plant Cell *22*, 1961-1976.

Takahashi, T., and Komeda, Y. (1989). Characterization of two genes encoding small heat-shock proteins in Arabidopsis thaliana. Mol Gen Genet *219*, 365-372.

Takasu, Y., Kobayashi, I., Beumer, K., Uchino, K., Sezutsu, H., Sajwan, S., Carroll, D., Tamura, T., and Zurovec, M. (2010). Targeted mutagenesis in the silkworm Bombyx mori using zinc finger nuclease mRNA injection. Insect Biochem Mol Biol *40*, 759-765.

Taverniers, I., Papazova, N., Bertheau, Y., De Loose, M., and Holst-Jensen, A. (2008). Gene stacking in transgenic plants: towards compliance between definitions, terminology, and detection within the EU regulatory framework. Environ Biosafety Res *7*, 197-218.

Tovkach, A., Zeevi, V., and Tzfira, T. (2009). A toolbox and procedural notes for characterizing novel zinc finger nucleases for genome editing in plant cells. Plant J *57*, 747-757.

Tovkach, A., Zeevi, V., and Tzfira, T. (2010). Validation and expression of zinc finger nucleases in plant cells. Methods Mol Biol *649*, 315-336.

Tovkach, A., Zeevi, V., and Tzfira, T. (2011). Expression, purification and characterization of cloning-grade zinc finger nuclease. J Biotechnol *151*, 1-8.

Towler, M.J., and Weathers, P.J. (2007). Evidence of artemisinin production from IPP stemming from both the mevalonate and the nonmevalonate pathways. Plant Cell Rep *26*, 2129-2136.

Townsend, J.A., Wright, D.A., Winfrey, R.J., Fu, F., Maeder, M.L., Joung, J.K., and Voytas, D.F. (2009). High-frequency modification of plant genes using engineered zinc-finger nucleases. Nature *459*, 442-445.

Tzfira, T., Frankman, L.R., Vaidya, M., and Citovsky, V. (2003). Site-specific integration of Agrobacterium tumefaciens T-DNA via double-stranded intermediates. Plant Physiol *133*, 1011-1023.

Tzfira, T., Jensen, C., Wang, W., Zuker, A., Vinocur, B., Altman, A., and Vainstein, A. (1997). Transgenic Populus tremula: a step-by-step protocol for its Agrobacterium-mediated transformation. Plant Molecular Biology Reporter *15*, 219-235.

Tzfira, T., Tian, G.W., Lacroix, B., Vyas, S., Li, J., Leitner-Dagan, Y., Krichevsky, A., Taylor, T., Vainstein, A., and Citovsky, V. (2005). pSAT vectors: a modular series of plasmids for autofluorescent protein tagging and expression of multiple genes in plants. Plant Mol Biol *57*, 503-516.

Urnov, F.D., Miller, J.C., Lee, Y.L., Beausejour, C.M., Rock, J.M., Augustus, S., Jamieson, A.C., Porteus, M.H., Gregory, P.D., and Holmes, M.C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature *435*, 646-651.

van Engelen, F.A., Molthoff, J.W., Conner, A.J., Nap, J.P., Pereira, A., and Stiekema, W.J. (1995). pBINPLUS: an improved plant transformation vector based on pBIN19. Transgenic Res *4*, 288-290.

van Herpen, T.W., Cankar, K., Nogueira, M., Bosch, D., Bouwmeester, H.J., and Beekwilder, J. (2010). Nicotiana benthamiana as a production platform for artemisinin precursors. PLoS One *5*, e14222.

Vishnevetsky, M., Ovadis, M., Itzhaki, H., Levy, M., Libal-Weksler, Y., Adam, Z., and Vainstein, A. (1996). Molecular cloning of a carotenoid-associated protein from Cucumis sativus corollas: homologous genes involved in carotenoid sequestration in chromoplasts. Plant J *10*, 1111-1118.

Vishnevetsky, M., Ovadis, M., and Vainstein, A. (1999). Carotenoid sequestration in plants: the role of carotenoid-associated proteins. Trends Plant Sci *4*, 232-235.

Wakasa, Y., Yasuda, H., and Takaiwa, F. (2006). High accumulation of bioactive peptide in transgenic rice seeds by expression of introduced multiple genes. Plant Biotechnol J *4*, 499-510.

Weinthal, D., Tovkach, A., Zeevi, V., and Tzfira, T. (2010). Genome editing in plant cells by zinc finger nucleases. Trends Plant Sci *15*, 308-321.

Weinthal, D.M., Taylor, R.A., and Tzfira, T. (2013). Nonhomologous end joining-mediated gene replacement in plant cells. Plant Physiol *162*, 390-400.

Wu, F.H., Shen, S.C., Lee, L.Y., Lee, S.H., Chan, M.T., and Lin, C.S. (2009). Tape-Arabidopsis Sandwich - a simpler Arabidopsis protoplast isolation method. Plant Methods *5*, 16.

Wu, G., Truksa, M., Datla, N., Vrinten, P., Bauer, J., Zank, T., Cirpus, P., Heinz, E., and Qiu, X. (2005). Stepwise engineering to produce high yields of very long-chain polyunsaturated fatty acids in plants. Nat Biotechnol *23*, 1013-1017.

Wu, S., Schalk, M., Clark, A., Miles, R.B., Coates, R., and Chappell, J. (2006). Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. Nat Biotechnol *24*, 1441-1447.

Zeevi, V., Tovkach, A., and Tzfira, T. (2008). Increasing cloning possibilities using artificial zinc finger nucleases. Proceedings of the National Academy of Sciences of the United States of America *105*, 12785-12790.

Zeevi, V., Tovkach, A., and Tzfira, T. (2010). Artificial zinc finger nucleases for DNA cloning. Methods Mol Biol *649*, 209-225.

Zhang, Y., Nowak, G., Reed, D.W., and Covello, P.S. (2011). The production of artemisinin precursors in tobacco. Plant Biotechnol J *9*, 445-454.

Zhang, Y., Teoh, K.H., Reed, D.W., Maes, L., Goossens, A., Olson, D.J., Ross, A.R., and Covello, P.S. (2008). The molecular cloning of artemisinic aldehyde Delta11(13) reductase and its role in glandular trichome-dependent biosynthesis of artemisinin in Artemisia annua. J Biol Chem *283*, 21501-21508.

Zhong, R., Richardson, E.A., and Ye, Z.H. (2007). Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. Planta *225*, 1603-1611.

Zhu, C., Naqvi, S., Breitenbach, J., Sandmann, G., Christou, P., and Capell, T. (2008). Combinatorial genetic transformation generates a library of metabolic phenotypes for the carotenoid pathway in maize. Proceedings of the National Academy of Sciences of the United States of America *105*, 18232-18237.

Zimmerman, K.A., Fischer, K.P., Joyce, M.A., and Tyrrell, D.L. (2008). Zinc finger proteins designed to specifically target duck hepatitis B virus covalently closed circular DNA inhibit viral transcription in tissue culture. J Virol *82*, 8013-8021.

CHAPTER 4

Saccharomyces Cerevisiae DNA Ligase IV Supports Imprecise End Joining Independently of Its Catalytic Activity

Abstract

DNA ligase IV (Dnl4 in budding yeast) is a specialized ligase used in non-homologous end joining (NHEJ) of DNA double-strand breaks (DSBs). Although point and truncation mutations arise in the human ligase IV syndrome, the roles of Dnl4 in DSB repair have mainly been examined using gene deletions. Here Dnl4 catalytic point mutants were generated that were severely defective in auto-adenylation *in vitro* and NHEJ activity *in vivo*, despite being hyper-recruited to DSBs and supporting wild-type levels of Lif1 interaction and assembly of a Ku- and Lif1-containing complex at DSBs. Interestingly, residual levels of especially imprecise NHEJ were markedly higher in a deletion-based assay with Dnl4 catalytic mutants than with a gene deletion strain, suggesting a role of DSB-bound Dnl4 in supporting a mode of NHEJ catalyzed by a different ligase. Similarly, next generation sequencing of repair joints in a distinct single-DSB assay showed that *dnl4*-K466A mutation conferred a significantly different imprecise joining profile than wild-type Dnl4 and that such repair was rarely observed in the absence of Dnl4. Enrichment of DNA ligase I (Cdc9 in yeast) at DSBs was observed in wild-type as well as *dnl4* point mutant strains, with both Dnl4 and Cdc9 disappearing from DSBs upon 5' resection that was unimpeded by the presence of catalytically inactive Dnl4. These findings indicate that

Dnl4 can promote mutagenic end joining independently of its catalytic activity, likely by a mechanism that involves Cdc9.

Introduction

DNA double-strand breaks (DSBs) are potentially catastrophic chromosomal lesions. Accordingly, many proteins are recruited to DSBs for repair by non-homologous end joining (NHEJ) or homologous recombination (HR). NHEJ, a pathway conserved from yeast to humans, repairs DSBs by processing and directly ligating the DNA ends, often with little or no nucleotide loss (Lieber, 2010; Chiruvella *et al.*, 2013). In *Saccharomyces cerevisiae*, the core components of the canonical NHEJ (c-NHEJ) machinery are the Ku (Yku70–Yku80), MRX (Mre11–Rad50–Xrs2) and DNA ligase IV (Dnl4–Lif1–Nej1) complexes (Daley *et al.*, 2005). The final step in c-NHEJ is DNA ligation, carried out by the DNA ligase IV catalytic subunit (Dnl4 in yeast, Lig4 in humans) in complex with its scaffolding protein XRCC4 (Lif1 in yeast) and supported by the XRCC4-like factor XLF (Nej1 in yeast) (Lieber, 2010; Daley *et al.*, 2005; Ellenberger and Tomkinson, 2008).

Dnl4/Lig4 is an ATP-dependent DNA ligase that functions exclusively in NHEJ (Lieber, 2010; Daley *et al.*, 2005; Ellenberger and Tomkinson, 2008). The ATP-dependent DNA ligation mechanism has three distinct catalytic steps (Ellenberger and Tomkinson, 2008). In step 1, the enzyme displaces pyrophosphate from ATP leading to covalent auto-adenylation of the ligase on its active site lysine (K282 in Dnl4), a reaction that can be reversed by incubating adenylated enzyme with excess pyrophosphate. This intermediate is long-lived such that most cellular ligase molecules are adenylated. In step 2, the activated AMP is transferred to a DNA 5’ phosphate via a 5’-5’ phosphoanhydride bond, which is finally cleaved in step 3 by the attack of an adjacent 3’ hydroxyl leading to DNA ligation and release of AMP. DNA ligases undergo profound conformational changes upon ATP and DNA binding (Pascal *et al.*, 2004; Pascal *et al.*, 2006).

While the above properties are common to all DNA ligases, Dnl4/Lig4 has other unique but poorly understood properties presumably related to its role in DSB ligation, including an ability to ligate incompatible ends (Ma *et al.*, 2004) and a slow rate of auto-adenylation (Riballo *et al.*, 2009).

There are many potential consequences of failed c-NHEJ ligation. Most importantly, 5' resection of persistent DSB ends generates 3'-terminated single-stranded DNA tails that are essential for HR (Ira *et al.*, 2004; Symington and Gautier, 2011). Loss of c-NHEJ proteins in yeast, including Dnl4, has been reported to lead to increased rates of 5' resection and HR (Zhang *et al.*, 2007; Clerici *et al.*, 2008), suggesting that either the bound c-NHEJ protein complex and/or c-NHEJ ligation is in competition with HR. Additionally, because error-free HR is not always possible, DSB-derived deletions and other mutations catalyzed by alternative NHEJ (alt-NHEJ) pathways (Mladenov and Iliakis, 2011; Simsek *et al.*, 2011; Della-Maria *et al.*, 2011; Boboila *et al.*, 2012) or catastrophic chromosome loss and cell death can result from c-NHEJ ligation failure. The importance of these negative consequences is underscored by the DNA ligase IV deficiency syndrome, in which human patients with impaired Lig4 function display microcephaly, dysmorphology, developmental delay, bone marrow and immune deficiency, radiosensitivity and malignancy (O'Driscoll *et al.*, 2001; Chistiakov, 2010).

The balance of the above outcomes might depend substantially on the reason for c-NHEJ ligation failure. However, most genetic studies of Dnl4/Lig4 function have been performed with gene deletion mutants, while most human mutations and polymorphisms are point changes (Rucci *et al.*, 2010). For these reasons, we have performed mutational analysis of Dnl4 in *S. cerevisiae*, where an extensive set of outcomes can be assessed. We describe a distinct and

mutagenic NHEJ repair behavior of a series of mutations in the ligase catalytic domain as compared to complete loss of the Dnl4 protein. Despite being severely catalytically defective, these mutants, along with the c-NHEJ complex including Ku and Lif1, hyper-accumulated at induced DSBs and supported imprecise NHEJ in some assays. Cdc9 likely represents the alternative ligase used in the presence of dysfunctional Dnl4, as evidenced by its appearance there in a time frame consistent with NHEJ catalysis. Finally, both Dnl4 mutants and Cdc9 were efficiently removed from DSBs and presented no impediment to 5' resection, suggesting an active switch from NHEJ to HR.

Materials and Methods

Protein structural analysis. Protein structural analysis was performed using The PyMOL Molecular Graphics System, Schrödinger, LLC.

Yeast growth and manipulation. Yeast strains used for NHEJ and chromatin immunoprecipitation (ChIP) assays (**Table 4-3**) were isogenic derivatives of BY4741 (Brachmann *et al.*, 1998) as previously described (Wu *et al.*, 2008; Palmbos *et al.*, 2008). Point mutations were constructed in the native *DNL4* gene using a pop-in/pop-out method (Palmbos *et al.*, 2005). L750* truncation was created as a stop codon after the indicated residue. All mutant alleles were confirmed by sequencing. Yeast were grown at 30^0C in either rich medium containing 1% yeast extract, 2% peptone and 40 μg/ml adenine (YPA) or synthetic defined (SD) medium with either 2% glucose, 2% galactose or 3% glycerol as the carbon source (Wu *et al.*, 2008; Palmbos *et al.*, 2008).

Yeast two-hybrid. The two hybrid indicator strain PJ69-4a/α, vectors pODB2 and pOAD, and full-length and truncated Dnl4 and Lif1 derivatives have been described (Palmbos *et al.*, 2005; Uetz *et al.*, 2000). Dnl4 point mutant derivatives were made by transferring mutant chromosomal alleles into plasmids by gap repair (Palmbos *et al.*, 2005). Dnl4-Lif1 interaction was monitored by mating to strains carrying Lif1 constructs and spotting overnight cultures to plates lacking either histidine or adenine followed by 3 and 5 days growth, respectively.

Plasmid recircularization assay. pRS316 (Brachmann *et al.*, 1998) was cut with *Nco*I to create a 5' overhanging DSB within its *URA3* marker. pK1827 (Karathanasis and Wilson, 2002) was cut with *Kpn*I to create a 3' overhanging DSB within its *ADE2* marker, leaving its *URA3* marker intact. Each cut plasmid (100 ng) was co-transformed into YW1228 derivatives along

with 10 ng of supercoiled *HIS3*-marked pRS413 (Brachmann *et al.*, 1998), as previously described (Palmbos *et al.*, 2008). c-NHEJ efficiency is reported as the ratio of Ura^+ (pRS316) or $Ura^{+/}Ade^+$ colonies (pK1827) to His^+ colonies.

Suicide deletion assay. The suicide deletion assay for monitoring NHEJ was previously described (Karathanasis and Wilson, 2002; Palmbos *et al.*, 2005; Wilson, 2002). Here all strains (**Table 4-3**) bore the *ade2*::SD2-::*STE3-MET15* allele for which HR is not possible. Cells were pre-grown either to stationary phase for 2 days in SD medium, our standard protocol (Karathanasis and Wilson, 2002), or to log phase overnight in YPA-Glycerol. NHEJ was scored as colony counts on galactose plates divided by counts on parallel glucose control plates. The nature of NHEJ events in Ade^+ colonies was determined by mating to YW2083 carrying the I-*Sce*I expression plasmid pTW334 (Karathanasis and Wilson, 2002). If precise NHEJ had created an intact I-*Sce*I site in the suicide deletion strain, it is recleaved and undergoes HR with the YW2083 *ade2*-M7 allele to give Ade^-/red diploids. Ade^- suicide deletion colonies were assessed by first ensuring an isolate was Met^-, and thus had deleted *STE3-MET15*, and then attempting to amplify a 1.2 kb PCR product symmetrically flanking the expected DSB junction. No product indicated a large deletion. Inferred imprecise NHEJ alleles were finally sequenced.

Chromosomal DSB dnduction. We previously described the *GAL1*-cs and *ILV1*-cs strains for creating single-site chromosomal DSBs (Wu *et al.*, 2008). In them, HO is expressed from the native chromosomal *GAL1* promoter so that the strain is rendered Gal^-. HO cut sites are present in either the same *GAL1* promoter that expresses HO (*GAL1*-cs) or in the native *ILV1* promoter (*ILV1*-cs). DSB induction and α-factor G1 arrest were as described (Wu *et al.*, 2008).

Epitope tagging and ChIP. *GAL1*-cs strains bore a 13Myc tag on the C-terminus of the native chromosomal *DNL4, CDC9, LIF1* and *YKU80* genes as previously described (Wu *et al.*, 2008). ChIP assays and parallel monitoring of DSB formation and repair were performed as described (Wu *et al.*, 2008) except that quantitative PCR rather than competitive PCR was used to monitor Dnl4 binding to the DSB site, with fold enrichment determined from the Ct difference between *GAL1* and the *ACT1* control gene.

Immunoprecipitation and adenylation assays. *GAL1*-cs strains were grown overnight in YPA-Dextrose. Lysates were prepared from 1 g wet cell pellet in lysis buffer (10 mM Tris pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$, 2 mM $CaCl_2$, 10% glycerol, and Roche complete mini tablet) using zirconia bead lysis followed by centrifugation at 15,000 x g. Dnl4 was immunoprecipitated by overnight incubation with 80 µl anti-Myc conjugated agarose beads (Sigma). Samples were washed 3 times with lysis buffer, divided into two aliquots, and pre-incubated with 1 ml AMP buffer (60 mM Tris pH 8.0, 10 mM $MgCl_2$, 2mM DTT, and Roche complete mini tablet) with or without 0.5 mM sodium pyrophosphate for 15 min at room temperature. Beads were washed with 1 ml AMP buffer before addition of 15µCi α-^{32}P ATP to the residual volume for 30 min at room temperature. Beads were washed twice before boiling and electrophoresis on a 7.5% SDS-PAGE gel and transfer to nitrocellulose. The same membrane was used to obtain a phosphorimager screen exposure followed by Western blotting using anti-Myc antibody with detection by a LiCor Odyssey scanner. Results are expressed as the ratio of phosphorimager to Western blot signal, setting the wild-type Dnl4 sample with pyrophosphate treatment to 100%.

Resection monitoring. Genomic DNA was extracted after galactose induction in the *ILV1*-cs system, digested with *Eco*RI and *Nco*I, and subjected to Southern blotting as previously

described (Wu *et al.*, 2008). Results are expressed as a ratio of the intensity of the intact (2.6 kb) or HO-cut (1.1 kb) *ILV1* fragment to that of the *APN1* control fragment (3.5 kb), normalized to the ratio at either 0 or 60 minutes of induction, as indicated.

Recombination assay. *ILV1*-cs yeast were transformed with pRS416 containing a 960bp homologous *ILV1* donor fragment. Cells were plated to glucose plates after varying times of DSB induction in galactose to determine cell survival relative to untreated cells.

Next Generation Sequencing. *ILV1*-cs yeast were pre-grown overnight in YPA-Glycerol. Genomic DNA was harvested before and 24 hours after galactose induction and subjected to PCR using primers flanking the *ILV1* HO DSB site that additionally bore unique 6-base sequence indexes for each input sample and a 5' extension matching Illumina paired-end sequencing adapters (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxxxxAGGGCAAAAAGAAAAAGCGCAG and 5'-CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTxxxxxxGTTTTATCAAGGAAGGTGACA, where "xxxxxx" was replaced with specific index sequences). After 15 cycles of amplification, products were purified and subjected to an additional 16 PCR cycles using Illumina paired-end second-round primers. Products from all samples were pooled, 50% of a PhiX control library was added to facilitate cluster calling, and the mixture subjected to Illumina HiSeq 100bp paired-end sequencing at the University of Michigan DNA Sequencing Core. Only forward reads crossed the *ILV1* HO cut site and were used for further analysis. Reads matching PhiX were removed using Bowtie (Langmead *et al.*, 2009). Custom Perl scripts (**File 4-2**) assigned the remaining reads to the input samples based on the index and counted the number

that corresponded to the intact *ILV1* HO cut site or any one of the possible blunt, microhomology, or insertion-mediated joints supported by the input DSB ends, allowing up to two inserted bases at the junction and one mismatched read base. Reads corresponding to more than one joint were called ambiguous. Joint counts were input into DESeq (Anders and Huber, 2010) as is, or expressed as a fraction of the total *ILV1* read counts for graphical comparisons.

Dnl4 mass spectrometry. pTW644 (a kind gift Patrick Sung) expresses Dnl4 and His6-tagged Lif1 from divergent *GAL1/10* promoters. Dnl4-K466A was introduced into this plasmid by PCR amplification and ligation into *BamH*I/*Sal*I sites and confirmed by sequencing. Plasmids were transformed into YW1230 for suicide deletion analysis as described above and into protease-deficient YW2189 for protein purification. Yeast were grown to log phase in 1000 ml of YPA-Glycerol and expression induced by addition of 2% galactose for 4 hours. 5g of cells were harvested, resuspended in buffer A (Palmbos *et al.*, 2008), and lysed using zirconium beads (Biospec). The cleared lysate was incubated with 0.3 ml of nickel-NTA agarose beads (Qiagen) in the presence of 10 mM imidazole for 1h at 4 ^{0}C. The beads were washed three times with buffer A containing 20 mM imidazole and boiled in 1X SDS loading buffer. Following 10% SDS-PAGE and Novex colloidal blue staining (Invitrogen), Dnl4 bands were excised and subjected to trypsin digestion. Resulting peptides were resolved on a nano-capillary reverse phase column (Picofrit column, New Objective) using a 1% acetic acid/acetonitrile gradient, introduced into a linear ion-trap mass spectrometer (LTQ Orbitrap XL, ThermoFisher), and searched against the *S. cerevisiae* protein database using X!Tandem/TPP (Yates *et al.*, 1995). Oxidation of methionine (+15.9949 Da), carbamidomethylation of cysteines (+57.0214 Da), and AMP-lysine (+329.0525 Da) were allowed modifications. Precursor and fragment mass tolerance

were 50 ppm and 0.8 Da, respectively. Matches to Dnl4 with a ProteinProphet probability of >0.9 (fdr <2%) were used to determine the adenylation state of K282-containing peptides.

Results

Dnl4 catalytic mutants are stable and bind Lif1 normally. We introduced mutations into the *DNL4* chromosomal locus to explore the DSB repair behavior of catalytically defective Dnl4 (**Figure 4-1A**). Mutations in ligase Motif 1 (K282R, D284A) and Motif V (K466A) affect universally conserved residues predicted to be involved in ligation catalysis based on comparisons to structure-function studies of viral DNA ligases(Sriskanda and Shuman, 1998, 2002; Nair *et al.*, 2007) and the crystal structure of human DNA ligase I (Pascal *et al.*, 2004) (**Figure 4-1B,C**). These residues were chosen on the hypothesis that they might differentially affect the three steps of ligation. K282R alters the lysine predicted to form the covalent AMP adduct and should be an obligatory step 1 mutant, a behavior validated by the complete inability of immunoprecipitated Dnl4-K282R to become adenylated *in vitro* (**Figure 4-2A**). By analogy to viral ligase mutants (Sriskanda and Shuman, 1998, 2002) it was possible that D284A and K466A might be step 2 and step 3 mutants, respectively. However, each of these mutant proteins showed severely reduced *in vitro* adenylation following pre-incubation with excess pyrophosphate to remove pre-existent AMP adducts (**Figure 4-2A**). Because adenylation and pyrophosphorolysis are reverse reactions, these data indicate that Dnl4-D284A and Dnl4-K466A are also severely defective in ligation step 1.

We next examined the dependence of ligase stability and the Dnl4-Lif1 interaction on catalytic function. Western blots of whole cell extracts demonstrated that Dnl4-K282R was expressed at moderately reduced levels relative to wild-type Dnl4 (**Figure 4-2B**), suggesting a partial instability of this mutant form. In contrast, Dnl4-D284A and Dnl4-K466A were expressed normally (**Figure 4-2B**). These expression results mirrored the apparent strength of the Dnl4-

Lif1 interaction as revealed by two-hybrid analysis (**Figure 4-2C**), but in fact all three catalytic mutants showed a good ability to co-immunoprecipitate Lif1 (**Figure 4-2D**). We interpret that the reduced Dnl4-K282R-Lif1 two-hybrid signal is most likely due to reduced Dnl4 expression rather than a defect in Lif1 binding, and thus that Dnl4 catalytic function is not required for Lif1 binding.

For comparison in repair studies below, we also created truncation mutant L750*, which lacks the core of the Lif1 interaction region (Dore *et al.*, 2006) (**Figure 4-1A**). As expected, Dnl4-L750* was unable to bind Lif1 (**Figure 4-2C**).

Dnl4 catalytic mutants are strongly defective in c-NHEJ. To determine the c-NHEJ phenotype of the catalytic Dnl4 mutants, corresponding strains were transformed by linearized plasmids bearing 5' (*Nco*I) or 3' (*Kpn*I) overhanging DSBs, which is known to depend on Dnl4-mediated simple-religation c-NHEJ (Daley and Wilson, 2005; Wilson *et al.*, 1997; Karathanasis and Wilson, 2002). Consistent with their severe adenylation defect (**Figure 4-2A**), all three mutants were impaired in plasmid repair to nearly the same degree as a *dnl4*Δ mutant (**Figure 4-2E,F**), confirming the expected finding that catalysis by Dnl4 is required for c-NHEJ.

Catalytically defective Dnl4 binds to, and is removed from, chromosomal DSBs. To determine whether catalytically defective Dnl4 mutants are recruited to DSBs *in vivo*, we employed a previously described system, *GAL1*-cs, in which a single DSB is introduced into the chromosomal *GAL1* promoter by transient induction of HO endonuclease (Wu *et al.*, 2008). In this system, the accumulation of 13Myc-tagged Dnl4 at the DSB is measured by chromatin immunoprecipitation (ChIP) and any associated NHEJ repair is monitored by a parallel quantitative PCR assay. Unlike wild-type Dnl4, which showed an increase in intact HO cut sites

from 24% at one hour to 72% at four hours after DSB induction, Dnl4-K282R, -D284A and -K466A were all unable to repair the cleaved allele (**Figure 4-3A**), consistent with results above. ChIP demonstrated that the NHEJ defect of the catalytic mutants was not a result of failed recruitment to the DSB. The three mutant proteins each accumulated at the DSB at one and two hours and in fact achieved even higher levels of binding than wild-type (**Figure 4-3B**), presumably because DSBs in these strains were not being repaired.

Further ChIP analysis was used to explore whether an intact c-NHEJ complex, or just Dnl4, was accumulating at DSBs. Similar to Dnl4, Lif1 was hyper-recruited to DSBs at one and two hours in *dnl4*-K466A as compared to the wild-type strain, in marked contrast to the greatly impaired accumulation of Lif1 observed in our system and others in a *dnl4*Δ strain (**Figure 4-3C**) (Zhang *et al.*, 2007). Yku80 showed a similarly enhanced ChIP signal in the *dnl4*-K466A mutant, while signal was slightly reduced at two hours in the *dnl4*Δ strain in a manner that likely reflects the reported stabilization of Ku by the ligase complex (**Figure 4-3D**) (Zhang *et al.*, 2007). Thus, we conclude that typical, stable, but non-productive c-NHEJ complexes are assembled at DSBs in strains bearing catalytically defective Dnl4.

ChIP readouts also provided key information regarding the dissociation of Dnl4. We previously observed that Dnl4 and other c-NHEJ proteins begin rapidly disappearing from DSBs after two-hours in an MRX-dependent fashion (Wu *et al.*, 2008; Palmbos *et al.*, 2008). **Figure 4-3B-D** demonstrate that this disappearance is independent of successful c-NHEJ because non-productive c-NHEJ complexes were removed from DSBs with the same time course as when Dnl4 had successfully catalyzed DSB repair in many cells.

NHEJ, but not bound Dnl4, restricts DSB resection. We previously showed that disappearance of c-NHEJ proteins from DSBs correlates with the onset of 5' resection in preparation for HR (Wu *et al.*, 2008). Others have suggested that the presence of DNA ligase IV and the c-NHEJ complex at a DSB is inhibitory to resection (Zhang *et al.*, 2007; Clerici *et al.*, 2008). However, the gene deletion mutations used in such studies both prevent productive c-NHEJ and disrupt c-NHEJ assembly. This distinction is important because experiments that rely on continued expression of HO allow for futile cycles in which DSBs are created, repaired and recreated. Observed effects of *dnl4*Δ mutants might result from interruption of this repair cycle rather than direct inhibition of resection by Dnl4. Dnl4 catalytic mutants provided a means of distinguishing these phenomena.

For monitoring resection we used a previously described system, *ILV1*-cs, in which a single HO DSB is introduced into the chromosomal *ILV1* promoter (Wu *et al.*, 2008). Here a one-hour induction period was followed by the addition of glucose to turn off HO expression, so that either NHEJ or 5' resection might ensue. These two downstream events were monitored and distinguished by Southern blotting. Surprisingly, glucose addition proved to be essential for DSB resection to occur (compare **Figures 4-4A** and **4-4B**). We attribute this to the fact that our strains are *gal1* due to fusion of HO to the native *GAL1* promoter and therefore cannot utilize galactose as a carbon source. Consistently, reintroduction of the *GAL1* gene on a plasmid restored resection without the addition of glucose (**Figure 4-8**), typical of an extensive yeast literature at various DSB sites. These results demonstrate that the metabolic state of a cell substantially influences resection efficiency.

Figure 4-4B-D shows *ILV1*-cs results with asynchronous cultures of yeast expressing wild-type Dnl4, no Dnl4 (*dnl4*Δ) or Dnl4-K466A. Once again, substantial NHEJ occurred upon glucose addition to the wild-type strain that was absent with both *dnl4* mutants (**Figure 4-4B,C**). This repair contributed to the disappearance of the HO-cut alleles (**Figure 4-4B,D**). However, loss of the HO-cut band also reflects competing 5' resection, and indeed it also disappeared from the *dnl4* mutant strains. The rate of loss of the HO-cut band was equivalent regardless of whether productive c-NHEJ was occurring. This result indicates that resection did occur more robustly in the *dnl4* mutant strains but that this difference was accounted for by the DSBs being repaired by c-NHEJ in wild-type. Moreover, the rate of HO-cut band loss, and thus resection, was identical when comparing *dnl4*Δ and K466A strains (**Figure 4-4D**). Similar to previous results (Ira *et al.*, 2004; Zhang *et al.*, 2007; Clerici *et al.*, 2008), G1 arrest inhibited resection, although with *ILV1*-cs system this effect was incomplete (**Figure 4-4E,F**). In G1, we did observe a mildly accelerated loss of the HO-cut band in *dnl4* mutant strains as compared to wild-type, similar to prior observations (Zhang *et al.*, 2007). However, the resection rate was again no different between *dnl4*Δ and K466A (**Figure 4-4F**). Similarly, there was no difference in HR-mediated repair when a homologous donor allele was introduced into the various strains (**Figure 4-9**). We conclude that the increased rate of resection in *dnl4* mutants can be attributed to the loss of competing NHEJ and is not correlated with the presence of Dnl4 at a DSB.

Dnl4 promotes imprecise NHEJ independently of its catalytic function. We further examined the NHEJ function of the catalytic mutants at two nearby chromosomal DSBs using the well-described suicide deletion assay (Karathanasis and Wilson, 2002; Palmbos *et al.*, 2005; Wilson, 2002). Suicide deletion involves galactose-induced formation of two I-*Sce*I DSBs

flanking an internal cassette that are joined to recreate the *ADE2* gene (**Figure 4-5A**). With this assay, *dnl4*-L750* was as deficient as *dnl4*Δ yeast (**Table 4-1**), confirming the known Dnl4 dependence of the assay (Karathanasis and Wilson, 2002; Wilson, 2002). In marked contrast, we observed measureable rates of Ade^+ colony formation with all of the Dnl4 catalytic mutants above that observed with *dnl4*Δ and L750* (**Table 4-1** and **Figure 4-5B**). Critically, Ade^+ colonies can only arise when the split halves of *ADE2* are restored by NHEJ. To verify this, we re-expressed I-*Sce*I in them and observed that, as expected, many repair joints could be re-cleaved by I-*Sce*I and so occurred by simple religation (**Table 4-1**). These results indicate that residual end joining capacity exists that depends on Dnl4 protein but not its catalytic activity.

Sequencing the *ADE2* alleles from a number of I-*Sce*I-resistant Ade^+ colonies revealed a similar and typical collection of in-frame but imprecise NHEJ junctions, including microhomologies, insertions and deletions, for both wild-type Dnl4 and the catalytic mutants (**Figure 4-10A**). Indeed, the absolute frequency of imprecise NHEJ was not substantially different for the catalytic point mutants as compared to wild-type Dnl4, in marked contrast to *dnl4*Δ and L750* yeast that showed a complete loss of imprecise junctions (**Table 4-1**). Thus, the joining pathway dependent on the Dnl4 protein, but not its activity, was substantially more mutagenic than that being catalyzed by Dnl4.

Ade^- suicide deletion colonies are much rarer than Ade^+ colonies and cannot be selected for when plating, preventing statistically meaningful frequency comparisons. However, we performed allelic analysis on the Ade^-colonies we did recover (**Table 4-2**), first by scoring for status of the *STE3-MET15* marker contained within the cassette that is lost upon suicide deletion (**Figure 4-5A**). We then attempted to amplify the DSB junctions from Ade^- Met^- colonies to

determine whether they corresponded to imprecise NHEJ or large deletions of the *ADE2* locus. Strikingly, *dnl4*Δ colonies never arose due to imprecise NHEJ but instead showed high frequencies of large deletions (**Table 4-2**). In contrast, the catalytic mutants showed imprecise NHEJ in a pattern that mirrored the Ade^+ suicide deletion results (**Table 4-2**, **Figure 4-10**). We conclude that imprecise NHEJ in the suicide deletion assay depends on Dnl4, but not its catalytic activity.

The different result pattern when comparing the suicide deletion (**Figure 4-5**) and plasmid recircularization (**Figure 4-2E,F**) assays was striking, especially since the I-*Sce*I and *Kpn*I 3' DSBs are in the same position in *ADE2*. Importantly, our standard suicide deletion protocol (Karathanasis and Wilson, 2002) entails pre-growth of cells to post-diauxic phase for two days, unlike the log-phase cells used in plasmid recircularization assays. Indeed, the residual suicide deletion repair with Dnl4 catalytic mutants was almost entirely suppressed when the experiment was repeated with log-phase cells (**Figure 4-5C**), demonstrating that, like resection, the mutagenic end joining pathway dependent on Dnl4 protein is strongly influenced by the metabolic state of the cell.

Residual NHEJ in the above suicide deletion experiments cannot have been catalyzed by Dnl4-K282R as it cannot be adenylated (see **Figure 4-2A**). We further assessed the *in vivo* adenylation state of Dnl4-K466A. Overexpressed Dnl4 proteins were excised from a gel (**Figure 4-11A**) and subjected to LC-MS/MS. Importantly, enhanced suicide-deletion NHEJ was still seen with overexpressed Dnl4-K466A (**Figure 4-11B**). Twelve adenylated peptides were recovered with wild-type Dnl4 out of 163 K282-containing peptides (**Figure 4-11C**). This fraction is smaller than expected and likely reveals instability of the adenylation intermediate

during fragmentation. Nevertheless, no adenylated peptides were recovered out of 128 K282-containing peptides from Dnl4-K466A, consistent with the *in vitro* results that it is deficient in ligation step 1.

The above results with catalytically defective Dnl4 were reminiscent of prior observations that yeast lacking the Dnl4 accessory protein Nej1 display substantial residual suicide deletion NHEJ, unlike Ku, Lif1 and MRX mutants that phenocopy the near complete defect seen with *dnl4*Δ (Wilson, 2002). However, the residual NHEJ observed with *nej1*Δ and catalytically defective Dnl4 do not reflect the same phenomenon since *nej1 dnl4* double mutants showed a synergistic decrease in suicide deletion survival (**Figure 4-5D**). Thus, Nej1 is required for the residual NHEJ observed with catalytically defective Dnl4. In turn, Dnl4 likely catalyzes the residual NHEJ observed with *nej1*Δ, demonstrating that c-NHEJ is differentially dependent on Nej1 and Lif1 in at least this assay.

A final possibility was that the Dnl4 catalytic mutants might act as dominant negatives if they could preclude the action of competent Dnl4 molecules. To test this, we transformed *dnl4* mutant strains with a *DNL4* plasmid. The same degree of complementation was observed with the three catalytic point mutants as with the *dnl4*Δ strain (**Figure 4-5B**), indicating that point mutant and wild-type Dnl4 had equal access to the DSB.

Dnl4-K466A confers a different imprecise joining profile than wild-type. We next pursued high-throughput next-generation sequencing of NHEJ occurring at HO-induced DSBs in the *ILV1*-cs assay (i) to determine whether the findings regarding imprecise suicide deletion NHEJ could be confirmed using a single-DSB assay, and (ii) to dramatically improve the capacity with which imprecise joints could be scored to allow a more accurate comparison of the

joining profiles of different strains. Here we did not add glucose following initial DSB induction in order to suppress 5' resection (see **Figure 4-4A**) and to allow for continuous (re)cleavage of intact HO cut sites and thereby enrich for imprecise joining. Genomic DNA was collected from wild-type, *dnl4*-K466A, and *dnl4*Δ strains at 0-hour and 24-hour time points and PCR products flanking the HO DSB site were subjected to Illumina HiSeq sequencing.

As expected, all strains showed ≥98% intact cut sites from among all recovered reads before HO induction (**File 4-1**). At 24 hours, wild-type and K466A strains each showed an average across two independent replicates of 40% imprecise joints, consistent with the design of the experiment (**Figure 4-6A**). In marked contrast, the *dnl4*Δ strain yielded 97% precise joints and only 3% imprecise joints at 24 hours, a pattern more like the 0-hour time point than the other strains. The recovery of mainly intact cut sites in the most highly NHEJ deficient strain is explained by the presence of uncleaved alleles in dead cells combined with the near absence of *bona fide* repair events in live cells, given that results represent the frequency of sequences recovered from a sample, not their absolute efficiency. The inferred severe impairment of imprecise NHEJ only in the *dnl4*Δ strain closely parallels the findings from the suicide deletion assay.

To compare the imprecise joining profiles of wild-type and Dnl4-K466A, joints were codified to names containing all information needed to uniquely define them (explained in **Figure 4-12**). The DESeq program (Anders and Huber, 2010) was then used to apply the negative binomial distribution to identify joints for which the frequency was significantly different between the two 24-hour replicates of K466A as compared to wild-type (**Figure 4-13**). To help ensure statistical validity, we further restricted our attention to only those joints that had

a frequency >0.1% in any tested strain, for which robust counts had thus been obtained (**File 4-1**). Although wild-type and K466A strains showed a similar overall frequency for these major imprecise joints (**Figure 4-6A**), 15 (25%) and 16 (27%) individual joints showed significantly increased and decreased frequencies, respectively, when comparing K466A to wild-type (**Figure 4-6B**). These results establish that loss of Dnl4 catalytic activity confers differences in the quality of imprecise joints that are all nonetheless dependent on the Dnl4 protein.

Cdc9 is recruited to DSBs coincident with Dnl4. A possible explanation for the above results is that Cdc9/DNA ligase I, the only other DNA ligase in *S. cerevisiae*, catalyzes the residual NHEJ when Dnl4 is present but inactive. It is difficult to perform NHEJ assays in *cdc9* yeast as it is an essential gene, so to begin to assess its potential role in residual NHEJ we performed Cdc9 ChIP. A first experiment revealed the appearance of Cdc9 at the *GAL1*-cs DSB after one hour of galactose induction and its disappearance following glucose addition (**Figure 4-7A,B**). This pattern was very similar to Dnl4 (**Figure 4-3**), although the level of Cdc9 enrichment was lower. To make the signal more sustained we repeated the experiment by pre-growing the cells to post-diauxic phase and did not add glucose at one hour, which was expected to prevent 5' resection and promote residual NHEJ according to results above. DSB induction was delayed in this protocol (**Figure 4-7C**), but Cdc9 enrichment was again observed that now persisted for four hours (**Figure 4-7D**). Given the lower DSB levels, this pattern suggests a more robust Cdc9 enrichment in post-diauxic than log phase cells. Controls lacking the anti-Myc ChIP antibody confirmed the specificity of this result, for which DSB induction in fact lowered the DSB locus signal relative to the *ACT1* control locus, perhaps due to reduced non-specific binding of the DSB fragment resulting from local chromatin alterations upon DSB induction (**Figure 4-**

7D) (Shim *et al.*, 2007). Finally, a pattern emerged in post-diauxic cells wherein Cdc9 signals were lowest in the complete absence of Dnl4 and highest in the presence of Dnl4-K466A (**Figure 4-7D**). This trend was again not predicted by the higher levels of DSB formation in the *dnl4*Δ strain, suggesting a promotion of Cdc9 retention by Dnl4 in post-diauxic cells.

Discussion

The central findings of this work are that (i) catalytically inactive yeast Dnl4/DNA ligase IV can promote imprecise NHEJ when it persists at DSBs, (ii) Cdc9/DNA ligase I is present at DSBs at times consistent with it catalyzing the observed residual NHEJ in place of Dnl4, (iii) these phenomenon and DSB 5' resection are influenced by the metabolic state of the cells, and (iv) 5' resection efficiency is reduced by c-NHEJ activity but not by the presence of Dnl4 protein at a DSB, suggesting an active process that removes NHEJ proteins and initiates 5' resection.

Dnl4 supports mutagenic NHEJ independently of its catalytic activity. We made mutations in universally conserved Dnl4 residues that have been shown to create step-specific catalytic blocks in other ATP-dependent DNA ligases (Sriskanda and Shuman, 1998, 2002). Dnl4-K282R could not auto-adenylate, as expected, but we also observed severely impaired capacity of catalytic step 1 *in vitro* with Dnl4-D284A and Dnl4-K466A (**Figure 4-2A**) and commensurate undetectable auto-adenylation *in vivo* with Dnl4-K466A (**Figure 4-11**). This pattern of results emphasizes the integrated nature of the DNA ligase active site. Consistently, Dnl4-K282R, -D284A and -K466A mutants were all severely deficient in c-NHEJ *in vivo* in several assays (**Figures 4-2**). No mutant acted as a dominant negative (**Figure 4-5B**), despite the fact that all bound Lif1 (**Figure 4-2C,D**) and accumulated even more than wild-type at DSBs (**Figure 4-3**). Importantly, DSB recruitment must be distinguished from fully productive binding to DNA termini, which likely requires ligase auto-adenylation (Pascal *et al.*, 2004; Sriskanda and Shuman, 1998).

Given the expected severe effect of Dnl4 catalytic mutation on ligase activity and c-NHEJ, it was striking that these same mutants supported residual NHEJ activity in the suicide

deletion and *ILV1*-cs assays well above that observed in strains lacking Dnl4 (**Figures 4-5** and **4-6**). We considered a number of properties of the suicide deletion assay to account for this residual repair. The phenomenon did not depend on the use of two DSBs since a difference in residual NHEJ was also observed in the single-DSB *ILV1*-cs next-generation sequencing experiment (**Figure 4-6**). In contrast, increased survival due to residual suicide deletion NHEJ was observed only when cells were pre-grown to post-diauxic/early stationary phase (**Figure 4-5**). With the *ILV1*-cs assay, the detection of Dnl4 protein-dependent residual NHEJ entailed a growth protocol that promoted the recovery of imprecise joints by prolonged HO expression in the context of limited 5' resection due to the inability of the cells to ferment galactose.

Despite the technical differences in execution, the suicide deletion and *ILV1*-cs assays led to the same conclusion that a substantial extent of imprecise NHEJ is preserved with catalytically defective Dnl4 but lost in the absence of Dnl4 protein. Thus, the form of NHEJ supported by Dnl4 protein was more mutagenic than the proper c-NHEJ catalyzed by wild-type Dnl4. This observation is reminiscent of findings that Dnl4 participates in some microhomology mediated end joining (MMEJ) through a poorly understood mechanism (Ma *et al.*, 2003) and of mammalian studies documenting the mutagenic potential of DNA ligase IV-independent NHEJ (Boboila *et al.*, 2012; Chiruvella *et al.*, 2012). It also has implications for the human ligase IV syndrome in which mutations such as R278H, Q280R, and H282L (motif I) and G468E, G469E (motif V) are hypomorphic despite the fact that they severely alter the ATP binding properties of the enzyme (Riballo *et al.*, 2001; Girard *et al.*, 2004). This pattern is very similar to the yeast mutants studied here and suggests the possibility for affected humans that inactive ligase might accumulate at a DSB and support mutagenic NHEJ.

Cdc9 as a candidate imprecise NHEJ ligase. Joint analysis unequivocally established that residual repair observed with Dnl4 catalytic mutants occurred by NHEJ (**Figure 4-6**, **4-10** and **Table 4-1**), demonstrating that NHEJ can be catalyzed by a ligase other than Dnl4. We cannot entirely rule out that residual repair was catalyzed by some Dnl4 mutants, but K282R is incapable of becoming adenylated (**Figure 4-2A**) and we found no evidence to support adenylation and c-NHEJ ligation by K466A (**Figures 4-2** and **4-11**). In contrast, Cdc9 was recruited to DSBs *in vivo* in a time frame consistent with its participation in NHEJ and in a manner that was enhanced by Dnl4 protein (**Figure 4-7**). The difference in the imprecise joint profile obtained with wild-type and Dnl4-K466A (**Figures 4-6B** and **4-13**) provides further indirect support for the participation of a different ligase in the residual imprecise NHEJ. Dnl4-independent NHEJ is not especially surprising since even some precise NHEJ occurs in the complete absence of Dnl4 (Wilson *et al.*, 1997) and can become quite efficient as overhang lengths increase (Daley and Wilson, 2005). Moreover, DNA ligase III, which is missing from yeast, and to a lesser extent DNA ligase I have been shown to catalyze alt-NHEJ in higher eukaryotes (Simsek *et al.*, 2011; Della-Maria *et al.*, 2011). The key finding here is that use of an alternative ligase leading to especially imprecise NHEJ can be stimulated by the presence of Dnl4 protein at a DSB, to the point that one must consider whether many mutagenic NHEJ events are catalyzed by DNA ligase I/III even when DNA ligase IV is present, an idea supported by mammalian studies (Simsek *et al.*, 2011).

The exact mechanism by which Dnl4 protein might promote imprecise NHEJ by Cdc9 is unknown, but numerous observations suggest that it is the c-NHEJ complex that is responsible even if the events are not catalyzed by Dnl4. All tested c-NHEJ proteins accumulate at DSBs in

strains bearing Dnl4 catalytic mutations in a manner that parallels Dnl4 itself (**Figure 4-3**), so all are likely present when residual NHEJ occurs. Further, it is well established for suicide deletion and other assays that loss of Ku, Lif1 and MRX alone leads to a severe loss of all NHEJ, including imprecise events, in the same fashion as *dnl4*Δ (Wilson, 2002). In the one case where we could specifically test a role of another c-NHEJ protein, we indeed found that the residual NHEJ observed in the absence of Dnl4 catalytic activity depends on Nej1 (**Figure 4-5D**). The mode of mutagenic NHEJ we describe here thus appears to be distinct from yeast MMEJ (Ma *et al.*, 2003) or alt-NHEJ more generally, which are inhibited by the c-NHEJ assembly containing Ku (Chiruvella *et al.*, 2013). Instead, Dnl4 protein likely promotes assembly of the c-NHEJ complex at the DSB, as has been proposed (Zhang *et al.*, 2007), for subsequent action not only by Dnl4 but also Cdc9 and likely many other enzymes. This concept is consistent with electron microscopy data that DNA ligase IV can execute end bridging in conjunction with Ku (Grob *et al.*, 2012), as well as mammalian studies suggesting that DNA ligase IV acts non-catalytically at DSBs in the early stage of NHEJ (Cottarel *et al.*, 2013), perhaps by supporting filament formation seen with XRCC4/XLF (Mahaney *et al.*, 2013; Hammel *et al.*, 2010).

Dnl4 indirectly inhibits DSB resection. We monitored the flux of DSBs into the competing NHEJ and HR pathways to determine whether resection is inhibited by Dnl4 as has been suggested (Zhang *et al.*, 2007). The results demonstrate that resection was inhibited by Dnl4 only because it catalyzed NHEJ that removed DSBs from the potential resection pool (**Figure 4-4**). DSBs that failed NHEJ were shunted to 5' resection with equal efficiency whether the failure was created by loss of Dnl4 or by a Dnl4 point mutant that accumulated at the DSB. This behavior is explained by the observation that both wild-type and mutant Dnl4 (**Figure 4-3**),

as well as Cdc9 (**Figure 4-7**), were efficiently removed from persistent DSBs. Although the mechanistic sequence remains to be established, these results support a model in which NHEJ protein removal is an active process initiated after an initial time period of end preservation to permit NHEJ (Chiruvella *et al.*, 2013; Wu *et al.*, 2008; Shibata *et al.*, 2011). The notion that it is an active process is supported by the observation that the utilizable carbon source, and by inference the energy status, had a large influence on the cell's ability to engage in 5' resection (**Figure 4-4**) and ligase removal (**Figure 4-7**).

The combined observations on residual NHEJ activity and resection emphasize the impact that cell state, including both cell cycle stage and metabolic influences, has on DSB repair outcomes. Such ideas are relevant to human DNA repair since most cells in an adult mammal are in the G0 state and under careful metabolic control that is disrupted in cancer (Oermann *et al.*, 2012). Indeed, recent studies demonstrate a strong influence of cell state on mutagenic NHEJ outcomes not unlike those observed here (Bindra *et al.*, 2013). These insights also suggest possible unintended mutagenic consequences of DNA ligase IV inhibitors (Srivastava *et al.*, 2012), agents which might lead to accumulation of catalytically ineffective c-NHEJ complexes at DSBs in a manner similar to the mutants studied here.

Acknowledgments

All the contents in Chapter 4 have been published in 2013 in *PLoS Genet* 9(6): e1003599 doi: 10.1371/journal.pgen.1003599 with the title: *Saccharomyces cerevisiae* DNA ligase IV supports imprecise end joining independently of its catalytic activity. I am the second author in the published research article. My main contributions to this project were study of the imprecise joining profile of catalytic mutant K466A by the next-generation sequencing assay I developed (**Figure 4-6**) and the study of CDC9 recruitment to the induced DSBs (**Figure 4-7**). Results of my conducted experiments are key findings in this study which help elucidate potential mechanism of the observed imprecise joining in DNL4 catalytic mutants.

I would like to thank Thomas E Wilson for his mentorship and help in design of the experiments and analysis of the data, and Kishore Chiruvella (first-author) for his work in testing the studied Dnl4 catalytic mutants for their cellular stability, interaction with Lif1, end-joining capacity, 5' resection and protein recruitment, which constitutes the main body of our findings.

I would also like to thank Venkatesha Basrur for the initial screening and profiling of the Dnl4 catalytic mutants partially shown in **Figure 4-1**, Shanda Birkeland for testing the auto-adenylation of the studied Dnl4 catalytic mutants in **Figure 4-2** and Dongliang Wu for initial help with strain construction and protocol of the ChIP assay.

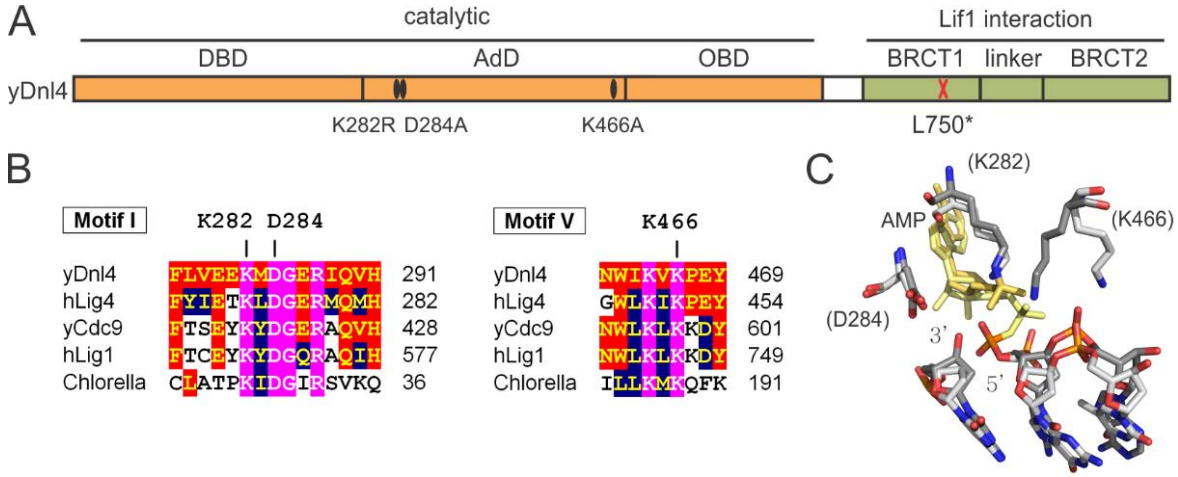


Figure 4-1. Dnl4 mutations under study. (**A**) Location of mutations made in this study relative to the functional domains of *S. cerevisiae* Dnl4 (yDnl4). DBD, DNA binding domain; AdD, adenylation domain; OBD, oligonucleotide binding domain; BRCT, BRCA1 C-terminal repeat; black oval, point mutation; red cross, stop codon. (**B**) Multiple sequence alignments surrounding conserved mutated yDnl4 positions. hLig4, human DNA ligase IV; yCdc9, *S. cerevisiae* DNA ligase I; hLig1, human DNA ligase I; Chlorella, chlorella virus DNA ligase. Magenta, identical among all proteins; red, identical to yDnl4; blue, conserved relative to yDnl4. (**C**) DNA ligase catalytic active site showing a structural alignment of hLig1 bound to a 5'-adenylated DNA nick (PDB 1X9N (Pascal *et al.*, 2004), shaded more lightly) and adenylated Chlorella virus ligase bound to a nick (PDB 2Q2T (Nair *et al.*, 2007), shaded more darkly). Shown are the AMP (yellow), the substrate DNA strand with labeled 3' and 5' nick termini, and the universally conserved residues under study, labeled as the homologous positions in yDnl4. Protein and DNA are shaded by element.

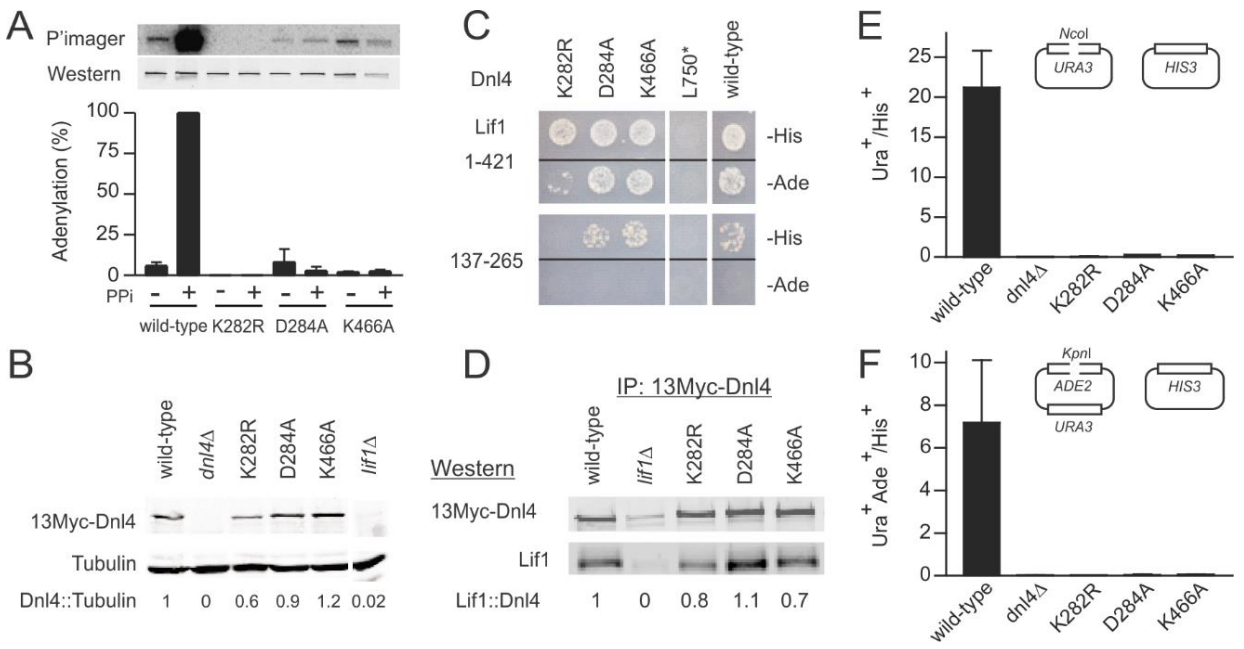


Figure 4-2. Dnl4 catalytic mutations severely impair auto-adenylation and c-NHEJ efficiency. (**A**) Immunoprecipitated 13Myc-Dnl4 was incubated with α-^{32}P ATP to form a radiolabeled enzyme-adenylate complex, with or without sodium pyrophosphate (PPi) pre-treatment to remove pre-existent covalently bound AMP. Phosphorimager (top panel) and immunoblot (bottom panel) exposures of an example blot are shown, as well as normalized results over three independent experiments (bottom panel). (**B**) Immunoblot of whole cell extracts from yeast expressing 13Myc-tagged Dnl4 using anti-Myc and anti-tubulin antibodies (Santa Cruz Biotechnology and Thermo Scientific, respectively). The normalized Dnl4 to tubulin signal ratio is shown. (**C**) Dnl4-Lif1 yeast two-hybrid analysis. Haploid strains expressing the full-length Dnl4 constructs (baits) were mated with strains expressing Lif1 1-421 or 137-265 (preys). The Dnl4-Lif1 interaction was scored by spotting to plates lacking either histidine (-His) or adenine (-Ade). (**D**) 13Myc-Dnl4 protein was immunoprecipitated from yeast in the same manner as (**A**) and immunoblotted for co-immunoprecipitated native Lif1 (antibody a kind gift Alan Tomkinson). Numbers are the Lif1/Dnl4 ratio normalized to wild-type Dnl4, after correcting for the background band co-migrating with Lif1. (**E**) and (**F**) Plasmid recircularization assay with 5' and 3'overhangs, respectively. Results are the mean ± standard deviation of at least two independent experiments.

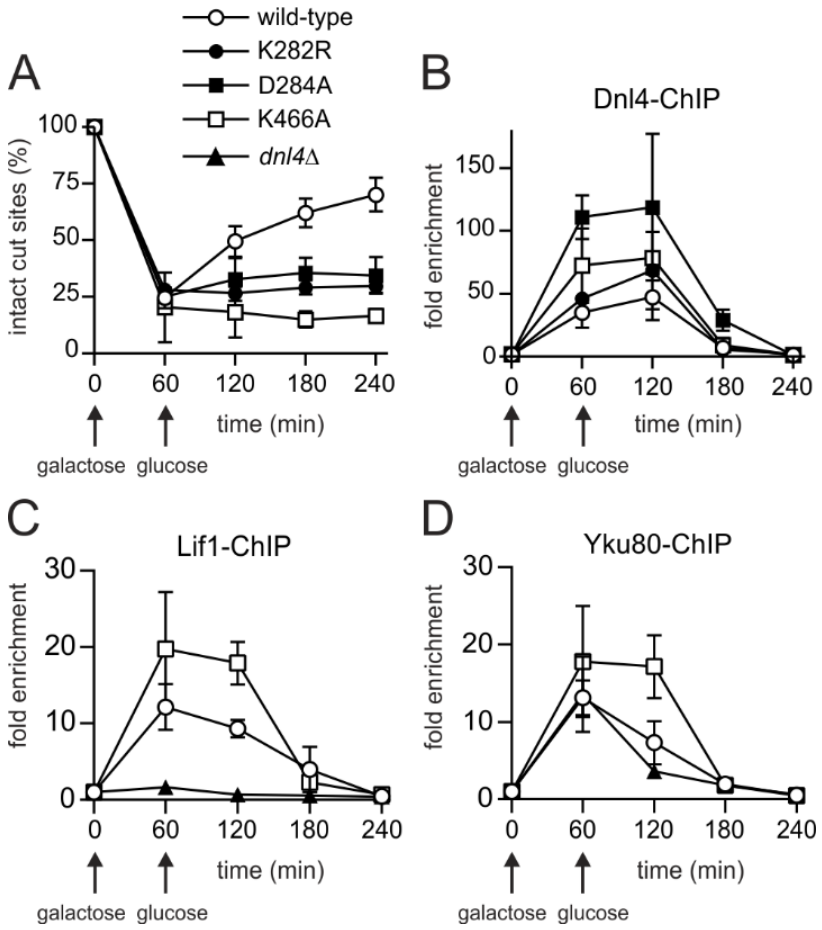


Figure 4-3. Extensive recruitment of catalytically inactive Dnl4 and associated c-NHEJ factors to a chromosomal DSB. Yeast strains bearing the indicated Dnl4 mutations and the *GAL1*-cs allele were grown in galactose medium for 60 min to induce HO expression. Cells were then transferred to glucose to allow repair by NHEJ. (**A**) The fraction of intact *GAL1*-cs HO cut sites showing the extent of DSB formation and repair over time, determined by flanking PCR. (**B**) The corresponding enrichment of 13Myc-tagged Dnl4 at the *GAL1*-cs DSB relative to the *ACT1* control locus as determined by ChIP from the same samples as (**A**). (**C**) Enrichment of 13Myc-tagged Lif1 and (**D**) 13Myc-tagged Yku80 at the *GAL1*-cs DSB. Results are the mean ± standard deviation of at least two independent experiments.

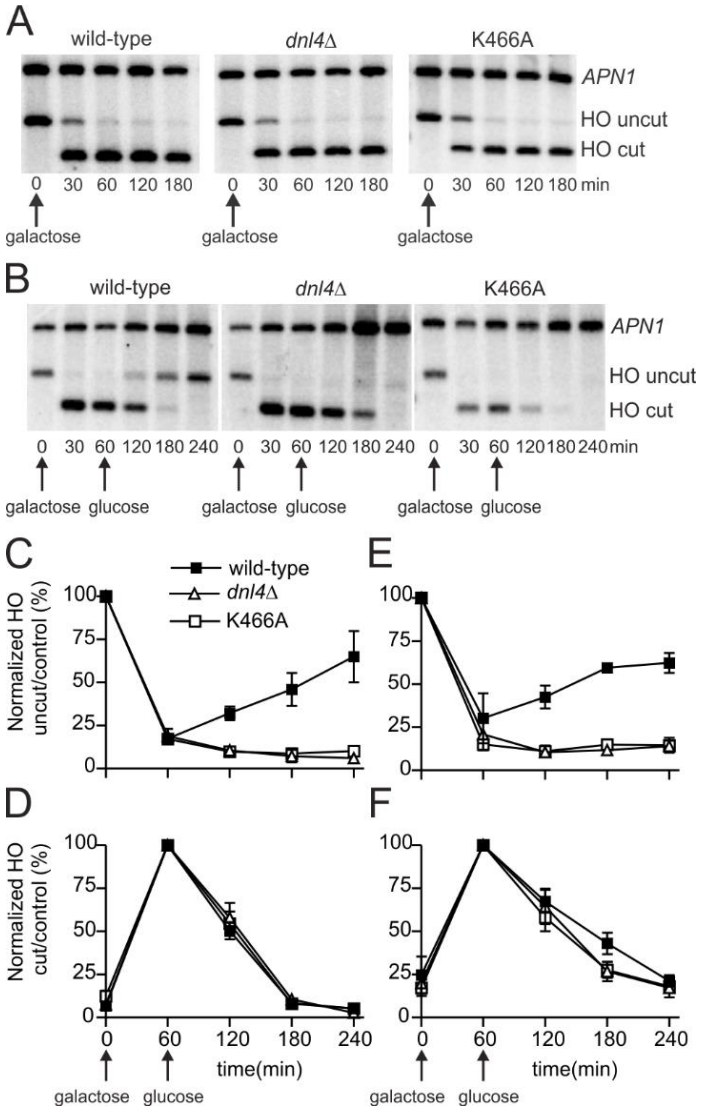


Figure 4-4. Catalytically inactive Dnl4 protein does not impede DSB resection. DSBs were induced in yeast strains bearing the *ILV1*-cs allele similarly to **Figure 4-3** and monitored by Southern blotting. (**A**) Blot showing that DSB resection is not observed without the addition of glucose at 60 min. (**B**) Example blot showing formation of the HO-cut band and its subsequent disappearance by a combination of NHEJ and DSB resection. (**C**) and (**E**) The ratio of the HO-uncut band to the *APN1* control was normalized to the ratio at time 0 to allow monitoring of DSB formation and repair by NHEJ. (**D**) and (**F**) The ratio of the HO-cut band to the *APN1* control was normalized to the ratio at 60 min when DSB formation was maximal. Disappearance of the HO-cut band at subsequent times results from NHEJ (wild-type strain only) and/or DSB resection. (**C**) and (**D**) show results from asynchronous cells while (**E**) and (**F**) show results from cells arrested in G1 with α-factor. Results are the mean ± standard deviation of five independent experiments.

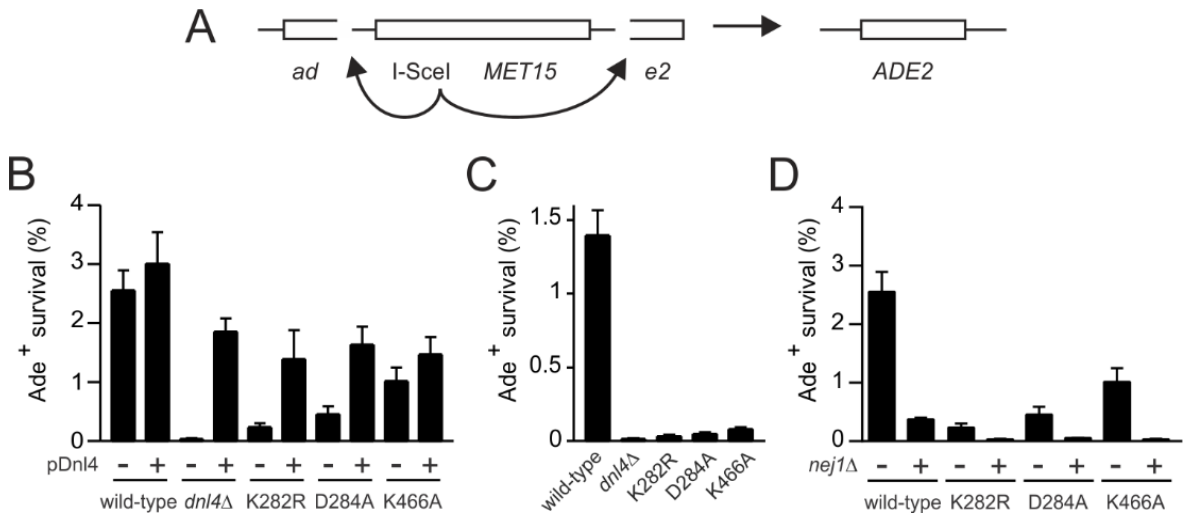


Figure 4-5. Residual NHEJ in the chromosomal suicide deletion assay with catalytically defective Dnl4. (**A**) Diagram of the suicide deletion chromosomal assay used to determine NHEJ efficiency in panels (**B**) to **(D)**. (**B**) Cells were pre-grown to stationary phase in synthetic defined medium prior to plating to galactose. When indicated, the chromosomal *dnl4* allele was complemented with a plasmid bearing wild-type *DNL4* (p*DNL4*). (**C**) Cells were pre-grown to log phase in YPA-Glycerol. (**D**) Epistasis analysis of Dnl4 catalytic mutants with *nej1*Δ. Results are the mean ± standard deviation of at least three independent experiments.

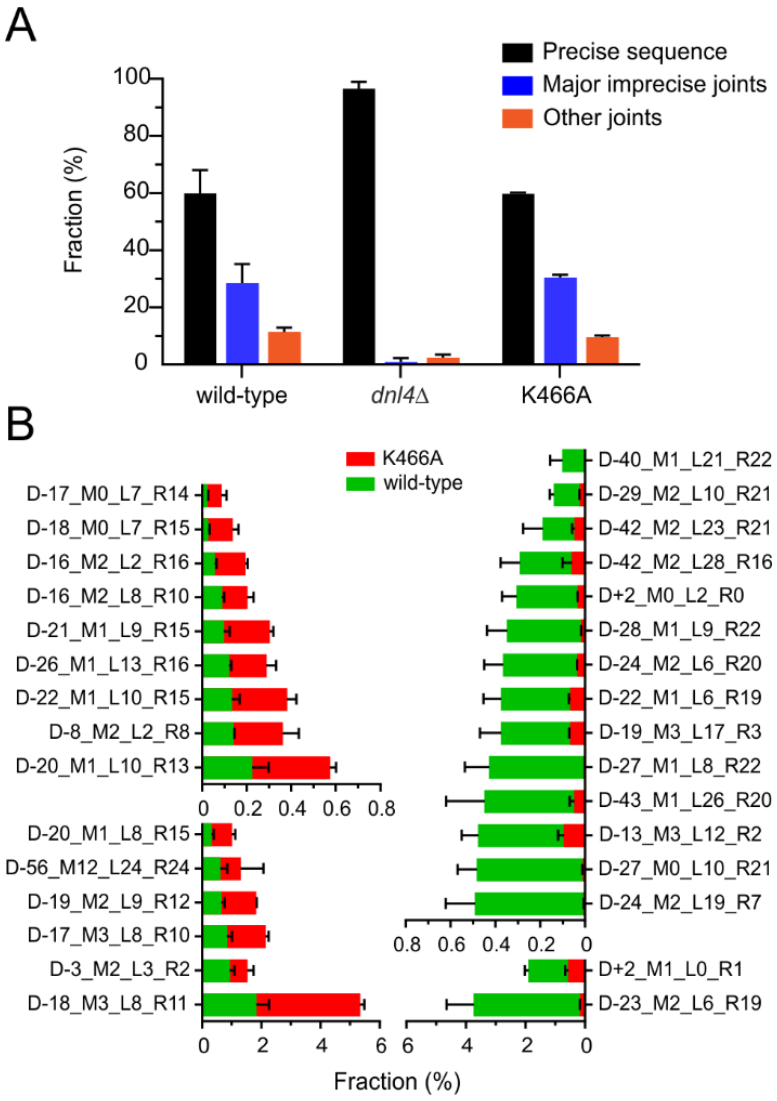


Figure 4-6. End joining profiles after extensive HO recleavage depend on Dnl4 status. (**A**) Fractions of different *ILV1*-cs joint categories after 24 hours of HO induction, showing preserved imprecise joining with K466A. "Precise sequence" is the same as the input allele prior to DSB formation, "major imprecise joints" had a frequency ≥0.1% in any tested strain, and "other joints" are the remainder. (**B**) Individual major imprecise joints that showed significantly different frequencies between wild-type and K466A. The left half shows joints that had increased frequency in K466A as compared to wild-type. The right half shows joints that had a decreased frequency. Red bars, joint fraction in K466A; green bars, joint fraction in wild-type. See **Figure 4-12** for a description of the joint identifiers and **File 4-1** for their sequences.

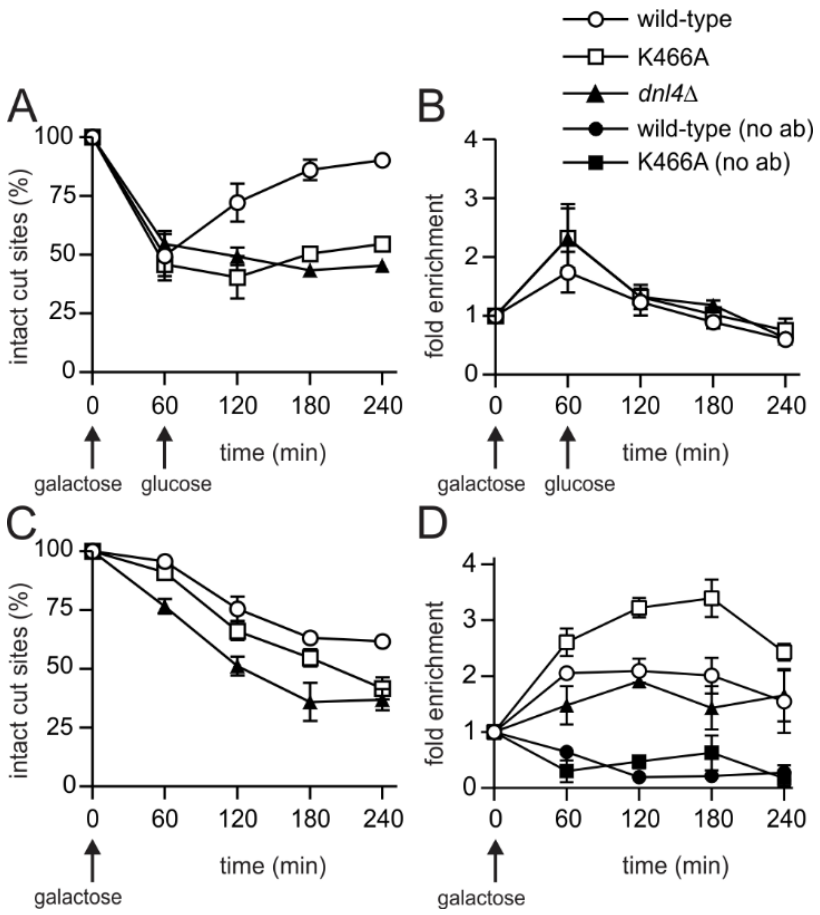


Figure 4-7. Cdc9 recruitment to a chromosomal DSB during the NHEJ repair phase. Yeast strains bearing 13Myc-tagged Cdc9 with the indicated Dnl4 mutations and the *GAL1*-cs allele were pre-grown to log phase in YPA-Glycerol (**A, B**) or post-diauxic phase (**C, D**) and treated with 2% galactose to induce HO expression and DSB formation. (**A**) and (**C**) show the fraction of intact *GAL1*-cs HO cut sites and thus the extent of DSB formation and repair over time as determined by flanking PCR. (**B**) and (**D**) show the corresponding enrichment of 13Myc-Cdc9 at the *GAL1*-cs DSB as determined by ChIP from the same samples as (**A**) and (**C**), respectively. 'No ab' indicates background signal obtained without antibody addition during immunoprecipitation. Results are the mean ± standard deviation of two independent experiments.

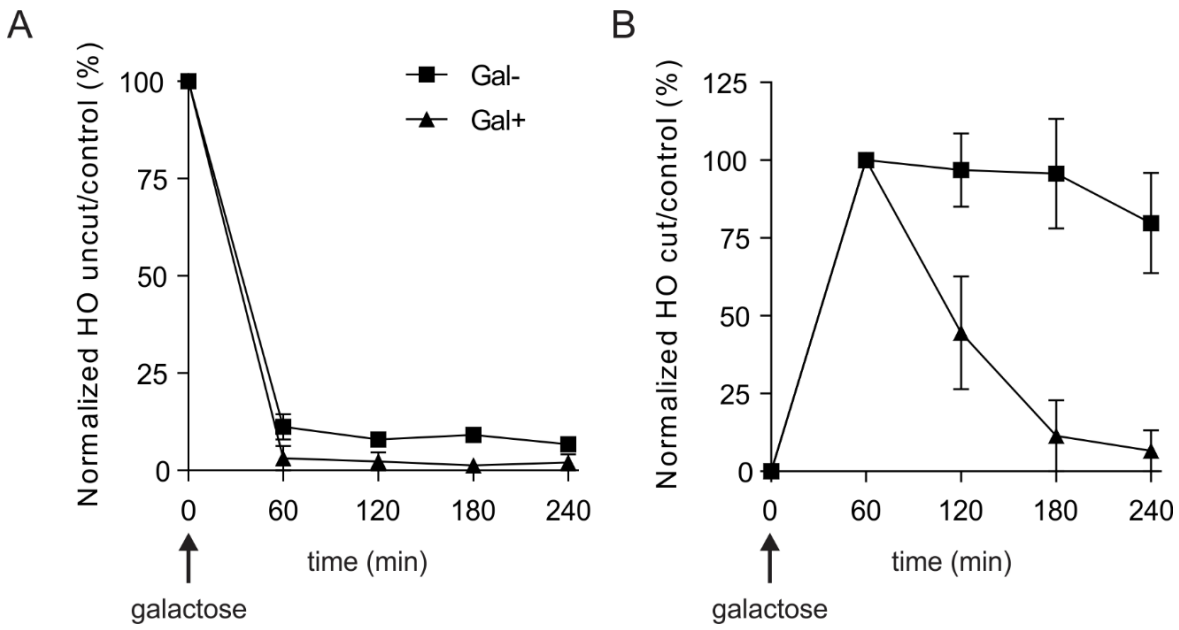


Figure 4-8. DSB resection is inefficient in the absence of a fermentable carbon source. Analysis of DSB resection at the *ILV1*-cs allele was monitored by Southern blotting in *gal1* yeast complemented with vector (Gal-) or p*GAL1* (Gal+) without the addition of glucose at 60 min, similar to **Figure 4-4A**. *gal1* yeast cannot metabolize the galactose added to induce HO expression and DSB formation, but glycerol was present as a carbon and energy source throughout the experiment. (**A**) The ratio of the HO-uncut band to the *APN1* control was normalized to the ratio at time 0 to allow monitoring of DSB formation and repair by NHEJ. (**B**) The ratio of the HO-cut band to the *APN1* control was normalized to the ratio at 60 min when DSB formation was maximal. Results are the mean ± standard deviation of two independent experiments. DSBs were formed but only very slowly resected in *gal1* yeast, even at the 180 min time point when resection was nearly complete in *GAL1* strain.

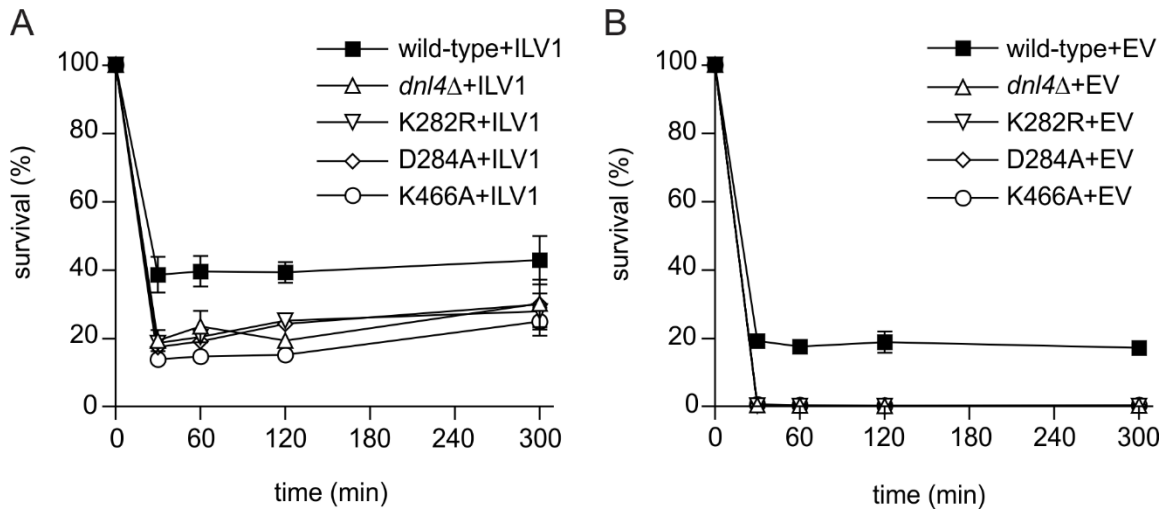


Figure 4-9. DSB repair by homologous recombination in Dnl4 catalytic mutant strains. *ILV1*-cs yeast bearing the indicated Dnl4 alleles, with (**A**) and without (**B**) a homologous *ILV1* donor fragment on a plasmid, were pre-grown to log phase in YPA-Glycerol and treated with 2% galactose for the indicated times to induce a DSB. Cells were then plated to glucose and survival was determined relative to untreated cells. Survival in wild-type with the homologous donor reflects DSB repair by both HR and c-NHEJ. The *dnl4* mutants are all defective in c-NHEJ so that survival with the donor reflects equivalent rates of HR. Results are the mean ± standard deviation of three independent experiments. EV, empty vector.

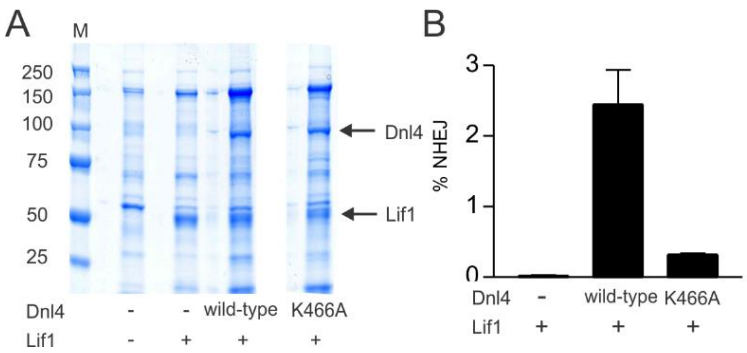
A

GATAAACGCGTGTATTACCCTG **TTAT**CCCTAGCGTCAGATCCTCTAGAA	WT	K282R	D284A	K466A
GATAAACGCGTGTATTA -----CCCT----AGCGTCAGATCCTCTAGAA	6/10	5/10	4/10	4/10
GATAAAC---------- GCGT---------CAGATCCTCTAGAA	3/10	1/10	4/10	1/10
GATAAACGCGTGTATTACCCTGT TA------GCGTCAGATCCTCTAGAA	1/10			
GATAAACGCGTGTA ---------TATCCCTAGCGTCAGATCCTCTAGAA		2/10		1/10
GATAAACGCGTGTATTACCC ---TATCCCTAGCGTCAGATCCTCTAGAA		1/10	1/10	2/10
GATAAACGCGTGTATTACCCTG ---------GCGTCAGATCCTCTAGAA		1/10		
GATAAACGCGT GT---------TTATCCCTAGCGTCAGATCCTCTAGAA				1/10
GATAAACGCGTGTATTACC ---a---CCCTAGCGTCAGATCCTCTAGAA			1/10	
GATAAACGCGTGTATTACCCTGTTAaCCCTAGCGTCAGATCCTCTAGAA				1/10

B

GATAAACGCGTGTATTACCCTG **TTAT**CCCTAGCGTCAGATCCTCTAGAA	WT	K282R	D284A	K466A
GATAAACGCGTGTATTACCCTGtTTATCCCTAGCGTCAGATCCTCTAGAA	2/5			
GATAAACGCGTGTATTACCC ----ATCCCTAGCGTCAGATCCTCTAGAA	1/5		1/3	
GATAAACGCGTGTA --------TTATCCCTAGCGTCAGATCCTCTAGAA	1/5			2/5
GATAAACGCGTGTATTACCCTGTTAT -CCTAGCGTCAGATCCTCTAGAA	1/5			
GATAAACGCGTG ----------TATCCCTAGCGTCAGATCCTCTAGAA		2/2		
GATAAACGCGTGTATTA -CCTGTTATCCCTAGCGTCAGAT CCTCTAGAA			1/3	
GATAAACGCGTGTATTACCCTGTTA -CCCTAGCGTCAGATCCTCTAGAA			1/3	
GATAAACGCGTGTAT TACCCTGT TATatCCCTAGCGTCAGATCCTCTAGAA				2/5
GATAAACGCGTGTATTACCCTG -TATCCCTAGCGTCAGATCCTCTAGAA				1/5

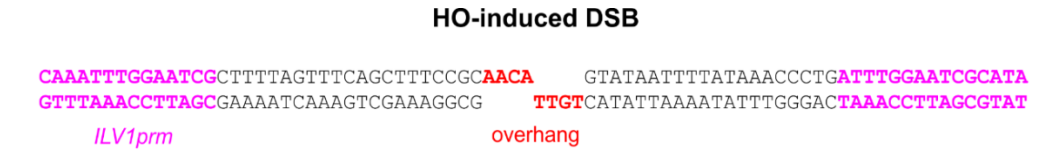
Figure 4-10. Imprecise joints in the suicide deletion assay. Sequenced imprecise DSB repair junctions from colonies formed in the suicide deletion assay. (**A**) Ade^{+}/white colonies that did not recleave the allele upon re-introduction of I-*Sce*I (see **Table 4-1**). (**B**) Ade^{-}/red colonies. In each, the sequence at the top left is the product of precise suicide deletion NHEJ, equivalent to an I-*Sce*I cut site. Bold type shows the location of the 4-base 3' overhang. Subsequent entries show recovered imprecise joints and the number of times they were observed in different strains. Underlined bases are microhomologies, dashes indicate deleted bases, and lower case indicates inserted bases.



C

	wild-type	K466A
TLHDDFLVEEK	115	91
TLHDDFLVEEKMDGER	36	37
TLHDDFLVEEK*MDGER	12	0
Total Dnl4 peptides	5653	6357

Figure 4-11. Impaired Dnl4-K466A auto-adenylation *in vivo*. (**A**) Protein gel showing expression of the Dnl4-Lif1 complex in strains transformed with plasmids bearing the indicated components. (**B**) NHEJ efficiency in the suicide deletion assay in yeast transformed with the same plasmids as in (**A**). Results are the mean ± standard deviation of three independent experiments. (**C**) Dnl4-K466A adenylation status *in vivo*. TLHDDFLVEEK is the fully cleaved and unadenylated tryptic peptide ending at K282. TLHDDFLVEEKMDGER and TLHDDFLVEEK*MDGER are not cleaved at K282, with K* indicating K282 adenylation. Results are from multiple lanes and mass spectrometry runs from two biological replicates for each of wild-type and Dnl4-K466A. Note that K466 is not contained within the peptides shown.



coded name	joint type	joint sequence
D+0_M4_L0_R0	precise	TTTTAGTTTCAGCTTTCCGCAACAGTATAATTTTATAAACCCTG AAAATCAAAGTCGAAAGGCGTTGTCATATTAAAATATTTGGGAC microhomology
D-4_M0_L4_R4	resected blunt	TTTTAGTTTCAGCTTTCCGCGTATAATTTTATAAACCCTG AAAATCAAAGTCGAAAGGCGCATATTAAAATATTTGGGAC
D+2_M1_L0_R1	+CA	templated fill-in TTTTAGTTTCAGCTTTCCGCAACACAGTATAATTTTATAAACCCTG AAAATCAAAGTCGAAAGGCGTTGTGTCATATTAAAATATTTGGGAC
D-3_M2_L3_R2	-ACA	TTTTAGTTTCAGCTTTCCGCAGTATAATTTTATAAACCCTG AAAATCAAAGTCGAAAGGCGTCATATTAAAATATTTGGGAC
D+4_M0_L0_R0	fill-in blunt	TTTTAGTTTCAGCTTTCCGCAACAAACAGTATAATTTTATAAACCCTG AAAATCAAAGTCGAAAGGCGTTGTTTGTCATATTAAAATATTTGGGAC
D+6_M0_L0_R0_I(AC)	fill-in plus	non-templated fill-in TTTTAGTTTCAGCTTTCCGCAACAACAACAGTATAATTTTATAAACCCTG AAAATCAAAGTCGAAAGGCGTTGTTGTTGTCATATTAAAATATTTGGGAC

D= delta, the number of bases lost/gained
M= the number of bases of microhomology
L= the number of bases deleted on the left overhang
R= the number of bases deleted on the right overhang
I= insertional nucleotides

Figure 4-12. Description of the joint identifiers used in next-generation sequencing. HO-induced DSB formation in the *ILV1*-cs system is shown at the top. The HO 3' overhangs are in red and the surrounding *ILV1* promoter sequence is in magenta. In the coded joint identifiers, "D" indicates the number of base pairs lost or gained in the joint overall, "M" indicates the number of microhomologous base pairs at the repair junction, "L" indicates the number of base pairs deleted from the left side of the DSB, as measured from the most distal base of the overhang, "R" similarly indicates the number of base pairs deleted from the right side, and "I" indicates any non-templated insertion nucleotides at the repair junction, read from the top strand. Common joint types are shown as examples. +CA and -ACA joint designations reflect the nomenclature used by Moore and Haber (Moore and Haber, 1996).

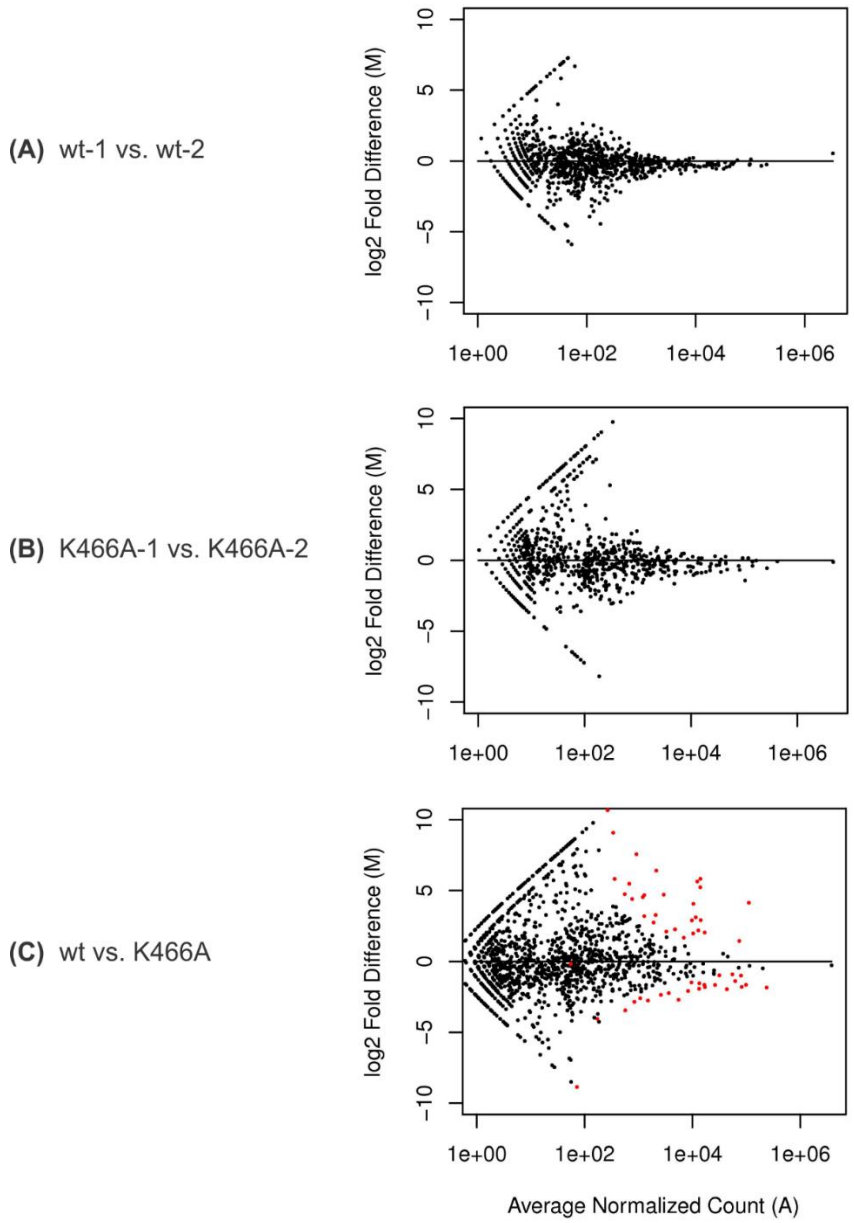


Figure 4-13. *ILV1*-cs next-generation DESeq analysis. MA plots of the output of DESeq runs that compared (**A**) the two wild-type 24-hour replicates to each other, (**B**) the two K466A 24-hour replicates to each other, and (**C**) the wild-type 24-hour replicates to the K466A 24-hour replicates. Data points represent individual joint types. Red points highlight joints with a Bonferroni-adjusted p-value <0.005. Results demonstrate a much higher concordance between replicates of the same sample than between wild-type and K466A.

Table 4-1. NHEJ precision in Ade$^+$ suicide deletion colonies. 'I-*Sce*I$^+$/total' indicates the fraction of Ade$^+$ survivors that were capable of being re-cleaved by I-*Sce*I, demonstrating that they had used precise NHEJ. 'Colony Survival' is the fraction of Ade$^+$ colonies that had survived by total, precise and imprecise NHEJ relative to all cells plated. 'Relative Efficiency' is the efficiency of total, precise and imprecise NHEJ for mutant strains relative to wild-type Dnl4.

	Joint analysis (%)			**Colony Survival (%)**			**Relative Efficiency (%)**		
***DNL4* genotype**	**I-*Sce*I$^+$/total**	**Precise**	**Imprecise**	**Total Ade$^+$**	**Precise**	**Imprecise**	**Total NHEJ**	**Precise**	**Imprecise**
Wild-type	112/120	93	7	2.2	2.0	0.1	100	100	100
***dnl4*Δ**	120/120	100	0	0.002	0.002	0	0.1	0.1	0
K282R	36/118	31	69	0.2	0.1	0.1	8	3	84
D284A	66/117	56	44	0.3	0.2	0.1	15	9	99
K466A	190/216	88	12	1.3	1.2	0.2	60	57	109
L750*	120/120	100	0	0.002	0.002	0	0.1	0.1	0

Table 4-2. Joint analysis of Ade$^-$ suicide deletion colonies. Ade$^-$ Met$^-$ colonies are inferred to have undergone suicide deletion, while Met$^+$ colonies survived galactose by other means. Met$^-$ colonies that were also positive for PCR flanking the expected DSB junction corresponded to imprecise NHEJ (see also **Figure 4-10B**), while PCR-negative colonies were inferred to correspond to large deletions. Percentages are calculated relative to all Ade$^-$ colonies analyzed for each strain.

***DNL4* genotype**	**Met$^-$**	**PCR$^+$**	**Large deletion (%)**	**Imprecise NHEJ (%)**	**Other (%)**
Wild-type	25/183	17/25	4	9	86
***dnl4*Δ**	32/141	0/26	23	0	77
K282R	37/140	3/24	23	3	74
D284A	21/140	5/20	11	4	85
K466A	27/190	20/25	3	11	86

Table 4-3. Genotype of yeast strains used in this study.

Strain	Genotype
YW1228	*MATα ade2*::*SD2-*::*STE3-MET15 his3*Δ1 *leu2*Δ0 *met15*Δ0 *ura3*Δ0
YW1230	YW1228 *dnl4*Δ::*kanMX4*
YW2042	YW1228 *dnl4*-K282R
YW2043	YW1228 *dnl4*-D284A
YW2044	YW1228 *dnl4*-K466A
YW2051	YW1228 *dnl4*-L750*
YW2304	YW1228 *nej1*::*kanMX4*
YW2305	YW2042 *nej1*::*kanMX4*
YW2306	YW2043 *nej1*::*kanMX4*
YW2307	YW2044 *nej1*::*kanMX4*
YW2083	*MATa ade2*-M7 *his3*Δ200 *leu2- lys*2-801 *trp1*Δ63 *ura*3-52
YW1993	*MAT*a-inc::AmpR-35S *can1*Δ::*GAL1*-QPCR *DNL4*-13Myc::*hisMX6 GAL1*prm-HOcs *gal1*::HO *his3*Δ1 *leu2*Δ0 *met15*Δ0 *ura3*Δ0
YW2033	YW1993 *dnl4*-K282R-13Myc::*hisMX6*
YW2034	YW1993 *dnl4*-D284A-13Myc::*hisMX6*
YW2035	YW1993 *dnl4*-K466A-13Myc::*hisMX6*
YW2121	YW1993 *lif1*Δ::*kanMX4*
YW2162	*MAT*a-inc::AmpR-35S *can1*Δ::*GAL1*-QPCR *CDC9*-13Myc::*hisMX6 GAL1*prm-HOcs *gal1*::HO *his3*Δ1 *leu2*Δ0 *met15*Δ0 *ura3*Δ0
YW2163	YW2162 *dnl4*Δ::*kanMX4*
YW2164	YW2162 *dnl4*-K466A
YW1858	*MAT*a-inc::*LEU2 can1*Δ::*ILV1*-QPCR *gal1*::HO *his3*Δ1 *ILV1*prm::HOcs *leu2*Δ0 *met15*Δ0 *ura3*Δ0
YW2107	YW1858 *dnl4*Δ::*kanMX4*
YW2100	YW1858 *dnl4*-K282R
YW2101	YW1858 *dnl4*-D284A
YW2102	YW1858 *dnl4*-K466A
YW2188	*MAT*a-inc::*LEU2 can1*Δ::*ILV1*-ctrl *gal1*::HO *his3*Δ1 *ILV1*prm::HOcs *leu2*Δ0 *met15*Δ0 *ura3*Δ0
YW2212	YW2188 *dnl4*-K466A
YW2213	YW2188 *dnl4*Δ::*kanMX4*
YW126	*MAT*a *leu2*Δ *trp*1Δ *ura*3-52Δ *prb*1Δ *pep*4Δ *prc*1Δ
YW2189	YW126 *dnl4*Δ::*kanMX4*
YW2166	*MAT*a-inc::AmpR-35S *can1*Δ::*GAL1*-QPCR *LIF1*-13Myc::*hisMX6 GAL1*prm-HOcs *gal1*::HO *his3*Δ1 *leu2*Δ0 *met15*Δ0 *ura3*Δ0
YW2167	YW2166 *dnl4*Δ::*kanMX4*
YW2308	YW2166 *dnl4*-K466A
YW1750	*MAT*a-inc::AmpR-35S *can1*Δ::*GAL1*-QPCR *KU80*-13Myc::*hisMX6 GAL1*prm-HOcs *gal1*::HO *his3*Δ1 *leu2*Δ0 *met15*Δ0 *ura3*Δ0
YW1833	YW1750 *dnl4*Δ::*kanMX4*
YW2309	YW1750 *dnl4*-K466A

File 4-1. *ILV1*-cs next-generation sequencing results. Online Excel file (http://journals.plos.org/plosgenetics/article/asset?unique&id=info:doi/10.1371/journal.pgen.1003599.s007) containing data on the precise and major imprecise joints observed in the *ILV1*-cs next-generation sequencing experiments. A major imprecise joint was present in ≥0.1% of the reads of any individual sample. (**A**) Worksheet "Counts" tabulates the read counts for each sample replicate for the observed joints. Joint assignments were made as described in **Materials and Methods**. Joints are sorted by the frequency of occurrence. Columns are labeled in format "sample_timepoint_replicate". Joint identifiers are described in **Figure 4-12**. One read (row 20) did not match an expected joint and is listed as the observed read sequence. (**B**) Worksheet "Sequences" shows, for each observed joint, the joint structure with microhomologous bases in the middle, as well as the joint sequences without added spaces. Joints are in the same order as in (**A**). (**C**) Worksheet 'DESeq" shows the output of the DESeq analysis performed on the wt and K466A 24-hour data. "baseMean" is the read count after normalization to allow for inter-sample comparisons, "padj" is the Bonferroni-adjusted p-value of the joint, comparing wt replicates to K466A replicates, and "K466A effect" is the direction of statistically significant K466A mutant effects ($p<0.005$). Joints are sorted by padj.

File 4-2. Perl scripts used in ILV1-cs next-generation sequencing. Online Zip file (http://journals.plos.org/plosgenetics/article/asset?unique&id=info:doi/10.1371/journal.pgen.1003599.s008) containing three Perl scripts used to assign sequence reads to joint identifiers. See scripts for details.

Reference

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Bindra, R.S., Goglia, A.G., Jasin, M., and Powell, S.N. (2013). Development of an assay to measure mutagenic non-homologous end-joining repair activity in mammalian cells. Nucleic Acids Res *41*, e115.

Boboila, C., Oksenych, V., Gostissa, M., Wang, J.H., Zha, S., Zhang, Y., Chai, H., Lee, C.S., Jankovic, M., Saez, L.M., *et al.* (2012). Robust chromosomal DNA repair via alternative end-joining in the absence of X-ray repair cross-complementing protein 1 (XRCC1). Proc Natl Acad Sci U S A *109*, 2473-2478.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P., and Boeke, J.D. (1998). Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast *14*, 115-132.

Chiruvella, K.K., Liang, Z., and Wilson, T.E. (2013). Repair of double-strand breaks by end joining. Cold Spring Harbor perspectives in biology *5*, a012757.

Chiruvella, K.K., Sebastian, R., Sharma, S., Karande, A.A., Choudhary, B., and Raghavan, S.C. (2012). Time-dependent predominance of nonhomologous DNA end-joining pathways during embryonic development in mice. J Mol Biol *417*, 197-211.

Chistiakov, D.A. (2010). Ligase IV syndrome. Adv Exp Med Biol *685*, 175-185.

Clerici, M., Mantiero, D., Guerini, I., Lucchini, G., and Longhese, M.P. (2008). The Yku70-Yku80 complex contributes to regulate double-strand break processing and checkpoint activation during the cell cycle. EMBO Rep *9*, 810-818.

Cottarel, J., Frit, P., Bombarde, O., Salles, B., Negrel, A., Bernard, S., Jeggo, P.A., Lieber, M.R., Modesti, M., and Calsou, P. (2013). A noncatalytic function of the ligation complex during nonhomologous end joining. J Cell Biol *200*, 173-186.

Daley, J.M., Palmbos, P.L., Wu, D., and Wilson, T.E. (2005). Nonhomologous end joining in yeast. Annu Rev Genet *39*, 431-451.

Daley, J.M., and Wilson, T.E. (2005). Rejoining of DNA double-strand breaks as a function of overhang length. Mol Cell Biol *25*, 896-906.

Della-Maria, J., Zhou, Y., Tsai, M.S., Kuhnlein, J., Carney, J.P., Paull, T.T., and Tomkinson, A.E. (2011). Human Mre11/human Rad50/Nbs1 and DNA ligase IIIalpha/XRCC1 protein

complexes act together in an alternative nonhomologous end joining pathway. J Biol Chem *286*, 33845-33853.

Dore, A.S., Furnham, N., Davies, O.R., Sibanda, B.L., Chirgadze, D.Y., Jackson, S.P., Pellegrini, L., and Blundell, T.L. (2006). Structure of an Xrcc4-DNA ligase IV yeast ortholog complex reveals a novel BRCT interaction mode. DNA Repair (Amst) *5*, 362-368.

Ellenberger, T., and Tomkinson, A.E. (2008). Eukaryotic DNA ligases: structural and functional insights. Annu Rev Biochem *77*, 313-338.

Girard, P.M., Kysela, B., Harer, C.J., Doherty, A.J., and Jeggo, P.A. (2004). Analysis of DNA ligase IV mutations found in LIG4 syndrome patients: the impact of two linked polymorphisms. Hum Mol Genet *13*, 2369-2376.

Grob, P., Zhang, T.T., Hannah, R., Yang, H., Hefferin, M.L., Tomkinson, A.E., and Nogales, E. (2012). Electron microscopy visualization of DNA-protein complexes formed by Ku and DNA ligase IV. DNA Repair (Amst) *11*, 74-81.

Hammel, M., Yu, Y., Fang, S., Lees-Miller, S.P., and Tainer, J.A. (2010). XLF regulates filament architecture of the XRCC4.ligase IV complex. Structure *18*, 1431-1442.

Ira, G., Pellicioli, A., Balijja, A., Wang, X., Fiorani, S., Carotenuto, W., Liberi, G., Bressan, D., Wan, L., Hollingsworth, N.M., *et al.* (2004). DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. Nature *431*, 1011-1017.

Karathanasis, E., and Wilson, T.E. (2002). Enhancement of Saccharomyces cerevisiae end-joining efficiency by cell growth stage but not by impairment of recombination. Genetics *161*, 1015-1027.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lieber, M.R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu Rev Biochem *79*, 181-211.

Ma, J.L., Kim, E.M., Haber, J.E., and Lee, S.E. (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.

Ma, Y., Lu, H., Tippin, B., Goodman, M.F., Shimazaki, N., Koiwai, O., Hsieh, C.L., Schwarz, K., and Lieber, M.R. (2004). A biochemically defined system for mammalian nonhomologous DNA end joining. Mol Cell *16*, 701-713.

Mahaney, B.L., Hammel, M., Meek, K., Tainer, J.A., and Lees-Miller, S.P. (2013). XRCC4 and XLF form long helical protein filaments suitable for DNA end protection and alignment to facilitate DNA double strand break repair. Biochem Cell Biol *91*, 31-41.

Mladenov, E., and Iliakis, G. (2011). Induction and repair of DNA double strand breaks: the increasing spectrum of non-homologous end joining pathways. Mutat Res *711*, 61-72.

Moore, J.K., and Haber, J.E. (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. Mol Cell Biol *16*, 2164-2173.

Nair, P.A., Nandakumar, J., Smith, P., Odell, M., Lima, C.D., and Shuman, S. (2007). Structural basis for nick recognition by a minimal pluripotent DNA ligase. Nat Struct Mol Biol *14*, 770-778.

O'Driscoll, M., Cerosaletti, K.M., Girard, P.M., Dai, Y., Stumm, M., Kysela, B., Hirsch, B., Gennery, A., Palmer, S.E., Seidel, J., *et al.* (2001). DNA ligase IV mutations identified in patients exhibiting developmental delay and immunodeficiency. Mol Cell *8*, 1175-1185.

Oermann, E.K., Wu, J., Guan, K.L., and Xiong, Y. (2012). Alterations of metabolic genes and metabolites in cancer. Semin Cell Dev Biol *23*, 370-380.

Palmbos, P.L., Daley, J.M., and Wilson, T.E. (2005). Mutations of the Yku80 C terminus and Xrs2 FHA domain specifically block yeast nonhomologous end joining. Mol Cell Biol *25*, 10782-10790.

Palmbos, P.L., Wu, D., Daley, J.M., and Wilson, T.E. (2008). Recruitment of Saccharomyces cerevisiae Dnl4-Lif1 complex to a double-strand break requires interactions with Yku80 and the Xrs2 FHA domain. Genetics *180*, 1809-1819.

Pascal, J.M., O'Brien, P.J., Tomkinson, A.E., and Ellenberger, T. (2004). Human DNA ligase I completely encircles and partially unwinds nicked DNA. Nature *432*, 473-478.

Pascal, J.M., Tsodikov, O.V., Hura, G.L., Song, W., Cotner, E.A., Classen, S., Tomkinson, A.E., Tainer, J.A., and Ellenberger, T. (2006). A flexible interface between DNA ligase and PCNA supports conformational switching and efficient ligation of DNA. Mol Cell *24*, 279-291.

Riballo, E., Doherty, A.J., Dai, Y., Stiff, T., Oettinger, M.A., Jeggo, P.A., and Kysela, B. (2001). Cellular and biochemical impact of a mutation in DNA ligase IV conferring clinical radiosensitivity. J Biol Chem *276*, 31124-31132.

Riballo, E., Woodbine, L., Stiff, T., Walker, S.A., Goodarzi, A.A., and Jeggo, P.A. (2009). XLF-Cernunnos promotes DNA ligase IV-XRCC4 re-adenylation following ligation. Nucleic Acids Res *37*, 482-492.

Rucci, F., Notarangelo, L.D., Fazeli, A., Patrizi, L., Hickernell, T., Paganini, T., Coakley, K.M., Detre, C., Keszei, M., Walter, J.E., *et al.* (2010). Homozygous DNA ligase IV R278H mutation in mice leads to leaky SCID and represents a model for human LIG4 syndrome. Proc Natl Acad Sci U S A *107*, 3024-3029.

Shibata, A., Conrad, S., Birraux, J., Geuting, V., Barton, O., Ismail, A., Kakarougkas, A., Meek, K., Taucher-Scholz, G., Lobrich, M., *et al.* (2011). Factors determining DNA double-strand break repair pathway choice in G2 phase. EMBO J *30*, 1079-1092.

Shim, E.Y., Hong, S.J., Oum, J.H., Yanez, Y., Zhang, Y., and Lee, S.E. (2007). RSC mobilizes nucleosomes to improve accessibility of repair machinery to the damaged chromatin. Mol Cell Biol *27*, 1602-1613.

Simsek, D., Brunet, E., Wong, S.Y., Katyal, S., Gao, Y., McKinnon, P.J., Lou, J., Zhang, L., Li, J., Rebar, E.J., *et al.* (2011). DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation. PLoS Genet *7*, e1002080.

Sriskanda, V., and Shuman, S. (1998). Mutational analysis of Chlorella virus DNA ligase: catalytic roles of domain I and motif VI. Nucleic Acids Res *26*, 4618-4625.

Sriskanda, V., and Shuman, S. (2002). Role of nucleotidyltransferase motifs I, III and IV in the catalysis of phosphodiester bond formation by Chlorella virus DNA ligase. Nucleic Acids Res *30*, 903-911.

Srivastava, M., Nambiar, M., Sharma, S., Karki, S.S., Goldsmith, G., Hegde, M., Kumar, S., Pandey, M., Singh, R.K., Ray, P., *et al.* (2012). An inhibitor of nonhomologous end-joining abrogates double-strand break repair and impedes cancer progression. Cell *151*, 1474-1487.

Symington, L.S., and Gautier, J. (2011). Double-strand break end resection and repair pathway choice. Annu Rev Genet *45*, 247-271.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Wilson, T.E. (2002). A genomics-based screen for yeast mutants with an altered recombination/end-joining repair ratio. Genetics *162*, 677-688.

Wilson, T.E., Grawunder, U., and Lieber, M.R. (1997). Yeast DNA ligase IV mediates non-homologous DNA end joining. Nature *388*, 495-498.

Wu, D., Topper, L.M., and Wilson, T.E. (2008). Recruitment and dissociation of nonhomologous end joining proteins at a DNA double-strand break in Saccharomyces cerevisiae. Genetics *178*, 1237-1249.

Yates, J.R., 3rd, Eng, J.K., McCormack, A.L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem *67*, 1426-1436.

Zhang, Y., Hefferin, M.L., Chen, L., Shim, E.Y., Tseng, H.M., Kwon, Y., Sung, P., Lee, S.E., and Tomkinson, A.E. (2007). Role of Dnl4-Lif1 in nonhomologous end-joining repair complex assembly and suppression of homologous recombination. Nat Struct Mol Biol *14*, 639-646.

CHAPTER 5

Overhang Polarity of Chromosomal Double-Strand Breaks Impacts Kinetics and Fidelity of Yeast Nonhomologous End Joining

Abstract

Nonhomologous end joining (NHEJ), which directly rejoins two ends of broken DNA, is the dominant repair pathway of double-stranded breaks (DSBs) in cells with limited or no 5' resection. Physiological DSBs often harbor diverse break structures, such as various overhang polarities, sequences, length and base lesions, which can complicate rejoining and lead to mutations. Using extrachromosomal plasmid-rejoining assays, previous studies reported that diverse break structures can influence the usage of repair factors and fidelity of NHEJ. But the lack of an efficient system to generate chromosomal 5'-overhanging DSBs (5' DSBs) limited further study. To better understand how overhang polarity of chromosomal DSBs affects the kinetics and fidelity of yeast NHEJ, we first developed an efficient system to induce site-specific 5' DSBs in the *S. cerevisiae* genome using an optimized zinc finger nuclease (ZFN), which shows comparable activity to the well-established HO endonuclease used to induce chromosomal 3' overhanging DSBs (3' DSBs). By monitoring the formation and repair of induced DSBs by qPCR, our results suggests that ZFN-mediated 5' DSBs are rejoined more rapidly than HO-mediated 3' DSBs. Consistently, results of chromatin immunoprecipitation (ChIP) show that more NHEJ factors are recruited with higher efficiency to ZFN-mediated 5' DSBs. Also, our ChIP results reveals that yeast Tdp1, an end-processing enzyme proposed to

control fidelity of NHEJ in 5’ DSBs, is recruited exclusively to 5’ DSBs, but not to 3’ DSBs. Results of ligation-mediated qPCR (LM-qPCR), a high-resolution approach to detect break-end modifications, shows that 5’ DSBs have more robust end-processing than 3’ DSBs. Conversely, next-generation sequencing (NGS) of the repair joints reveals that 5’ DSBs have a higher mutation frequency. In addition, our NGS results validate the differential requirement of Pol4 for gap-fillings in 3’ and 5’ DSBs. In summary, using several new experimental approaches, results of this study provide new insights into the influence of overhang polarity on yeast NHEJ of chromosomal DSBs.

Introduction

DNA double-strand breaks (DSBs) are one of the most cytotoxic DNA lesions which greatly threaten the integrity of the genome and can trigger detrimental genome translocations (Anand *et al.*, 2013; Ghezraoui *et al.*, 2014; Piganeau *et al.*, 2013). However, they are also essential substrates for normal cell activities such as meiotic chromosome exchange and immunoglobulin receptor production (Khanna and Jackson, 2001; van Gent *et al.*, 2001).

DSBs are repaired by two major pathways. The first pathway called homologous recombination (HR) repairs the DSBs with the usage of another homologues sequence. HR and its subpathways are initiated by 5' to 3' degradation of one strand of DNA, called 5' resection, by various endo/exonucleases (e.g. Mre11, Exo1, Sae2 and Dna2 in yeast) (Paull, 2010). The search for homology is mediated by the preserved and exposed 3' strand coated with Rad51 proteins (Jasin and Rothstein, 2013). In cells with limited or no resection, nonhomologous end joining (NHEJ), which directly rejoins two ends of the broken DNA, is the dominant repair pathway of DSBs (Chiruvella *et al.*, 2013b) The conserved actions of NHEJ in different kingdoms can be organized into three steps, taking yeast NHEJ for instance. (i) Rapid binding of the DSBs by the YKu70/80 heterodimers, which recruit other repair factors and protect the break-ends from degradation, marking the onset of end-joining. (ii) End processing that modifies the break-ends by enzymes such as the nuclease complex MRX (Mre1l-Rad50-Xrs2) and polymerase Pol4. This is a necessary step when direct religation of the two break-ends is not possible (e.g. with incompatible overhangs). (iii) Ligation of two strands of the DSBs by the specialized NHEJ ligase LigIV. Notably, although these three steps can be described in a logical order, the actual reaction of NHEJ is likely dynamic with various repair proteins rapidly

associating to and dissociating from the break, similar to the action demonstrated with single-molecule real-time microscopy in nucleotide excision repair (NER) (Ghodke *et al.*, 2014).

Because of the nature of NHEJ, repair progression and fidelity is greatly affected by the local structures of the two break-ends (e.g. various overhang polarities, sequences, length and base lesions), and the end-processing required for these structures (e.g. cleavage, polymerization, phosphorylation and ligation of terminal nucleotides) (Chiruvella *et al.*, 2013b) .Studies of diverse break structures have used mainly in vitro biochemistry approaches, due to the ease of generating defined break structures (Makharashvili *et al.*, 2014; Cannavo and Cejka, 2014). We previously studied the in-vivo rejoining of pre-made extrachromosomal DSBs with various break structures by transfection of linearized plasmid ligated with synthetic adapters in yeast (Daley *et al.*, 2005; Daley and Wilson, 2008). A similar approach was also used in recent study of end-joining in human cell lines (Waters *et al.*, 2014).

In this study, we further investigate the impact of overhang polarity of chromosomal DSBs by developing an efficient, inducible system to generate 5' overhanging DSBs using zinc finger nuclease (ZFN) in the yeast genome. Our results of the rejoining of the ZFN-mediated 5' overhanging DSBs (5' DSBs) and HO-mediated 3' overhanging DSBs (3' DSBs) in various assays, suggest that overhang polarity can greatly impact the repair kinetics and fidelity of yeast nonhomologous end joining.

Materials and Methods

Yeast growth and manipulation. Yeast strains used in this study were isogenic derivatives of BY4741 (Brachmann *et al.,* 1998). Gene disruptions and modified alleles were made using a PCR-mediated technique (Brachmann *et al.,* 1998) or a URA3 pop-in/pop-out method (McCormick *et al.,* 1995) (Peterson *et al.,* 1995). Genetic manipulations were confirmed by PCR and sequencing. The chromosomal ZFN cut sites was made by using a 24-bp sequence consist of two 9-bp ZFN recognition sites franking a 6-bp *BamH*I site to replace the 24-bp core HO recognition site in the previously described strain bearing an HO cut site in *ILV1* prompter (Wu *et al.,* 2008). The ZFN recognition sites for the four ZFNs used in this study are: 5'-GTTGGTGCT for StickyC.ZFNIII; 5'-GGGGAAGAA for QQR. ZFN; 5'-GAAGATGGT for GFP. ZFN1; 5'-GACGACGGC for GFP. ZFN2. Codon optimizaiotn of the GFP.ZFN2 coding sequence was made by the GeneArt® gene synthesis technique from Life Technologies Inc. Chromosomal integration of the *GAL1-V10prm::GFP.ZFN2(CO)* expression cassttee was made by first knocking out the native *GAL1* locus by a *LEU2* marker, and then replacing the LEU2 marker with PCR product carrying *GAL1-V10prm::GFP.ZFN2(CO)* and a *URA3* maker. Yeast were grown at 30 ^{0}C in either rich medium containing 1% yeast extract, 2% peptone and 40 μg/ml adenine (YPA) or synthetic defined (SD) medium with either 2% glucose, 2% galactose or 3% glycerol as the carbon source.

Survival assay. Overnight cultures in selection medium or rich dextrose liquid medium (YPAD) were inoculated into YPA medium with 3% glycerol as the carbon source and grown overnight to a final OD600 of 0.3–0.6. Galactose was then added to 2% final con-centration to induce endonuclease expression. At varying time-points after induction aliquots were serially

diluted, plated to YPAD medium, and incubated at 30 ^{0}C for 3 days. Survival rate was measured as the ratio of corrected colony counts at each time point to corrected counts before galactose addition.

DSB-monitoring assay. Yeast cultures were handled in the same way as in survival assay described above. But after induction of the endonuclease expression by 2% galactose, cells were frozen on dry ice at varying time-points. Genomic DNA extracted and purified from the frozen samples was subject to standard qPCR resection using the Agilent Mx3005P qPCR System. Cut site allele was amplified with qPCR primers 5' AAAAAGCGCAGCGGGTAG and 5'-CTCAAAGCAGCAACAACAAAAGT. A control allele in the CAN1 locus was amplified with qPCR primers 5'-GTGGCCTTTGCTGTTTGC and 5'-CGAGATACGATTACTCCAGTTCC. The percentage of cells with intact or rejoined cut sites was calculated by normalizing the ΔCT of the time-point sample in question to the one in the 0-hour time point.

Chromatin immunoprecipitation (ChIP). 13Myc sequence was inserted right before the stop codon in the C-terminus of the native chromosomal *YKU80, DNL4, POL4, XRS2, EXO1* and *TDP1* genes as previously described (Wu *et al.*, 2008). ChIP used protocols as previously described (Wu *et al.*, 2008).

Linker preparation and ligation-mediated qPCR (LM-qPCR). Linkers used in LM-qPCR assay were made by annealing two PAGE-purified synthetic oligonucleotides from IDT. Two oligos were mixed with the same final concentration of 10uM in annealing buffer (10 mM Tris, pH 7.5–8.0, 50 mM NaCl, 1 mM EDTA). The mixture was placed in a heat block at 90-95 ^{0}C for 5 minutes, followed by slow cooling to RT for 1-2 hours by turning off the heat block.

Annealed linkers were stored at -20 ^{0}C at 10uM. Linker with a 4-nt overhang shown in **Figure 5-7A** were annealing product of 5'-CTTGAGACGA/3ddC/ and 5'-/5Phos/GATCGTCGTCTCAAGTCTAGCCTTCTCCGTGCA; Linker with a 5-nt overhang shown in **Figure 5-7B** were annealing product of 5'-CTTGAGACGA/3ddC/ and 5'-/5Phos/GGATCGTCGTCTCAAGTCTAGCCTTCTCCGTGCA; Linker with 4-nt overhang shown in **Figure 5-9A** were annealing product of 5'-TGCACGGAGAAGGCTAGAGTAGATAGTTGAGTCGACAACA and 5'-/5Phos/GTCGACTCAACTATCTAC/3ddC/.

T4 PNK treatment of DNA substrates was carried out prior to ligation. Ligation reactions were carried out with the same condition using the Promega quick ligation system for 2 hours at 16 ^{0}C on thermal cycler with 1nM of linker and GeneClean purified genomic DNA. The ligation reactions were precisely terminated at 2 hours by heat inactivation at 95 ^{0}C for 5 minutes. Diluted ligation reaction were then subject to standard qPCR using the Agilent Mx3005P qPCR System. Ligated products were amplified with qPCR primers 5'-TGCACGGAGAAGGCTAGA and 5'-GAACTCAAAGCAGCAACAACA. A control allele in the CAN1 locus was amplified with qPCR primers 5'-GTGGCCTTTGCTGTTTGC and 5'-CGAGATACGATTACTCCAGTTCC. Ligation efficiency, which represents the percentage of unmodified break-ends, was calculated by normalizing qPCR signal of ligated products, by (i) the percentage of cells with a broken cut site (generated by the DSB-monitoring assay discussed above), (ii) ligation efficiency of linkers to restriction enzyme linearized plasmids mixed with genomic DNA (e.g. in **Figure 5-8B**).

Next generation sequencing and bioinformatics analysis of the joint types. Primers with specific 3' end sequence of 5'-AGGGCAAAAAGAAAAAGCGCA and 5'-GTTTTATCAAGGAAGGTGACA were used to amplify the cut site allele in the *ILV1* promoter, so that one primer was close enough to the cut site that the 101-nt Illumina sequencing reads were sufficient to cross the cut site, and thus also any repair joints at the cut site, as long as this priming sequence was not deleted. Primers also included a sample-specific 6-nt barcode sequence to allow multiplexing of multiple samples in a single lane, and 4 random nucleotides on the 5' end of the barcode to improve the initial cluster detection accuracy for the sequencer. The same primers were able to amplify another artificial allele with an *Arabidopsis* genomic sequence in the *CAN1* locus. Adapter sequences were then added with further PCR with primers 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT and 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT. PCR used to enrich the cut site allele and control allele was optimized to have equal efficiency and it was ended many cycles prior to saturation to prevent bias. The sequencing library was then subject to Illumina 100-bp paired end sequencing using the HiSeq2000 platform.

The target sequence read was first filtered using custom Perl scripts for the presence of a recognizable barcode index sequence at the 5' end, allowing up to one base mismatch, which still ensured appropriate sample identification. Next, reads were filtered for a recognizable PCR primer sequence, which is defined by the primer itself and thus uninformative. The remaining

read bases were assumed to define a valid amplicon of the intact or rejoined cut site. To identify joint types, the sequence surrounding the input DSB site was pre-processed into a list of all possible blunt and microhomology joints, including the starting sequence itself. When the remaining portion of the sequence read matched one of these reads exactly, it was assigned to that joint. If it did not match exactly, but did match exactly any of the possible one-base substitutions of a given joint (to allow for sequencing errors), it was again assigned to that joint. Reads that matched more than one possible joints were called "ambiguous". Reads that matched no joints were declared "unknown" and propagated as raw sequence for manual inspection, which generally showed them to be lower quality reads with more than one sequencing error. Because such reads were rare, confidence is high that the large majority of reads were properly matched to their correct joint sequence. Tallies of joint counts were made and normalized with the read counts of the control alleles to allow comparison between samples.

Yeast two-hybrid. Yeast two-hybrid construct with full-length Tdp1 were created by gap repair in yeast. The Tdp1 bait construct mated with cells of our NHEJ two hybrid prey array as previously described (Palmbos *et al.*, 2005; Palmbos *et al.*, 2008).

Plasmid-rejoining assay. *Bgl*II linearized plasmids carrying a 5' DSB in the *ADE2* coding sequencing were mixed in a 10 to 1 ratio with another circular plasmid with *HIS3* marker. Yeast cells were transformed with the mixture of two plasmids and then subjected to selection on medium lacking uracil or histidine. Precise rejoining of the *Bgl*II site will restore a functional coding sequence of the *ADE2*, while rejoining with deletion or insertion, even in frame, will not. Thus the ratio of the number of red cells (Ade-) over number of total cells in –Ura plate represent the overall mutation frequency, while the ratio the number of Ura-positive cells over the number

of His-positive cells can be used to compare overall repair efficiency among cells that were transformed by the same mixture of plasmids.

Results

Efficient induction of 5' overhanging DSBs in *S. cerevisiae* genome by ZFN. The well-established yeast mating-type switch endonuclease HO can induce chromosomal 3' overhanging DSBs (3' DSBs) (Moore and Haber, 1996). We previously established an efficient HO system by inserting the HO coding sequence downstream of the native *GAL1* promoter, and placing the HO cut site (HOcs) in a well-characterized nucleosome-free region in the *ILV1* promoter (Wu *et al.*, 2008) (**Figure 5-1A**). To study the influence of overhang polarities of chromosomal DSBs on NHEJ, we developed an efficient system to generate chromosomal 5' overhanging DSBs (5' DSBs). Engineered endonucleases such as zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) can generate chromosomal 5' DSBs upon dimerization of two FokI DNA cleavage domains after binding to their target sites (DNA binding sites) (Gaj *et al.*, 2013). Both ZFNs and TALENs, as important genome-editing tools, have been successfully used in various organisms and cell types to induce site-specific mutagenesis through DSB repair by NHEJ (Gaj *et al.*, 2013). However, the cleavage efficiency of the currently available ZFN and TALEN systems is much lower than the HO system, taking many hours to generate a small fraction of breaks. Notably, because the HOcs is an essential part of the DNA binding sequence of HO, the 3' DSBs induced by HO will always have the fixed overhang sequence (5'-AACA) on one break-end. Unlike HO, the cut site and DNA binding site of ZFN are separated and the FokI DNA cleavage domain of ZFN is sequence non-specific. Therefore, ZFN can generate any 4-nt overhang sequence (**Figure 5-1B**).

In order to generate chromosomal 3' and 5' DSBs with comparable efficiency, we first selected the most active ZFN from four candidates in yeast using a previously reported survival

assay (Wu *et al.*, 2008). These ZFNs include StickyC-ZFNIII and QQR.-ZFN, which successfully induced site-specific mutagenesis in model plant *Arabidopsis* and tobacco with relatively high efficiency (Weinthal *et al.*, 2013; Tovkach *et al.*, 2009; Lloyd *et al.*, 2005)(and unpolished data), and two ZFNs engineered to target the *GFP* gene GFP, ZFN1 and GFP. ZFN2. We first tried to express ZFN using an extrachromosomal plasmid (**Figure 5-1C**). All plasmids used the same *GAL1* promoter to drive the ZFN coding sequences, and CEN/ARS centromere to yield similar 1-2 copies of plasmid per cell. Following galactose induction, ZFNs, would presumably bind to their target sites and cleave the ZFN cut site (ZFNcs) placed in the same *ILV1* promoter locus as HOcs. Notably, throughout this study, unless otherwise specified, we used exclusively haploid strains to avoid repair of the induced DSBs by HR. In addition, we used in this experiment end-joining defective *yku70*Δ strains to prevent NHEJ of the induced DSBs (**Figure 5-1C**). Therefore, the decline of the survival rate in **Figure 5-1C** directly reflected the cleavage of ZFNs.

Out of the four ZFN candidates tested, GFP.ZFN2 showed the highest activity (**Figure 5-1C**). Intriguingly, the co-expression of both GFP. ZFN1 and GFP. ZFN2 similar to their previous application in targeting the *GFP* gene in yeast [Citation] did not yield more efficient cleavage (**Figure 5-1C**). This may likely came from the fact that GFP. ZFN1 is less efficient than GFP.ZFN2 (**Figure 5-1C**). Because of that, and also for the simplicity of strain construction, we chose GFP.ZFN2 as the ZFN for further optimization. Importantly, the survival assay cannot tell where in the genome cleaved by ZFNs, because non-specific cleavage in other places of the genome will also contribute to the decline of the survival rate. To assess the potential non-specific cleavage activity of the selected GFP.ZFN2, we carried out similar survival assay for a

strain lacking the ZFNcs of GFP.ZFN2. The results in **Figure 5-1C** showed that non-specific cleavage, although still possible, is not significant when expressing GFP.ZFN2 from a low-copy-number plasmid. However, strong non-specific cleavage can be detected when expressing GFP. ZFN2 from a high-copy-number 2 micron plasmid (**Figure 5-2**). Furthermore, strains carrying the high-copy-number plasmid with *GAL1prm::GFP.ZFN2* expression cassette showed slow-growth phenotype even in glucose containing medium YPAD, suggesting the leaky expression of GFP.ZFN2 in the presence of glucose was toxic for cell growth (data not shown).

To optimize the cleavage activity of the selected GFP.ZFN2 to be more comparable to the efficiency of our HO system shown in **Figure 5-1B**, we integrated the GFP.ZFN2 coding sequence after the native *GAL1* promoter replacing the native *GAL1* gene in a way similar to our HO system. Therefore, the strains that we used in our studied with chromosomal integration of the endonuclease genes are *gal* and cannot ferment galactose. Chromosomal integration of GFP.ZFN2 coding sequencing improved its cleaving activity (compare **Figure 5-1C**, **D**), and also simplified the experimental protocol by removing the need of maintaining the plasmid in the yeast cells prior to galactose induction. To further improve our ZFN system, we optimized codon usage of the GFP.ZFN2 coding sequence for better expression in *S. cerevisiae* (**Figure 5-2B**). GFP.ZFP2(CO), the codon optimized GFP.ZFN2, showed further improved cleaving activity coincident with a stronger non-specific cleavage (**Figure 5-1D**). Although concerning, the fact that the killing effect in cells lacking the cut site started to appear after 60 min and most of the cells with the cut site have been killed by this time (**Figure 5-1D**), clearly suggested that the ZFNcs in the *ILV1* promoter is cut more efficiently. Furthermore, because of the focus of our study was to investigate the immediate repair process after the induction of the site-specific

DSBs, moderate non-specific cleavage occurring much later outside of the studied ZFNcs will have a minor effect on results. However this did limit our ability to apply our ZFN system for assays which require cell survival after extensive induction of the endonuclease, such as mutant selection by incubating cells on induction medium for days (Wu *et al.*, 2008). When we tried to do so, all twenty survivors had random point mutations in the *GAL1prm-GFP.ZFN2(CO)* expression. Surprisingly, although *yku70*Δ cells with the *GAL1prm-GFP.ZFN2(CO)* expression cassette grew normally, we experienced consistent difficulty in construction of a viable wild-type strain with the same *GAL1prm-GFP.ZFN2(CO)* expression cassette (data not shown). This is probably because wild-type cells could slowly develop a growth defect by accumulating random mutations in the genome after rejoining of the DSBs cleaved non-specifically by leaky ZFN expression. *yku70*Δ cells with DSBs cannot survive, so the viable *yku70*Δ cells are free of mutations and grow normally. To overcome this problem, we replaced the native *GAL1* promoter in GFP.ZFN2(CO) expressing strain with a mutagenic *GAL1* promoter (*GAL1-V10prm*) created and characterized by Dr. Lewis's group. (Lewis *et al.*, 2005), which showed greatly reduced leaky expression activity in the present of glucose and comparable induction activity to the wild-type *GAL1* promoter in response to galactose. With the use of the *GAL1-V10prm*, we were able to construct and maintain viable wild-type strains with GFP.ZFN2(CO) coding sequence, which grew at the same rate to a control strain (data not shown). Importantly, the *GAL1-V10prm* not only ensured the feasibility of constructing our strains, but also prevented pre-incorporation of mutations in the ZFNcs before intentional induction of the DSBs.

Considering the designs of our assays below which demand rapid generation of site-specific DSBs and intact cut sites before DSB induction, we chose the *GAL1-V10prm-*

GFP.ZFN2(CO) expression cassette as the final optimized ZFN expressing system in this study. For simplicity, unless otherwise specified, the genotype labeled as 'HO' represents chromosomal expression of HO endonuclease with the wild-type *GAL1* promoter (Chr[*GAL1prm::HO*]) and the genotype labeled as 'ZFN' represents chromosomal expression of the GFP.ZFN2(CO) endonuclease with the *GAL1-V10* promoter (Chr[*GAL1-V10prm::GFP.ZFN2(CO)*]), both bearing their cognate cut sites in the same *ILV1* promoter locus (**Figure 5-1A**).

Notably, ZFNcs used in this study, unless specified otherwise, cut at the same sequence as BamHI (5'-GGAATC). We restricted the ZFNcs to be 6-bp (a length that most commonly used by others) to generate a defined overhang sequence (5'-GATC) (**Figure 5-1B**), without the concern of the potential, although inefficient, alternative cleavage by ZFN which may create secondary overhang sequences in a longer cut site (Cradick *et al.*, 2011).

To better compare our HO and ZFN systems, we next monitored the formation of site-specific DSBs by qPCR. Normalization of qPCR signal of primers flanking the cut site by qPCR signal to primers amplifying a control allele can reveal the percentage of the cells that have the intact or rejoined cut sites (given that one cell has only one cut site) (**Figure 5-1A**). We call this the 'DSB-monitoring assay'. Because we collect and froze real-time samples from a time-course experiment, the actual information of the DSBs can be preserved and analyzed without the need of cell growth. This overcomes the complication that a survival-based assay may cause, especially when studying the ZFN.

End-joining defective strains with *yku80*Δ (similar to *yku70*Δ discussed above) showed that our optimized ZFN can rapidly induce DSBs in about 70% of cells during a one-hour window stating from 60 minutes in continuous galactose induction, and can induce DSBs in

almost all the cells within the first 3 hours (**Figure 5-3A**). Our optimized ZFN system showed a great improvement of the cleaving activity from the previously reported ZFN and TALEN systems (Christian *et al.*, 2010; Cermak *et al.*, 2011), allowing kinetic study of chromosomal 5' DSBs. Our optimized ZFN system also showed comparable efficiency of DSB induction to our HO system, although HO was still more active (**Figure 5-3A**). The reason that there was an approximately one-hour delay (0-60 min) before rapid cleavage by ZFN might reflect slower kinetics of ZFN in protein folding and/or binding-site recognition in yeast, as compared to the native endonuclease HO (**Figure 5-3A**). Nonetheless, the efficiency of our HO and ZFN systems reported here provides improved experimental conditions for comparative study of chromosomal 3' and 5' DSBs.

ZFN-mediated 5' DSBs are rejoined more rapidly than HO-mediated 3' DSBs. First, we tested whether overhang polarities of chromosomal DSBs will affect NHEJ efficiency. We compared the DSB-monitoring profile of the wild-type strain and the end-joining defective *dnl4*Δ strain lacking the essential DNA ligase (LigIV) for yeast NHEJ, after a short period of galactose induction. The reason that we only induced endonuclease expression for short period is to help avoid extensive recleavage of the cut sites after precise rejoining, which might affect the estimate of the repair efficiency. DSB-monitoring profiles between the wild-type and *yku80*Δ strains with HO expression overlapped, suggesting the recleavage is more profound for HO (**Figure 5-3A**). Because of that, we intentionally induced the endonuclease for a short period of time (5 minutes for HO and 30 minutes for ZFN, respectively), then changed the medium to glucose containing YPAD to turn off further expression. Under these conditions, the maximum percentage of *dnl4*Δ cells with broken cut sites reached and maintained at about 60-70% for both HO and ZFN strains

(**Figure 5-3B**, **C**). This suggested that only a small amount of endonucleases had been expressed and would not support extensive recleavage of the precisely rejoined cut sites, especially for HO which has a short 10-minute half-life for degradation (Kaplun *et al.*, 2000; Kaplun *et al.*, 2003). This ensures the difference of the DSB-monitoring profiles between the wild-type and *dnl4*Δ cells (highlighted by the red arrow in **Figure 5-3B**, **C**) in the post-cleavage time period (i.e. 60-240 min for HO and 120-240 min for ZFN) can represent the NHEJ efficiency. Our results suggest that rejoining of the 5' breaks might be more efficient than the 3' breaks, and certainly not less efficient.

More NHEJ factors, with higher efficiency, are recruited to ZFN-mediated 5' DSBs than to HO-mediated 3' DSBs. With the efficient HO systems, several labs including ours have reported findings of the recruitment of yeast DNA repair factors to chromosomal 3' DSBs using chromatin immunoprecipitation (ChIP) (Zhang *et al.*, 2007; Chen *et al.*, 2011; Chiruvella *et al.*, 2013a) . With the efficient ZFN system reported here, we are now able study the recruitment of NHEJ factors to chromosomal 5' DSBs.

To test whether overhang polarity can influence the recruitment of end-joining factors, cut sites of our HO and ZFN systems reported here are placed in the same nucleosome–free locus in the *IVL1* promoter to ensure they have the same chromatin structures prior to induction. We monitored recruitment of NHEJ factors by our previous established ChIP protocol (Wu *et al.*, 2008) under continuous galactose induction in order to generate substantial amount of DSBs for the less active ZFN (**Figure 5-3A**). If overhang polarity does not alter recruitment of end-joining factors, we would expect to see higher ChIP signal in the HO strains, simply because there are more HO expressing cells carrying the induced breaks than ZFN expressing cells over a large

part of the tested time course (**Figure 5-3A**). Strikingly, we observed just the opposite, as shown in **Figure 5-4A-D**, with a much higher ChIP signal of the tested NHEJ factors to the ZFN-mediated 5' DSBs as compared to HO-mediated 3' DSBs. The saturated enrichment of recruited NHEJ factors (as shown in the later time points) to the 5' DSBs, as compared to 3' DSBs, was approximately 5-fold for Yku80 and Pol4, and 15-fold for Dnl4 and Xrs2, while Xrs2 had the tendency to further increase its recruitment after 240 minutes (**Figure 5-4A-D**). Some of the fold enrichment numbers that we observed in the ZFN strain (e.g. those of Dnl4 and Xrs2) were much higher than what we and others have ever reported with HO strains. Therefore, we were skeptical of this intriguing, but somewhat surprising observation and decided to carry out various control experiments to validate our finding.

First of all, to rule out the stronger recruitment of NHEJ factors in the ZFN strains may results from NHEJ factors' protein-protein interaction with the ZFN, we repeated our ChIP assay using a catalytic inactive ZFN mutant with a point mutation D450A in its FokI DNA cleavage domain. The D450A mutant binds to its DNA recognition site normally, but fails to dimerize two FolK domains and therefore cannot cleave its cut site (Bitinaite *et al.*, 1998; Waugh and Sauer, 1993). We could not detect any ChIP signal for any of the four NHEJ factors when using the D450A mutant (**Figure 5-4A-D**), suggesting the recruitment of the NHEJ factors are specific to the induced breaks. In addition, we did not detect any ChIP signal of Dnl4 without the use of the anti-Myc antibody (**Figure 5-4C**), validating our ChIP protocol. Next, we repeated ChIP of Pol4 and Dnl4 using strains with *yku80*Δ. As expected, absence of the Ku complex abolished the recruitment of both NHEJ factors to the 5' DSBs, confirming that the factors were recruited to the 5' DSBs as part of the NHEJ complex which relies on the Ku heterodimer, and not from non-

specific protein aggregation. Furthermore, the ChIP signal of the down-stream 5' resection exonuclease Exo1 were similar between 5' and 3' DSBs (**Figure 5-4E**), again supporting the notion that the higher recruitment of repair factors were specific to those participating in NHEJ and was not artifact. Consistent with the widely observed competition between the Ku complex and Exo1, and the MRX complex, deletion of *EXO1* or *RAD50* increased the amount YKu80 to the both 3' and 5' DSBs (**Figure 5-5B**, **C**), suggesting DSBs induced by both endonucleases are bona fide substrates for the repair factors. Finally, although we don't think there will be substantial non-specific cleavage of our ZFN in the initial hours after the induction, there is a small possibility that such non-specific cleavage of the ZFN in other places of the genome may trigger a more acute damage response which hyper activates the NHEJ activity for the DSB in the *ILV1* promoter. To address this concern, we transformed the HO strain with either a plasmid carrying the ZFN expression cassette or an empty vector. Upon galactose induction, this strain can co-express HO and ZFN, while there is only cut site for HO to cleave. So if our observed higher recruitment of NHEJ factors did come from a secondary effect of the potential non-specific cleavage by the ZFN, we would expect to see similar up-regulated recruitment in the strain carrying the ZFN plasmid. However, we did not observed such an increase when we carried out the ChIP for Dnl4 **(Figure 5-5A)**. Results from these control experiments strongly oppose the idea that the higher recruitment of NHEJ factors to 5' DSBs comes from secondary effects of the different endonucleases used. Therefore, the results of our ChIP assay provide evidence that more NHEJ factors are recruited to the chromosomal 5' DSBs than to 3' DSBs, consistent with the higher NHEJ efficiency of 5' DSB in the DSB-monitoring assay. In addition, the fact that we can already detect similar or even slightly higher ChIP signal in the 60-min time

point (**Figure 5-4A-D**) when there is only a small fraction of the DSBs being induced in ZFN expressing strains as compared to full DSB induction in HO expressing strains (**Figure 5-3A**), further suggested that the recruitment of NHEJ factors to the 5' DSBs was more efficient than to the 3' DSBs.

Yeast Tdp1 is recruited to ZFN-mediated 5', but not HO-mediated 3', DSBs and its recruitment is restricted by Ku. The tyrosyl-DNA phosphodiesterase (Tdp1) is a general 3' phosphoesterase capable of removing lesions such as stalled topoisomerases and 3' phosphoglycolates (Zhou *et al.*, 2009). Loss of yeast Tdp1 was shown to increase the frequency of insertional mutations of NHEJ in 5' DSBs, but in 3' DSBs, using recirculation assay of linearized plasmids transformed in *S. cerevisiae* (Bahmed *et al.*, 2010). A potential role of Tdp1was proposed in controlling the repair fidelity of 5' DSBs in a mechanism by which Tdp1 removes the terminal nucleoside from the recessive 3' end of a 5' DSB, leaving a 3' phosphoryl group which can temporarily inhibit undesirable filling of the overhang by polymerases prior to (Bahmed *et al.*, 2010). With our ZFN system, we decided to examine the proposed role of Tdp1 in chromosomal 5' DSBs.

First we examined the recruitment of Tdp1 to the induced 3' DSBs and 5' DSBs by ChIP. Notably, ChIP results of Tdp1 to site-specific DSBs has not been reported prior to our study. This may partially come from the fact that Tdp1 does not recruit to the 3' DSBs induced by the widely used HO (**Figure 5-4F**). In contrast, we were able to detect small yet significant Tdp1 ChIP signal above the background to the ZFN-mediated 5' DSBs (**Figure 5-4F**). Remarkably, deletion of *YKU80* significantly increased the ChIP signal of Tdp1 to the 5' DSBs, while the ChIP signal of Tdp1 to the 3' DSBs remained in the background level (**Figure 5-4F**). This

results demonstrated that yeast Tdp1 is recruited exclusively to 5' DSBs and its recruitment is restricted by the Ku or the NHEJ complex. In addition, deletion of *TDP1* slightly increased the recruitment of Yku80 and Pol4 to 5' DSBs, but had no effect on the recruitment of Yku80 to 3' DSBs (**Figure 5-4A**, **B**). Conversely, we also observed that overexpression of Tdp1 can slightly impede end joining (**Figure 5-6A**). We further investigated whether Tdp1 can interact with the NHEJ complex by yeast two-hybrid using full-length Tdp1 as bait against our previously published NHEJ two-hybrid array which contains full-length and important domains of all major yeast NHEJ factors (Palmbos *et al.*, 2005; Palmbos *et al.*, 2008). We did not detect any protein-protein interaction between Tdp1 and other components of the NHEJ complex with the caveat that a positive interacting control for our full-length Tdp1 was not available from the current literatures (**Figure 5-6B**). In summary, our results suggests that yeast Tdp1 is exclusively recruited to chromosomal 5' DSBs, but not to 3' DSBs, and it competes, but not interacts, with Ku and presumably the NHEJ complex.

High-resolution detection of break-end modifications by ligation-mediated qPCR (LM-qPCR). One of the most important questions in NHEJ is whether and how the break-ends are processed prior to the religation. To examine whether yeast Tdp1 can serve as a nucleosidase in removing the 3' nucleoside of a chromosomal 5' DSB, a novel activity suggested by the previous study using plasmid-based and biochemical assays (Bahmed *et al.*, 2010), we developed a high-resolution approach to detect break-end modifications. The idea is that break-ends modified by DNA processing enzymes, such as various exonuclease and polymerases, often have gain, loss or chemical modifications of the terminal nucleotides, which will greatly reduce the ligation efficiency to an artificial linker with complementary overhang sequence to the

unmodified break-end. Therefore, we can interrogate the presence or absence of a certain overhang structure with known or predicted overhang sequence, polarity and length, by analyzing the ligation efficiency to a designed linker. **Figure 5-7A** shows an example in our experimental design to detect end-processing of a ZFN-mediated chromosomal 5' DSB. After ZFN induction, genomic DNA, extracted from frozen real-time samples, will harbor the induced 5' DSBs with a 4-nt overhang (5'-GATC) and a ligatable 3' hydroxyl in the recessive strand. An artificial linker, carrying a 4-nt 5'overhang with complementary sequence (5'-GATC) and terminal 5' phosphate, can ligate to the unmodified ZFN-mediated DSBs. In order to interrogate the modifications specific to recessive strand (one that with 3' hydroxyl) , we intentionally placed a terminal 3' Dideoxycytidine (3' ddC) in the upper (shorter) strand of the linker which can completely block DNA ligation as well as polymerization at this 3' end. With this design, ligation occurring in the lower strands between the linker and the genomic DNA can be detected by qPCR with primers in the configuration shown in **Figure 5-7A**. We called this assay ligation-mediated qPCR (or LM-qPCR).

The proposed nucleosidase activity of Tdp1 suggests that it can remove the 3' nucleoside of the chromosomal 5' DSB, leaving a terminal 3' phosphoryl (**Figure 5-7B**). To detect this hypothetical intermediate by Tdp1, we designed a similar artificial linker with a 5-nt complementary overhang which can ligate to the Tdp1 processed 5' DSBs after removal of the 3' phosphoryl by T4 PNK (**Figure 5-7B**). To validate our experimental approach, we synthesized two DNA duplexes with the same end structure as the hypothetical Tdp1 intermediate, with either 3' phosphoryl or hydroxyl group in the recessive strand (**Figure 5-8A**). As expected, linkers with 5-nt complementary overhangs can efficiently ligate to the synthetic duplex with 3'

hydroxyl but not to the duplex with 3’ phosphoryl. Importantly, duplex with 3’ phosphoryl showed similar ligation efficiency if incubated with T4 PNK prior to the ligation, confirming the feasibility of our experimental design for detection of the hypothetical Tdp1 intermediate in chromosomal 5’ DSBs (**Figure 5-8A**). As control, we did not detect significant ligation using linker with 4-nt overhang, or any ligation without the T4 ligase or linker, confirming the specificity of our assay.

After validation of our assay, we subjected genomic DNA, extracted from strains with various genome types including overexpression of Tdp1 and deletion of the 3’ phosphatase Tpp1, to the ligation reactions (**Figure 5-7C**). We induced the ZFN expression for 2 hours in order to allow enough time for break induction and potential Tdp1 end processing to occur. We did not detect the presence of hypothetical Tdp1 intermediates as the efficiency of ligations with 5-nt overhang was the same between the T4 PNK and mock treatment groups (**Figure 5-7C**). The fact that the efficiency of ligations with 5-nt was even lower than the background level we got using a linearized plasmid with BamHI (**Figure 5-8B**), further suggested that the hypothetical Tdp1 intermediates, either do not form in a substantial amount or are short-lived.

ZFN-mediated 5’ DSBs are processed more efficiently than HO-mediated 3’ DSBs. We next investigated the overall efficiency of end-processing in both ZFN-mediated 5’ DSBs and HO-mediated 3’ DSBs using the LM-qPCR assay we developed. We monitored the real-time progression of the end-processing by collecting time-course samples after endonuclease induction. Linkers with either 5’ or 3’ 4-nt overhangs were ligated to genomic DNA samples to detect unmodified break-ends from cleaved ZFNcs and HOcs, respectively (**Figure 5-7A** and **Figure 5-9A**). Ligation efficiency, which represents the percentage of unmodified break-ends,

was calculated by normalizing qPCR signal of the ligation event, to (i) the percentage of cells with the broken cut site (generated by the DSB-monitoring assay discussed above), (ii) ligation efficiency of linkers to restriction enzyme linearized plasmids mixed with genomic DNA (e.g. in **Figure 5-8B**) and (iii) other necessary normalizations of the qPCR condition used. Most of the ZFN-mediated 5 DSBs in wild-type cells were processed rapidly, showing about 25% unmodified break-ends at around 120 minutes after the induction (**Figure 5-7D**). As expected, absence of the end-protective Ku complex in *yku80*Δ cells led to massive end-processing, while absence of the nuclease complex MRX in *rad50*Δ cells increased the portion of the unmodified break-ends of the chromosomal 5' DSBs (**Figure 5-7D**). This is consistent with our observation that Xrs2 (part of the MRX complex) is heavily recruited to the 5' DSBs in our ChIP assay (**Figure 5-4D**).

Surprisingly, under the same continuous galactose induction, HO-mediated 3' DSBs were much more resistant to end-processing, with about 80% unmodified break-ends in early time points in wild-type, *rad50*Δ and even *yku80*Δ cells (**Figure 5-9B-D**). We previously reported that our HO-expressing strains have a much delayed progression of 5' resection due to the lack of the functional *GAL1* gene which was interrupted by the HO coding sequence (Chiruvella *et al.*, 2013b; Chiruvella *et al.*, 2014), and that a change to glucose-containing medium after the induction can strongly stimulate 5'resection (Chiruvella *et al.*, 2013a). We thus wondered if progression of the end-processing in HO-mediated 3' DSBs shared the same pattern as the 5' resection. In fact, that was what we observed when we changed the medium to YPAD after 30 minutes of induction (**Figure 5-9B-D**). To conclude, our results suggest that ZFN-mediated 5' DSBs are processed more efficiently than HO-mediated 3' DSBs.

High-throughput kinetic analysis of NHEJ joint types by quantitative next-generation sequencing. Although end-processing of break-ends is not necessary to give rise to mutations and often promotes precise rejoining, NHEJ mutagenesis comes from end-processing. Therefore, analysis of the end-processing can provide information about the origins of mutations and analysis of joint types, especially those with mutations, can infer end-processing. So we next wanted to study the repair joint types of both 3' and 5' DSBs. However, limited by the nature of our ZFN system discussed above, extensive induction of ZFN is lethal. In order to study the joint types of ZFN-mediated 5' DSBs, we decided to subject genomic DNA extracted from real-time samples to next-generation sequencing (NGS). This not only alleviates the need to have viable yeast cells after the ZFN induction, but also greatly expands our ability to analyze joint types in high-throughput. **Figure 5-10A** shows the general steps of our NGS protocol. After collection of time-course samples, genomic DNA was extracted. Critically, we used a same pair of primers to amplify the joint allele and the control allele simultaneously. The PCR condition is optimized to have the same efficiency for two PCR amplicons. Therefore, the ratio between the two allele remains unchanged after the PCR. This design enables quantification of the frequency of any joint types after the sequencing. In addition, we labeled amplicons of different time-course samples with unique barcodes which allowed us to pool tens of samples into one sequencing run. Then we subjected the sequencing library to the Illumina 100-bp pair-read sequencing platform and implemented Bioinformatics analysis of the sequencing reads using our own program. Joint types from validated reads are categorized and assigned to a unique joint identifiers as previously described (Chiruvella *et al.*, 2013a). Using our quantitative NGS approach, we were able to generate DSB-monitoring profiles, similar to the ones from our DSB-monitoring assay,

for various HO and ZFN strains (**Figure 5-10B**, **C**). This confirmed the accuracy of our quantitative design in the NGS assay.

Pol4 is strictly required for gap-filling at HO-mediated 3' DSBs, but partially dispensable at ZFN-mediated 5' DSBs. Because joints with insertional mutations require fillings by polymerases, they can therefore be used to examine the involvement of polymerase during the rejoining process. Our results showed that as a specialized end-joining polymerase, yeast Pol4 is the exclusive polymerase to generate insertional mutations in 3' DSBs, while other polymerases can catalyze insertional mutations in 5' DSBs with a lower efficiency (**Figure 5-11A**, **B** and **Figure 5-12A**). These results in the chromosomal break system recapitulated our previous findings using the plasmid-based system (Daley *et al.*, 2005). The reason that there is a differential requirement of Pol4 in gap-fillings of 3' and 5' DSBs may be that a 3' DSB can only be gap-filled after synapsis of two 3' overhangs which will require specialized end-joining polymerase Pol4 for this disrupted template, whereas one 3' recessive end in a 5' DSB can be polymerized independently of the other break-end, and thus can utilize other polymerases.

Consistent with our LM-qPCR results of the Tdp1 intermediates, we did not observe significant difference between wild-type and *tdp1*Δ cells in two independent NGS experiments (**Figure 5-11A**, **B** and **Figure 5-12A**), arguing against a major role of the yeast Tdp1 in NHEJ of chromosomal 5' DSBs. We further validated our NGS results of the Tdp1 by using a plasmid-rejoining assay as previously described (**Materials and Methods**). Our results shows that there is no significant difference between wild-type and *tdp1*Δ in mutation frequency and overall repair efficiency (**Figure 5-6C**, **D**).

ZFN-mediated 5' DSBs yield more frequent mutagenesis than HO-mediated 3' DSBs. Because we observed that 5' DSBs underwent more rapid end-processing than the 3' DSBs in the LM-qPCR assay, we wondered if this would lead to more frequent mutagenesis. To address that, we first analyzed NGS reads of samples collected after 6 hours of continuous endonuclease induction. The reason that we chose this time-point was that it enriched joint types with mutations after extensive recleavage of the precise joint (**Figure 5-11C**). The sequencing reads from these samples can be categorized into three groups: (i) precise joint, (ii) mutagenic joints and (iii) intact cut sites which have never been cleaved. The third type of reads can come from dead cells or cells that have rare mutations in the endonuclease expression cassette. After extensive recleavage, we assume that the majority of the reads with no mutation comes from (iii), while only a small fraction of them are bona fide precise joints. But because (i) and (iii) have identical sequence, they cannot be distinguished in the sequencing result. Nonetheless, mutagenic joints must be products of actual rejoining **(Figure 5-11C**). Consistent with our prediction from our LM-qPCR assay, we observed significantly higher frequency of mutation in ZFN strain as compared to HO strain with the same genotype (**Figure 5-11D**). Similar differences in mutation frequency can also be seen in other samples with short period of galactose induction (i.e. 30 minutes for HO and 60 minutes for ZFN) (**Table 5-1**). It is worth mentioning that unlike HO, the cleavage of the ZFN is not sequence specific in the cut site (6-bp in this study), ZFN has the potential to recleaved joints that have minor mutations (e.g. mutations with few nucleotide deleted or inserted), until a major mutation occurs that abolishes the zinc finger binding sties franking the 6-bp cut site. However, this does not affect the conclusion of our above analysis because once mutations are formed, they would almost never to be converted

back to precise joints even after recleavage. So the potential recleavage of the mutated ZFNcs will not make us overestimate the mutations frequency. Therefore, our analysis indicates that ZFN-mediated 5’ DSBs yield more frequent mutagenesis than HO-mediated 3’ DSBs.

Discussion

The central findings of this study are (i) the development of an efficient system to generate chromosomal 5' DSBs by optimizations of the selected ZFN in yeast, (ii) that ZFN-mediated 5' DSBs are rejoined more rapidly as compared to the HO-mediated 3' DSBs in the same genomic locus, consistent with the more efficient recruitment of NHEJ factors in 5' DSBs (**Figure 5-13**), (iii) the development of new high-resolution and high-throughput experimental approaches such as the LM-qPCR and quantitative NGS assays and using which, (iv) the observations of more robust end-processing and more frequent mutagenesis in ZFN-mediated 5' DSBs as compared to HO-mediated 3' DSBs (**Figure 5-13**).

Induction of chromosomal 5' DSBs by ZFN. All the widely used endonucleases in the field of DSB repair, such as HO and I-SceI, generate 3' DSBs. Although there were recent reports using the novel genome editing nucleases ZFN, TALEN and CRISPR/Cas9 to study genome translocations and modification of chromatin structure upon break induction (Piganeau *et al.*, 2013; Ghezraoui *et al.*, 2014; Ayrapetov *et al.*, 2014), the achieved kinetics of the nucleases in these studies were too slow to be used in real-time monitoring of rapid NHEJ processes. Here we report an efficient system to generate chromosomal 5' DSBs by optimization of selected ZFNs in yeast. Combining the well-established *GAL1* inducible/repressible system, codon optimization and genome integration of the ZFN coding sequence, our ZFN system has rapid cleavage kinetics comparable to the HO system. However, the current ZFN system reported here remains less efficient, although much improved, than the HO system, suggesting room for further optimizations. One of the major difference is that the ZFN cutting occurs in approximately 60 minutes after the induction whereas HO shows cutting in as soon as 15

minutes. We reason that this may results from a longer time the ZFN will take to be properly folded and/or to locate its DNA binding site. A potential way to shorten the response time of ZFN is to pre-express the endonuclease fused with an nuclear transportation signal which will 'store' the fusion endonucleases in the cytoplasm until the activation of the nuclear influx. This might create a system that can induce almost 'instant' chromosomal DSBs. The Mig1 nuclear transportation domain (NTD) represents a good candidate for this purpose (De Vit *et al.*, 1997; DeVit and Johnston, 1999). However, preliminary work using ZFN construct with the replacement of the original nuclear localization signal (NLS) by the Mig1 NTD showed cleavage after galactose induction, but prior to the activation of nuclear transportation by adding the glucose signal (unpublished data). This suggests further engineering of ZFN fusion protein will be needed to achieve this plausible design. Importantly, DSB-monitoring results of the wild-type ZFN strain for a long period of time (up to 6 hours after galactose induction for 2 hours) showed that there was a delay in the percentage increase of cells with rejoined cut sites after turning off the ZFN expression (data not shown), presumably due to the long lasting recleavage by ZFN. This implies that the ZFN, especially after codon optimization, likely has a higher stability (longer half-life) than the rapidly-degraded HO in our yeast cells. A potential approach to overcome this problem is to fuse the ZFN with a degradation signal (degron). This degron can be covered by interacting with a provided ligand (e.g. the Shield 1 ligand of the Clontech ProteoTuner system (Egeler *et al.*, 2011)) to allow efficient induction of DSBs after galactose induction. After enough breaks are induced, removal of the ligand (with changing medium to YPAD) can destabilize the ZFN-degron fusion protein and leads to rapid proteolysis.

Notably, the regular DNA cleavage domain of ZFNs are sequence non-specific, which provides flexibility to create 5' DSBs with any wanted overhang sequence. Taking this advantage, we are currently investigating the influence of overhang sequence on rejoining by using various ZFN cut sites, including one that has the same 4-bp sequence as the HOcs and one that will generate a stable microhomology-mediated annealing of two 5' overhangs. In addition, mutagenesis of the FokI DNA cleavage domain created the ZFNickase which will induce a DNA single-strand break (SSB) instead of a DSB (Kim *et al.*, 2012; Ramirez *et al.*, 2012). Interestingly, replacing the FokI domain with PvuII created the ZFN-PvuII which was shown to induce blunt-ended DSBs in vitro (Schierling *et al.*, 2012). So it should be possible to create an efficient system to induce site-specific chromosomal SSB and blunt-ended DSBs. Unfortunately, a similar Zinc-finger-based fusion nuclease with DNA cleavage domain that can cleave 3' DSBs, is not available. Otherwise, it could be a better choice to use with our ZFN to study overhang polarity. Nonetheless, with careful controls in our experimental designs, we were able to carry out various assays to compare the ZFN and HO-mediated DSBs. Also, we are expanding our study here of a single DSB in the nucleosome-free locus of *ILV1* promoter by placing the cut site in various genomic loci, as well as by placing multiple cut sites of individual endonucleases or of both HO and ZFN to study genome translocations mediated by simultaneous induction of DSBs.

5' DSBs have more rapid rejoining, facilitated by more robust end-processing, but at the cost of more frequent mutagenesis. Our results of the DSB-monitoring assay and LM-qPCR assay suggest that chromosomal 5'DSBs are rejoined and processed more robustly in yeast cells (**Figure 5-3C**,**D**, **Figure 5-7D** and **Figure 5-9B-D**), consistent with results of the ChIP assay showing abundant association of rejoining (e.g. Yku80, Dnl4) and processing factors (e.g.

Pol4, Xrs2) to the 5' DSBs (**Figure 5-4A-D**). Although precise rejoining is still the most favorable outcome in both 3' and 5' DSBs with compatible overhangs. 5' DSBs are processed at a higher rate to explore other joining possibilities. This directly leads to the more frequent mutagenesis we observed in the NGS assay (**Figure 5-11D** and **Table 5-1**). These findings argue the notion that 5' DSB may be a more malicious lesion as compared to 3' DSBs in yeast cells and therefore it needs to be resolved as quickly as possible, even at the cost of mutations. Importantly, analysis of joint types by high-throughput sequencing not only alleviates the need to have viable cells after endonuclease induction, but also allows detection of rare mutations in DSBs with compatible overhangs. It would be interesting to examine if this is a universal effect of end-joining by carrying out similar experiments in other cell types which is possible given that the ZFN was shown to work as a universal nuclease in various organisms and cell types including mice and human cell lines. But further customizations and optimizations will definitely be required to achieve the rapid cutting kinetics of ZFN we reported in yeast. Interestingly, FokI catalytic domain was selected to be the DNA cleaving domain of ZFN and TALEN mainly because of its cut-site non-specific nature. This provides the potential for engineered nucleases to target different loci in the genome. Our findings of the more frequent mutagenesis in the ZFN-mediated 5' DSBs provide new evidence to support the usage of FokI as cleavage domain in genome editing applications.

The role of Tdp1 in 5' DSB processing. With our efficient ZFN system, we were able to reveal the exclusive recruitment of yeast Tdp1 to ZFN-mediated 5' DSBs (**Figure 5-4F**), suggesting that yeast Tdp1 might have a specialized role at 5' DSBs but not at 3' DSBs.

However, our results also shows that yeast Tdp1 is not part of the NHEJ complex, but rather competes with NHEJ (**Figure 5-4F** and **Figure 5-6A**, **B**).

Notably, our study of end-processing focuses on those initial end modifications (with few nucleotide being processed) which directly affect NHEJ, instead of large-scale end-processing by the 5' resection in the HR pathway. We previously characterized that the 5' resection of HO-mediated 3' DSBs is inefficient in our *gal* strains (Chiruvella *et al.*, 2013a; Chiruvella *et al.*, 2014), which provides a specialized condition to study initial modifications such as the proposed Tdp1 intermediates. However, when using our validated LM-qPCR assay **(Figure 5-8A**), we did not detect evidence of the proposed Tdp1 intermediates (**Figure 5-7C**). Consistently, we did not detect significant differences in the frequency of insertional mutations of both chromosomal and plasmid 5' DSBs when comparing wild-type and *tdp1*Δ cells (**Figure 5-11B**, **Figure 5-12** and **Figure 5-6C**, **D**), contradictory to the previously proposed role of Tdp1in controlling repair fidelity of 5' DSBs (Bahmed *et al.*, 2010). The reason for the discrepancy is unclear, and we wonder the electroporation transformation protocol the other group used may create an unusual condition in which the effect of Tdp1 can be seen.

Recognition of overhang polarity by repair factors. Althoug the results of different assays are consistent and indicate that the overhang polarity of chromosomal DSBs can greatly impact the kinetics and fidelity of yeast NHEJ, we have not fully understood the mechanism(s) by which overhang polarity are recognized by the NHEJ factors. Some of the end-processing enzymes, such as Exo1 and Pol4, were shown to have different activity on 3' and 5' DSBs using biochemistry and plasmid-rejoining assays (Cannavo *et al.*, 2013; Mantero *et al.*, 1977; Bahmed *et al.*, 2011; Daley *et al.*, 2005). We were able to confirm that the differential requirement of

Pol4 in chromosomal 3 and 5' DSBs using our NGS assay (**Figure 5-11A**, **B** and **Figure 5-12**). So one potential explanation of our observed difference between 5'DSBs and 3'DSBs may be mediated by the recognition of different end-processing factors. In addition, the Ku heterodimer were shown to have the potential to recognize overhang polarity, although the reported observations are not always consistent among studies using different or even similar experimental approaches (Mimori and Hardin, 1986; Paillard and Strauss, 1991; Blier *et al.*, 1993; Ristic *et al.*, 2003; Clerici *et al.*, 2008; Foster *et al.*, 2011; Balestrini *et al.*, 2013). In addition, the recently discovered lyase activity of human and yeast Ku complex supports the notion that Ku can recognize complex break end structures (Roberts *et al.*, 2010; Strande *et al.*, 2012; Strande *et al.*, 2014).Interestingly, study of the human DNA-PKcs (the C-terminus of human Ku80) shows that it can recognize overhang polarity to regulate its kinase activity (Hammarsten *et al.*, 2000; Pawelczak *et al.*, 2005; Pawelczak and Turchi, 2008; Pawelczak *et al.*, 2011). Yeast Ku80 also has a short C-terminus that was shown to interact with yeast LigIV (Palmbos *et al.*, 2008). Preliminary results of some ongoing work hint that this c-terminus of Yku80 might involve in the recognition of the 5' overhang to mediate some of the reported effect in this study, but it is not the only recognition mechanism involved (unpublished data). In summary, the impact of overhang polarity is likely mediated by several repair factors and a consequence of a complex regulatory network.

Acknowledgments

This chapter is modified from the research manuscript in preparation in which I am the first author.

I would like to thank Thomas E Wilson for his mentorship and help in design of the experiments and bioinformatics analysis of the NGS reads and Sivakumar Nallasivam for his work in construction of a substantial number of yeast strains and his helpful discussion for the NGS assay design.

I would also like to thank Dr. Tzvi Tzfira and Dr. John L Nitiss for their gifted ZFN constructs we screened in **Figure 5-1** and Dr. Kevin Lewis for his gifted *GAL1-V10* promoter we used in our optimized ZFN system.

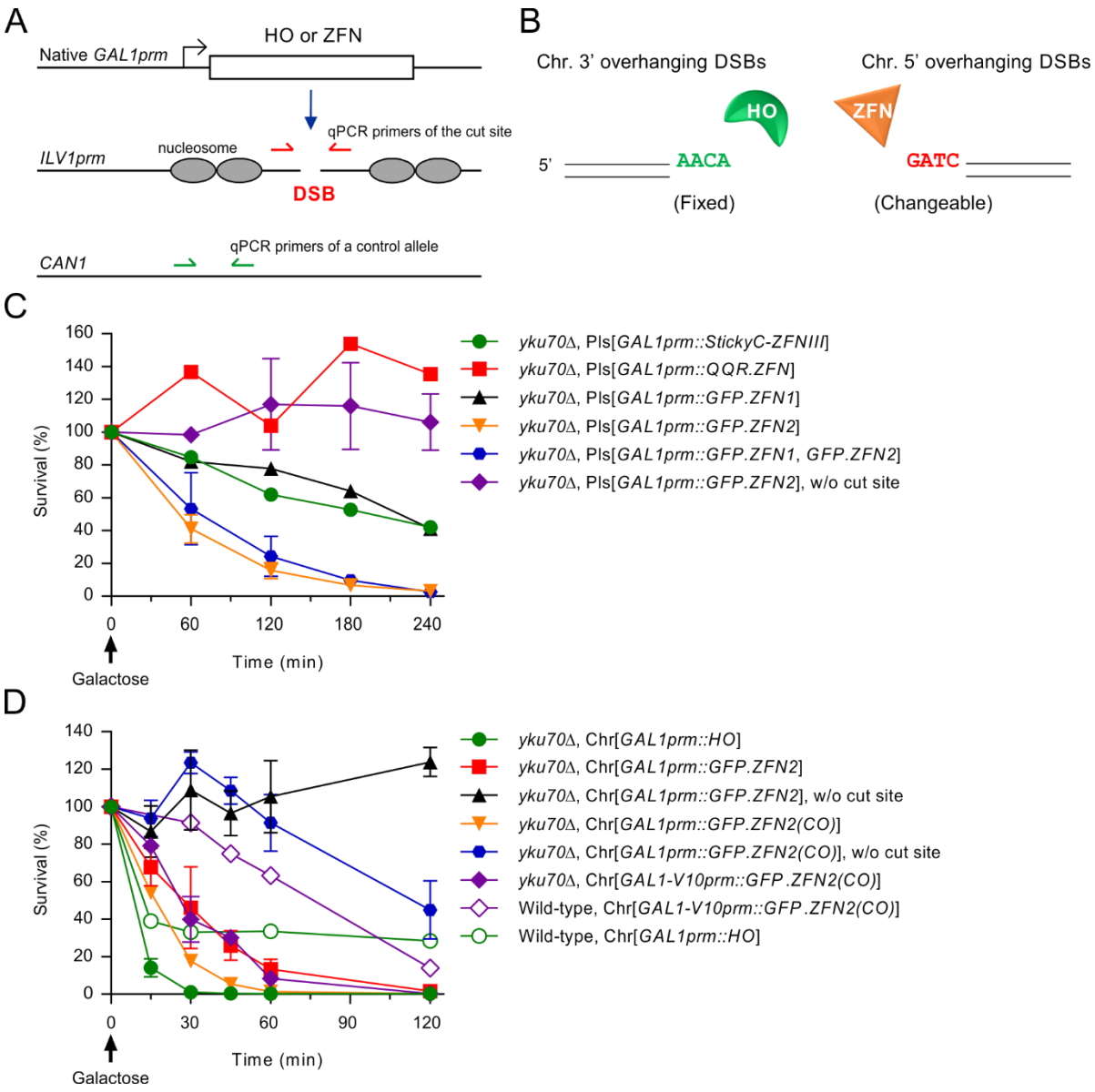


Figure 5-1. Induction of chromosomal 3' and 5' overhanging DSBs by HO and optimized ZFN. (**A**) DSB induction paradigm of HO and ZFN. Coding sequences of both endonucleases are placed down-stream of the native *GAL1* promoter. Induced endonucleases will cleave their cut sites in the same genomic locus of the *ILV1* promoter. Two pairs of qCPR primers for the DSB-monitoring assay, amplifying respectively the cut site and a control allele in the *CAN1* locus, are labeled with arrows. (**B**) Diagram of chromosomal 3' and 5' DSB generated by HO and ZFN, showing overhang polarity and sequence. (**C**, **D**) Survival assay of endonuclease expressing strains after galactose induction. Pls: Plasmid, Chr: Chromosomal.

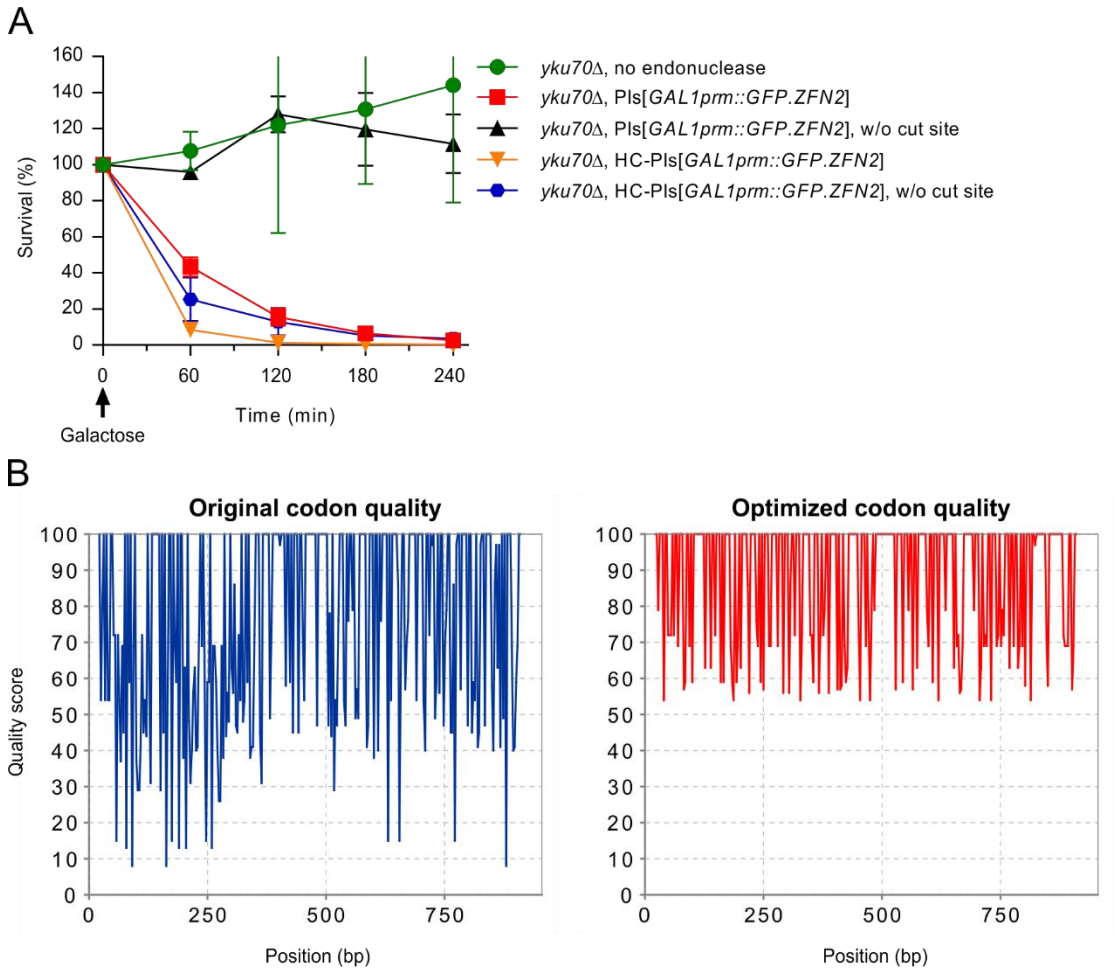


Figure 5-2. Optimization of GFP. ZFN2 cleaving activity. (**A**)Non-specific cleavage of GFP.ZFN2 when hyper-expressed from a high-copy-number plasmid. HC-Pls: High-copy-number plasmid. (**B**) Codon quality comparison of the original GFP.ZFN2 coding sequence (left) and the codon-optimized GFP.ZFN2(CO) coding sequence (right).

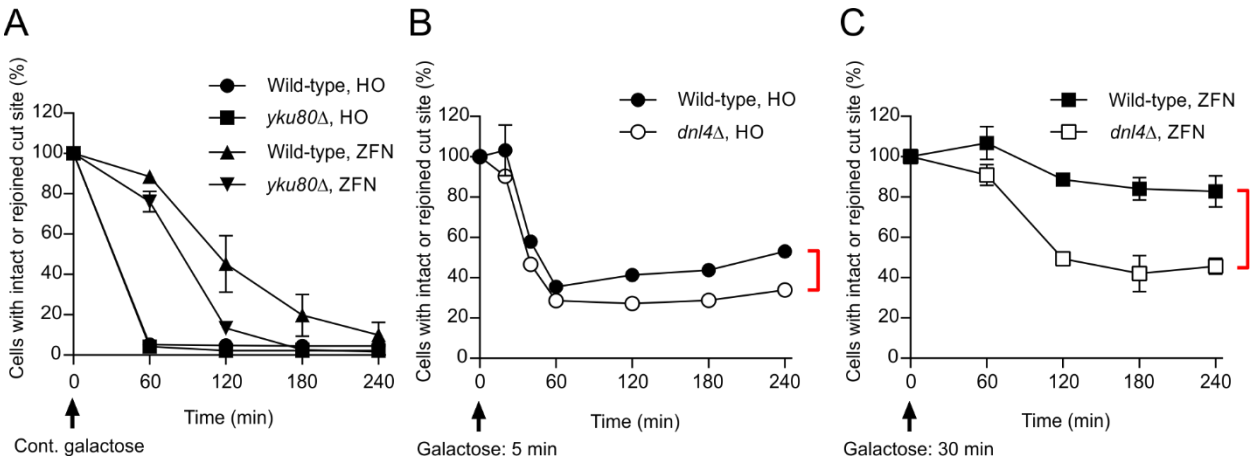


Figure 5-3. Formation and rejoining of induced DSBs monitored by qPCR. (**A**) DSB-monitoring of strains under continuous galactose induction. (**B**) DSB-monitoring of HO strains with 5-miniue galactose induction before changing the medium to YPAD. (**C**) DSB-monitoring of ZFN strains with 30-miniue galactose induction before changing the medium to YPAD.

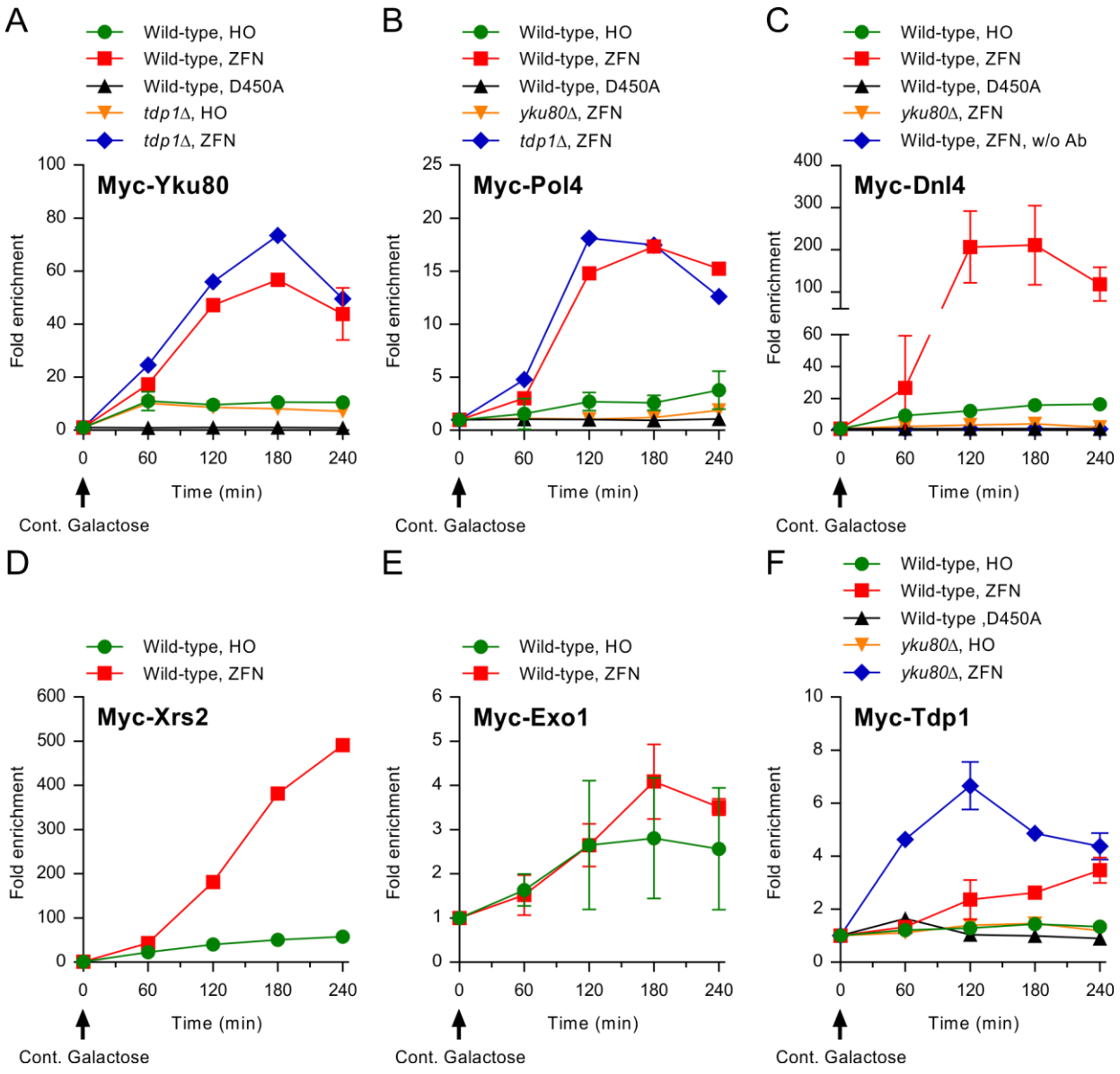


Figure 5-4. Recruitment of DNA repair factors to the induced DSBs monitored by chromatin immunoprecipitation (ChIP). (**A**-**F**) ChIP results of c-terminal 13Myc-tagged (A) Yku80; (B) Pol4; (C) Dnl4; (D) Xrs2; (E) Exo1; (F) Tdp1, under continuous galactose induction.

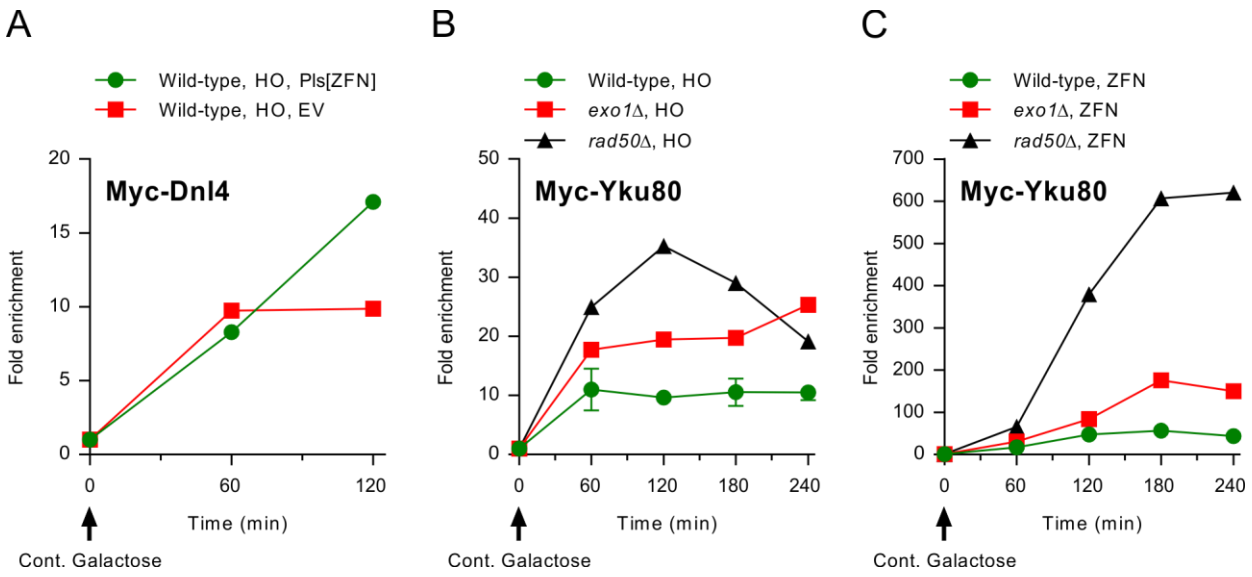


Figure 5-5. Additional ChIP results. (**A**) ChIP of c-terminal 13Myc-tagged Dnl4 with co-expression of HO and ZFN in the cells. Pls: Plasmid, EV: empty vector.

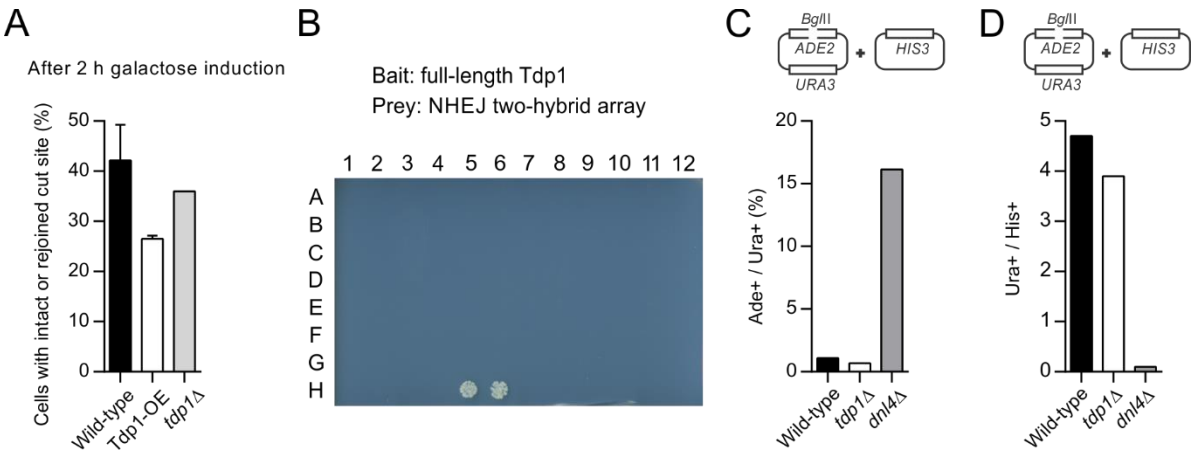


Figure 5-6. Involvement of Tdp1 in yeast NHEJ. (**A**) DSB-monitoring results of strains after 2-hour galactose induction. (**B**) Yeast two-hybrid study of protein-protein interactions between full-length Tdp1 (bait) and our published NHEJ two-hybrid array (prey) (Palmbos *et al.*, 2005; Palmbos *et al.*, 2008). Interaction between full-length Dnl4 and full-length Lif1in H5 and H6 wells serves as positive control for our assay. (**C**, **D**) Results of plasmid-rejoining assay with linearized plasmid carrying a *Bgl*II digested 5' DSB in the *ADE2* gene coding sequence.

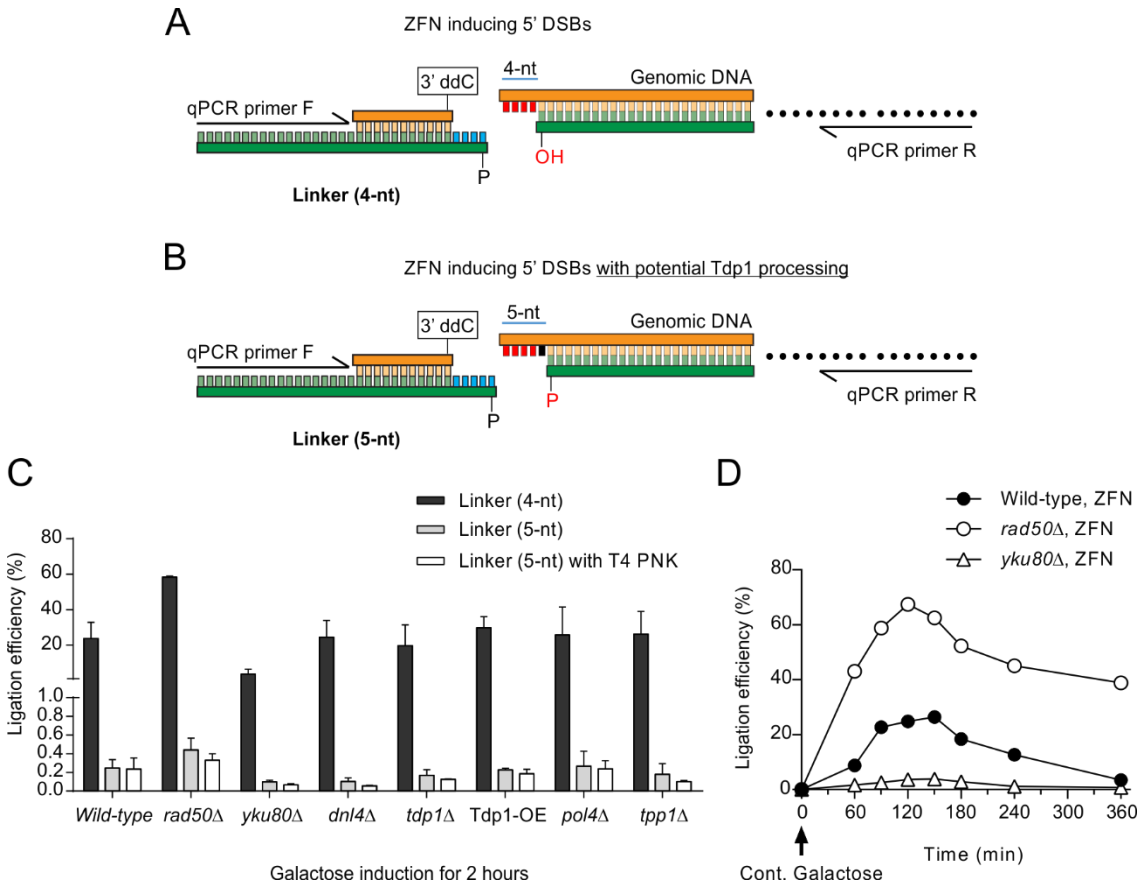


Figure 5-7. Detection of break-end modifications by ligation-mediated qPCR (LM-qPCR). (**A**) Diagram of LM-qPCR design to detect unmodified break-ends of ZFN-mediated 5'DSBs. (**B**) Diagram of LM-qPCR design to detect potential processed break-ends by Tdp1. (**C**) LM-qPCR results of ZFN-mediated 5'DSBs in various genotypes using different linkers and conditions. (**D**) Kinetics of end-processing ZFN-mediated 5'DSBs under continuous galactose induction.

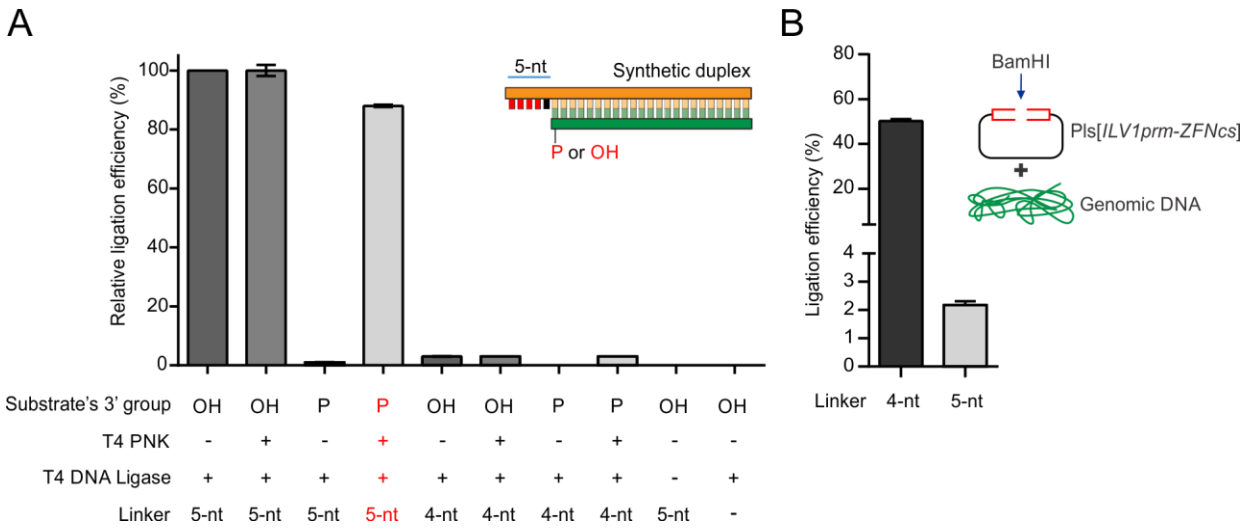


Figure 5-8. Control ligations to validate the design of the LM-qPCR assay. (**A**) Validation of the linkers and T4 PNK treatment used in LM-qPCR assay to detect Tdp1 intermediates (**Figure 5-7C**). (**B**) Validation of the ligation efficiency of linkers using restriction enzyme linearized plasmid mixing with genomic DNA.

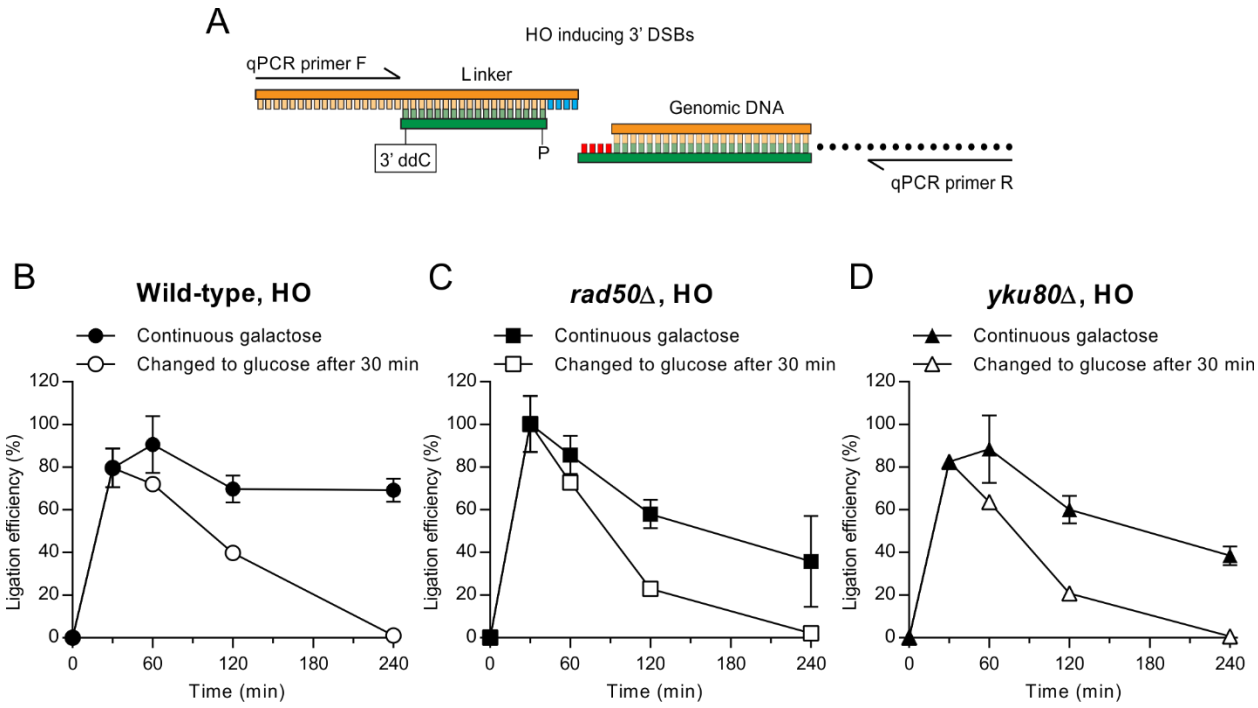


Figure 5-9. End-processing of HO-mediated 3' DSBs monitored by LM-qPCR. (**A**) Diagram of LM-qPCR design to detect unmodified break-ends of HO-mediated 3'DSBs. (**B-D**) Kinetics of end-processing of HO-mediated 3' DSBs in (B) wild-type; (C) *rad50*Δ; (D) *yku80*Δ cells with or without changing the medium to YPAD.

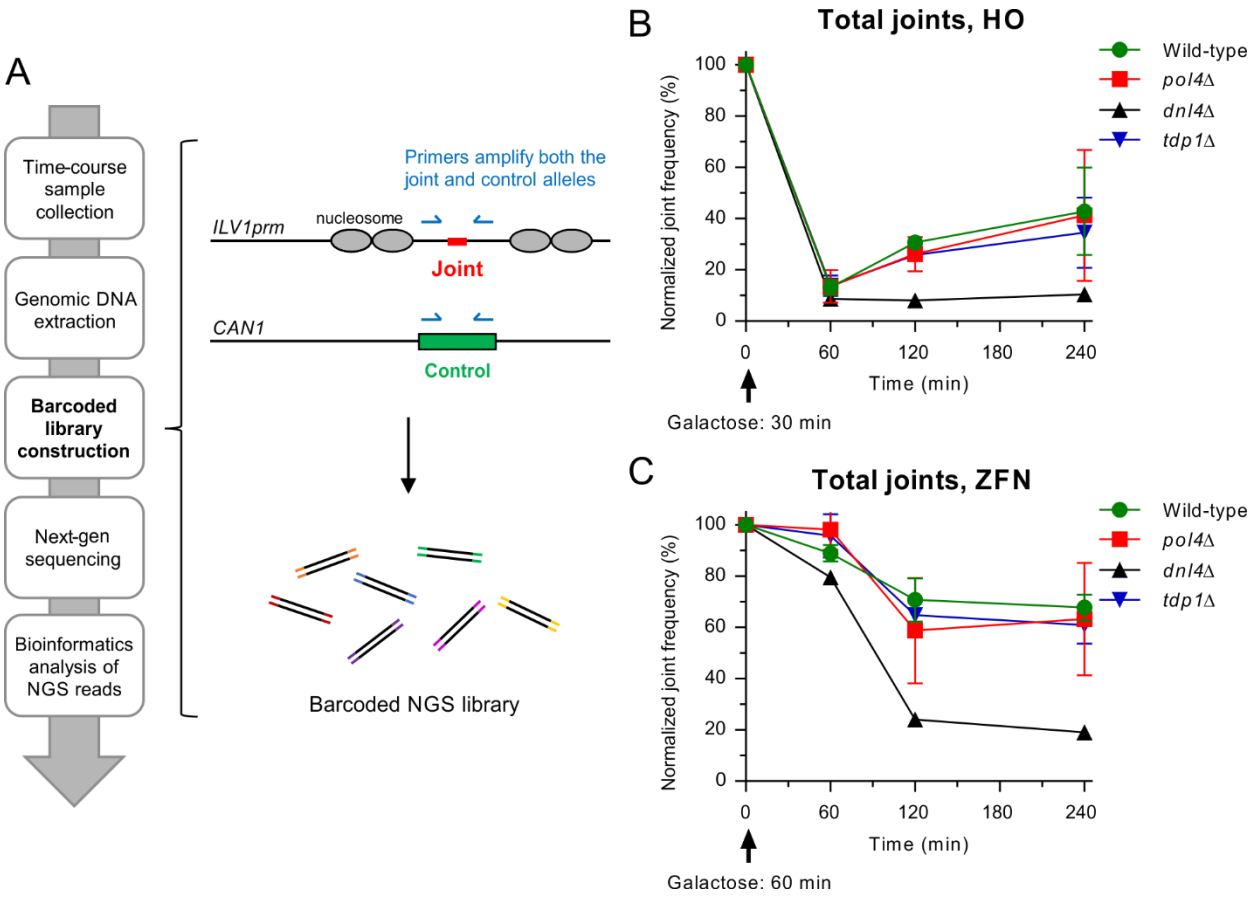


Figure 5-10. High-throughput kinetic analysis of NHEJ joint types by quantitative next-generation sequencing (NGS). (**A**) Diagram of the general steps of our NGS protocol. (**B**) Joint frequency of all joint types in HO expressing strains. (**C**) Joint frequency of all joint types in ZFN expressing strains.

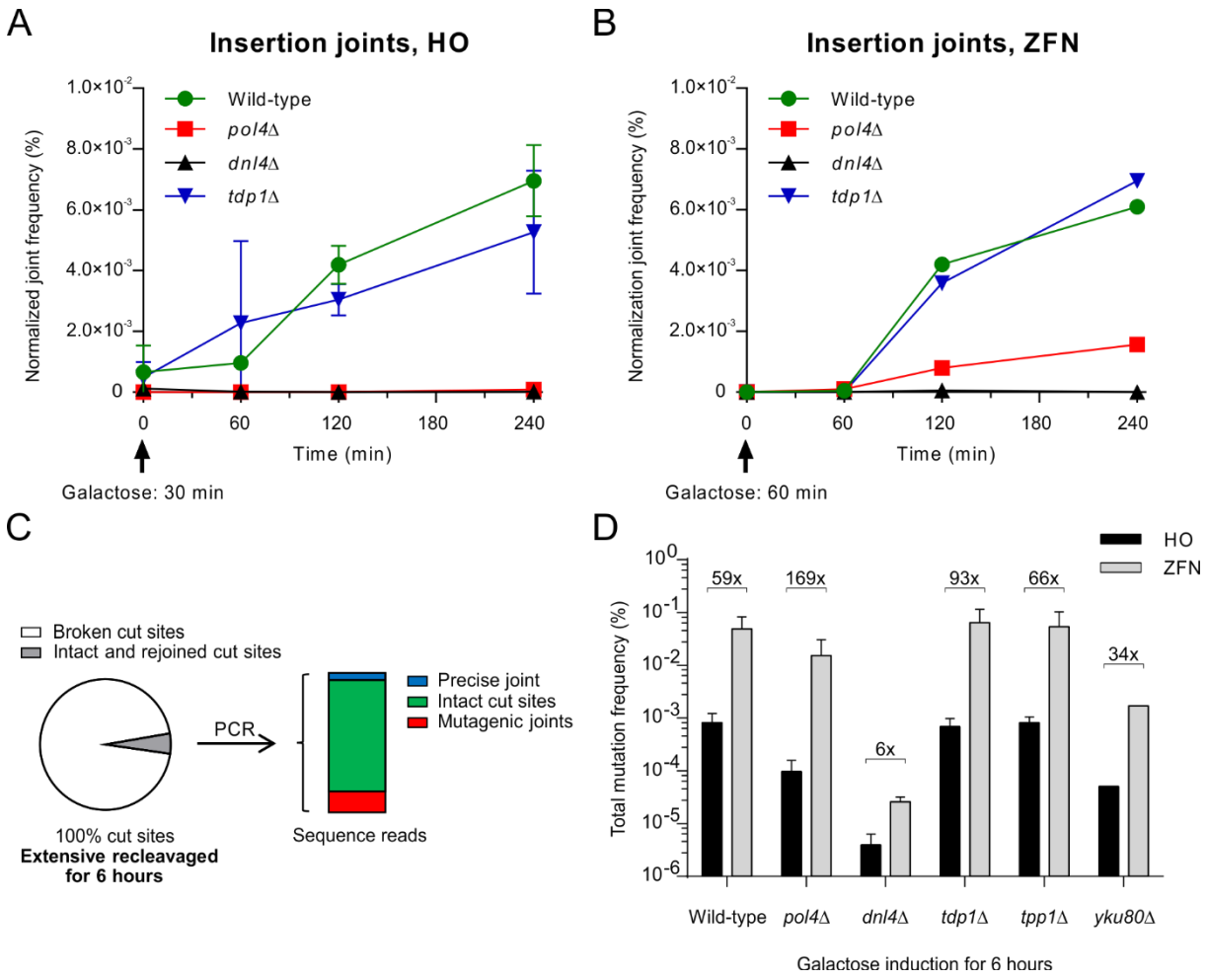


Figure 5-11. Mutagenesis analysis of NHEJ using NGS data. (**A**) Joint frequency of joint types with insertional mutation in HO expressing strains. (**B**) Joint frequency of joint types with insertional mutation in ZFN expressing strains. (**C**) Diagram showing the composition of sequencing reads from samples with 6-hour galactose induction. (**D**) Comparison of the overall mutation frequency between HO and ZFN expressing strains with 6-hour galactose induction.

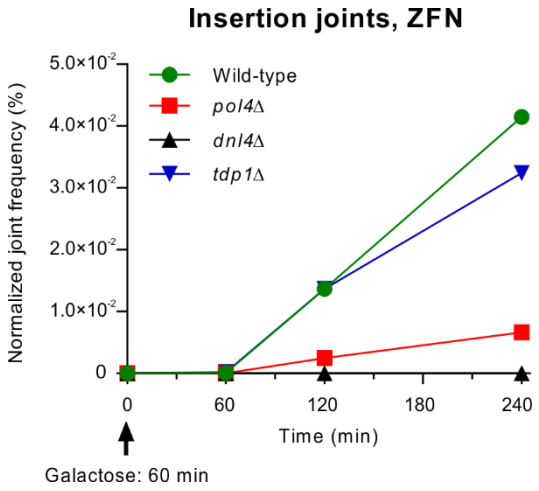


Figure 5-12. Joint frequency of joint types with insertional mutation in ZFN strains. Results from another independent replicate, analyzed in the same way as in **Figure 5-11B**.

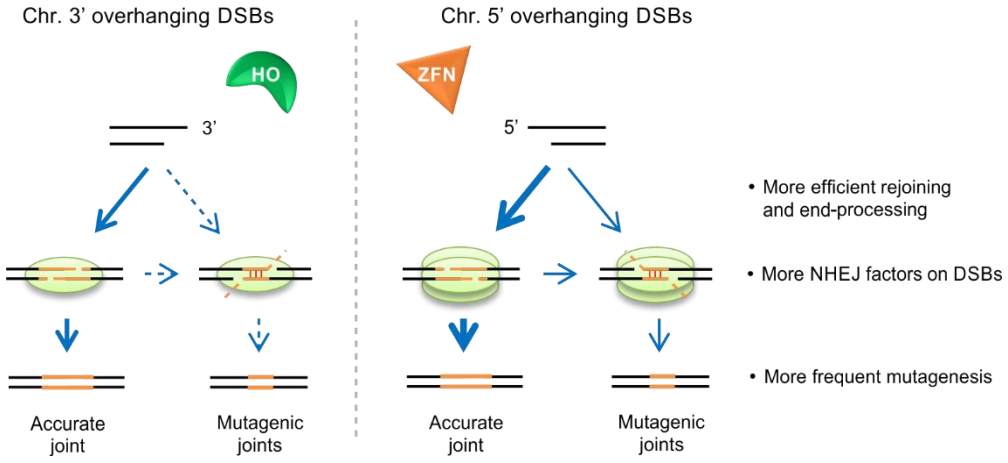


Figure 5-13. Working model summarizing key findings of our study.

Table 5-1. Total mutation frequency after endonuclease induction. For HO strains, medium was changed to YPAD after 30 minutes of galactose induction, expect the time point '360 (Gal)' which was under continuous galactose for 6 hours. For ZFN strains, medium was changed to YPAD after 60 minutes of galactose induction, expect the time point '360 (Gal)' which was under continuous galactose for 6 hours. The total mutation frequency shown here are average of two independent biological replicates Fold differences of the total mutation frequency between ZFN and HO strains are listed in the lower part of the table.

Total mutation frequency					
Time (min)	0	60	120	240	360 (Gal)
Wild-type, HO	4.40E-05	2.80E-05	1.42E-04	2.43E-04	8.24E-04
***pol4*Δ, HO**	2.44E-05	3.90E-06	2.21E-05	7.29E-05	9.54E-05
***dnl4*Δ, HO**	2.57E-05	2.22E-06	6.73E-06	1.25E-05	3.96E-06
***tdp1*Δ, HO**	4.98E-05	6.04E-05	1.01E-04	1.91E-04	6.95E-04
***tpp1*Δ, HO**	2.74E-05	1.28E-05	1.27E-04	2.80E-04	8.21E-04
***yku80*Δ, HO**	2.36E-05	3.07E-06	2.64E-06	8.55E-06	5.09E-05
Wild-type, ZFN	2.00E-05	2.52E-05	9.47E-04	3.83E-03	4.89E-02
***pol4*Δ, ZFN**	4.43E-05	5.49E-05	4.19E-04	1.92E-03	1.55E-02
***dnl4*Δ, ZFN**	2.02E-05	1.96E-05	3.12E-05	5.30E-05	2.57E-05
***tdp1*Δ, ZFN**	6.65E-05	3.84E-05	8.85E-04	3.54E-03	6.47E-02
***tpp1*Δ, ZFN**	3.00E-05	2.94E-05	8.95E-04	4.42E-03	5.43E-02
***yku80*Δ, ZFN**	2.33E-05	2.11E-05	1.96E-04	5.06E-04	1.72E-03

Fold difference (ZFN:HO)					
Time (min)	0	60	120	240	360 (Gal)
Wild-type	0.5	0.9	6.7	15.7	59.4
***pol4*Δ**	1.8	14.1	19.0	26.3	162.8
***dnl4*Δ**	0.8	8.8	4.6	4.2	6.5
***tdp1*Δ**	1.3	0.6	8.8	18.5	93.2
***tpp1*Δ**	1.1	2.3	7.0	15.8	66.1
***yku80*Δ**	1.0	6.9	74.4	59.2	33.8

Reference

Anand, R.P., Lovett, S.T., and Haber, J.E. (2013). Break-induced DNA replication. Cold Spring Harbor perspectives in biology *5*, a010397.

Ayrapetov, M.K., Gursoy-Yuzugullu, O., Xu, C., Xu, Y., and Price, B.D. (2014). DNA double-strand breaks promote methylation of histone H3 on lysine 9 and transient formation of repressive chromatin. Proceedings of the National Academy of Sciences of the United States of America *111*, 9169-9174.

Bahmed, K., Nitiss, K.C., and Nitiss, J.L. (2010). Yeast Tdp1 regulates the fidelity of nonhomologous end joining. Proceedings of the National Academy of Sciences of the United States of America *107*, 4057-4062.

Bahmed, K., Seth, A., Nitiss, K.C., and Nitiss, J.L. (2011). End-processing during non-homologous end-joining: a role for exonuclease 1. Nucleic Acids Res *39*, 970-978.

Balestrini, A., Ristic, D., Dionne, I., Liu, X.Z., Wyman, C., Wellinger, R.J., and Petrini, J.H. (2013). The Ku heterodimer and the metabolism of single-ended DNA double-strand breaks. Cell reports *3*, 2033-2045.

Bitinaite, J., Wah, D.A., Aggarwal, A.K., and Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. Proceedings of the National Academy of Sciences of the United States of America *95*, 10570-10575.

Blier, P.R., Griffith, A.J., Craft, J., and Hardin, J.A. (1993). Binding of Ku protein to DNA. Measurement of affinity for ends and demonstration of binding to nicks. J Biol Chem *268*, 7594-7601.

Brachmann, C.B., Davies, A., Cost, G.J., Caputo, E., Li, J., Hieter, P., and Boeke, J.D. (1998). Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast *14*, 115-132.

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

Cannavo, E., Cejka, P., and Kowalczykowski, S.C. (2013). Relationship of DNA degradation by Saccharomyces cerevisiae exonuclease 1 and its stimulation by RPA and Mre11-Rad50-Xrs2 to DNA end resection. Proceedings of the National Academy of Sciences of the United States of America *110*, E1661-1668.

Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res *39*, e82.

Chen, X., Niu, H., Chung, W.H., Zhu, Z., Papusha, A., Shim, E.Y., Lee, S.E., Sung, P., and Ira, G. (2011). Cell cycle regulation of DNA double-strand break end resection by Cdk1-dependent Dna2 phosphorylation. Nat Struct Mol Biol *18*, 1015-1019.

Chiruvella, K.K., Liang, Z., Birkeland, S.R., Basrur, V., and Wilson, T.E. (2013a). Saccharomyces cerevisiae DNA ligase IV supports imprecise end joining independently of its catalytic activity. PLoS Genet *9*, e1003599.

Chiruvella, K.K., Liang, Z., and Wilson, T.E. (2013b). Repair of double-strand breaks by end joining. Cold Spring Harbor perspectives in biology *5*, a012757.

Chiruvella, K.K., Renard, B.M., Birkeland, S.R., Sunder, S., Liang, Z., and Wilson, T.E. (2014). Yeast DNA ligase IV mutations reveal a nonhomologous end joining function of BRCT1 distinct from XRCC4/Lif1 binding. DNA Repair (Amst) *24*, 37-45.

Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. Genetics *186*, 757-761.

Clerici, M., Mantiero, D., Guerini, I., Lucchini, G., and Longhese, M.P. (2008). The Yku70-Yku80 complex contributes to regulate double-strand break processing and checkpoint activation during the cell cycle. EMBO Rep *9*, 810-818.

Cradick, T.J., Ambrosini, G., Iseli, C., Bucher, P., and McCaffrey, A.P. (2011). ZFN-site searches genomes for zinc finger nuclease target sites and off-target sites. BMC Bioinformatics *12*, 152.

Daley, J.M., Laan, R.L., Suresh, A., and Wilson, T.E. (2005). DNA joint dependence of pol X family polymerase action in nonhomologous end joining. J Biol Chem *280*, 29030-29037.

Daley, J.M., and Wilson, T.E. (2008). Evidence that base stacking potential in annealed 3' overhangs determines polymerase utilization in yeast nonhomologous end joining. DNA Repair (Amst) *7*, 67-76.

De Vit, M.J., Waddle, J.A., and Johnston, M. (1997). Regulated nuclear translocation of the Mig1 glucose repressor. Mol Biol Cell *8*, 1603-1618.

DeVit, M.J., and Johnston, M. (1999). The nuclear exportin Msn5 is required for nuclear export of the Mig1 glucose repressor of Saccharomyces cerevisiae. Curr Biol *9*, 1231-1241.

Egeler, E.L., Urner, L.M., Rakhit, R., Liu, C.W., and Wandless, T.J. (2011). Ligand-switchable substrates for a ubiquitin-proteasome system. J Biol Chem *286*, 31328-31336.

Foster, S.S., Balestrini, A., and Petrini, J.H. (2011). Functional interplay of the Mre11 nuclease and Ku in the response to replication-associated DNA damage. Mol Cell Biol *31*, 4379-4389.

Gaj, T., Gersbach, C.A., and Barbas, C.F., 3rd (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol *31*, 397-405.

Ghezraoui, H., Piganeau, M., Renouf, B., Renaud, J.B., Sallmyr, A., Ruis, B., Oh, S., Tomkinson, A.E., Hendrickson, E.A., Giovannangeli, C., *et al.* (2014). Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. Mol Cell *55*, 829-842.

Ghodke, H., Wang, H., Hsieh, C.L., Woldemeskel, S., Watkins, S.C., Rapic-Otrin, V., and Van Houten, B. (2014). Single-molecule analysis reveals human UV-damaged DNA-binding protein (UV-DDB) dimerizes on DNA via multiple kinetic intermediates. Proceedings of the National Academy of Sciences of the United States of America *111*, E1862-1871.

Hammarsten, O., DeFazio, L.G., and Chu, G. (2000). Activation of DNA-dependent protein kinase by single-stranded DNA ends. J Biol Chem *275*, 1541-1550.

Jasin, M., and Rothstein, R. (2013). Repair of strand breaks by homologous recombination. Cold Spring Harbor perspectives in biology *5*, a012740.

Kaplun, L., Ivantsiv, Y., Bakhrat, A., and Raveh, D. (2003). DNA damage response-mediated degradation of Ho endonuclease via the ubiquitin system involves its nuclear export. J Biol Chem *278*, 48727-48734.

Kaplun, L., Ivantsiv, Y., Kornitzer, D., and Raveh, D. (2000). Functions of the DNA damage response pathway target Ho endonuclease of yeast for degradation via the ubiquitin-26S proteasome system. Proceedings of the National Academy of Sciences of the United States of America *97*, 10077-10082.

Khanna, K.K., and Jackson, S.P. (2001). DNA double-strand breaks: signaling, repair and the cancer connection. Nat Genet *27*, 247-254.

Kim, E., Kim, S., Kim, D.H., Choi, B.S., Choi, I.Y., and Kim, J.S. (2012). Precision genome engineering with programmable DNA-nicking enzymes. Genome Res *22*, 1327-1333.

Lewis, L.K., Lobachev, K., Westmoreland, J.W., Karthikeyan, G., Williamson, K.M., Jordan, J.J., and Resnick, M.A. (2005). Use of a restriction endonuclease cytotoxicity assay to identify inducible GAL1 promoter variants with reduced basal activity. Gene *363*, 183-192.

Lloyd, A., Plaisier, C.L., Carroll, D., and Drews, G.N. (2005). Targeted mutagenesis using zinc-finger nucleases in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America *102*, 2232-2237.

Makharashvili, N., Tubbs, A.T., Yang, S.H., Wang, H., Barton, O., Zhou, Y., Deshpande, R.A., Lee, J.H., Lobrich, M., Sleckman, B.P., *et al.* (2014). Catalytic and noncatalytic roles of the CtIP endonuclease in double-strand break end resection. Mol Cell *54*, 1022-1033.

Mantero, R., Bertolotti, P., and Ferrari, G.L. (1977). Meuli's total prosthesis in severe rheumatoid osteoarthrosis of the wrist. Italian journal of orthopaedics and traumatology *3*, 47-51.

McCormick, S.P., Ng, J.K., Taylor, S., Flynn, L.M., Hammer, R.E., and Young, S.G. (1995). Mutagenesis of the human apolipoprotein B gene in a yeast artificial chromosome reveals the site of attachment for apolipoprotein(a). Proceedings of the National Academy of Sciences of the United States of America *92*, 10147-10151.

Mimori, T., and Hardin, J.A. (1986). Mechanism of interaction between Ku protein and DNA. J Biol Chem *261*, 10375-10379.

Moore, J.K., and Haber, J.E. (1996). Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. Mol Cell Biol *16*, 2164-2173.

Paillard, S., and Strauss, F. (1991). Analysis of the mechanism of interaction of simian Ku protein with DNA. Nucleic Acids Res *19*, 5619-5624.

Palmbos, P.L., Daley, J.M., and Wilson, T.E. (2005). Mutations of the Yku80 C terminus and Xrs2 FHA domain specifically block yeast nonhomologous end joining. Mol Cell Biol *25*, 10782-10790.

Palmbos, P.L., Wu, D., Daley, J.M., and Wilson, T.E. (2008). Recruitment of Saccharomyces cerevisiae Dnl4-Lif1 complex to a double-strand break requires interactions with Yku80 and the Xrs2 FHA domain. Genetics *180*, 1809-1819.

Paull, T.T. (2010). Making the best of the loose ends: Mre11/Rad50 complexes and Sae2 promote DNA double-strand break resection. DNA Repair (Amst) *9*, 1283-1291.

Pawelczak, K.S., Andrews, B.J., and Turchi, J.J. (2005). Differential activation of DNA-PK based on DNA strand orientation and sequence bias. Nucleic Acids Res *33*, 152-161.

Pawelczak, K.S., Bennett, S.M., and Turchi, J.J. (2011). Coordination of DNA-PK activation and nuclease processing of DNA termini in NHEJ. Antioxid Redox Signal *14*, 2531-2543.

Pawelczak, K.S., and Turchi, J.J. (2008). A mechanism for DNA-PK activation requiring unique contributions from each strand of a DNA terminus and implications for microhomology-mediated nonhomologous DNA end joining. Nucleic Acids Res *36*, 4022-4031.

Peterson, K.R., Li, Q.L., Clegg, C.H., Furukawa, T., Navas, P.A., Norton, E.J., Kimbrough, T.G., and Stamatoyannopoulos, G. (1995). Use of yeast artificial chromosomes (YACs) in studies of mammalian development: production of beta-globin locus YAC mice carrying human globin developmental mutants. Proceedings of the National Academy of Sciences of the United States of America *92*, 5655-5659.

Piganeau, M., Ghezraoui, H., De Cian, A., Guittat, L., Tomishima, M., Perrouault, L., Rene, O., Katibah, G.E., Zhang, L., Holmes, M.C., *et al.* (2013). Cancer translocations in human cells induced by zinc finger and TALE nucleases. Genome Res *23*, 1182-1193.

Ramirez, C.L., Certo, M.T., Mussolino, C., Goodwin, M.J., Cradick, T.J., McCaffrey, A.P., Cathomen, T., Scharenberg, A.M., and Joung, J.K. (2012). Engineered zinc finger nickases induce homology-directed repair with reduced mutagenic effects. Nucleic Acids Res *40*, 5560-5568.

Ristic, D., Modesti, M., Kanaar, R., and Wyman, C. (2003). Rad52 and Ku bind to different DNA structures produced early in double-strand break repair. Nucleic Acids Res *31*, 5229-5237.

Roberts, S.A., Strande, N., Burkhalter, M.D., Strom, C., Havener, J.M., Hasty, P., and Ramsden, D.A. (2010). Ku is a 5'-dRP/AP lyase that excises nucleotide damage near broken ends. Nature *464*, 1214-1217.

Schierling, B., Dannemann, N., Gabsalilow, L., Wende, W., Cathomen, T., and Pingoud, A. (2012). A novel zinc-finger nuclease platform with a sequence-specific cleavage module. Nucleic Acids Res *40*, 2623-2638.

Strande, N., Roberts, S.A., Oh, S., Hendrickson, E.A., and Ramsden, D.A. (2012). Specificity of the dRP/AP lyase of Ku promotes nonhomologous end joining (NHEJ) fidelity at damaged ends. J Biol Chem *287*, 13686-13693.

Strande, N.T., Carvajal-Garcia, J., Hallett, R.A., Waters, C.A., Roberts, S.A., Strom, C., Kuhlman, B., and Ramsden, D.A. (2014). Requirements for 5'dRP/AP lyase activity in Ku. Nucleic Acids Res *42*, 11136-11143.

Tovkach, A., Zeevi, V., and Tzfira, T. (2009). A toolbox and procedural notes for characterizing novel zinc finger nucleases for genome editing in plant cells. Plant J *57*, 747-757.

van Gent, D.C., Hoeijmakers, J.H., and Kanaar, R. (2001). Chromosomal stability and the DNA double-stranded break connection. Nat Rev Genet *2*, 196-206.

Waters, C.A., Strande, N.T., Pryor, J.M., Strom, C.N., Mieczkowski, P., Burkhalter, M.D., Oh, S., Qaqish, B.F., Moore, D.T., Hendrickson, E.A., *et al.* (2014). The fidelity of the ligation step determines how ends are resolved during nonhomologous end joining. Nature communications *5*, 4286.

Waugh, D.S., and Sauer, R.T. (1993). Single amino acid substitutions uncouple the DNA binding and strand scission activities of Fok I endonuclease. Proceedings of the National Academy of Sciences of the United States of America *90*, 9596-9600.

Weinthal, D.M., Taylor, R.A., and Tzfira, T. (2013). Nonhomologous end joining-mediated gene replacement in plant cells. Plant Physiol *162*, 390-400.

Wu, D., Topper, L.M., and Wilson, T.E. (2008). Recruitment and dissociation of nonhomologous end joining proteins at a DNA double-strand break in Saccharomyces cerevisiae. Genetics *178*, 1237-1249.

Zhang, Y., Hefferin, M.L., Chen, L., Shim, E.Y., Tseng, H.M., Kwon, Y., Sung, P., Lee, S.E., and Tomkinson, A.E. (2007). Role of Dnl4-Lif1 in nonhomologous end-joining repair complex assembly and suppression of homologous recombination. Nat Struct Mol Biol *14*, 639-646.

Zhou, T., Akopiants, K., Mohapatra, S., Lin, P.S., Valerie, K., Ramsden, D.A., Lees-Miller, S.P., and Povirk, L.F. (2009). Tyrosyl-DNA phosphodiesterase and the repair of 3'-phosphoglycolate-terminated DNA double-strand breaks. DNA Repair (Amst) *8*, 901-911.

CHAPTER 6

Concluding Remarks

Summary of results

Results of the research projects presented in this dissertation provide important new insights into the relationship between NHEJ and genome stability using systems of *Agrobacterium*-plant interactions and the budding yeast *Saccharomyces cerevisiae*, two diverse yet interconnected systems in the context of DSB repair, especially NHEJ. Findings in T-DNA double-stranded formations in **Chapter 2** provide experimental evidence of the hypothetical process governing the rapid biogenesis of potential integration substrates for plant genetic transformation. In **Chapter 3**, analysis of transgenic plants carrying genomic integration of mega T-DNA constructs following several generations provides important references for multiple transgene expression in model plants. Study of the yeast specialized c-NHEJ DNA ligase IV in **Chapter 4** reveal its previously mysterious function in promoting mutagenesis in the absence of proper catalysis, potentially informing the disease mechanism governing the human Ligase IV syndrome. Most recently, study of the role of overhang polarity in rapidly induced chromosomal double-strand breaks in **Chapter 5**, sheds light on the difference of the repair processes involved and mutagenesis outcomes. Along with the important findings, my thesis research also developed new biotechnical and experimental techniques for applications and research in the field of genome editing and DNA repair. In **Chapter 6**, I will discuss important implications and future directions of my completed thesis research.

Study the genetic requirements for T-strand conversion and its kinetics in yeast

In **Chapter 2**, formation of the dsT-DNA intermediates are likely products of the plant cell machinery. Referring to DNA replication and our results from this Chapter, I hypothesize that core replication proteins such as primase and polymerase are required for formation of dsT-DNA. To confirm this by experiment, one can use plant mutants and RNAi lines to alter key potential plant factors of T-strand conversion. However, there is a strong limitation to have viable mutants in essential DNA replication machinery despite although it is still possible. Alternatively, it is well established that Agrobacterium can infect yeast (Piers *et al.*, 1996; Bundock *et al.*, 2002), which have a good number of well-characterized temperature sensitive mutants in these pathways that we can use. Therefore, we can modify the published approaches using two novel readouts for dsT-DNA formation (**Figure 6-1**). First, we will create a T-DNA substrate with an internal CEN/ARS replication origin and URA3 marker gene, to allow for its maintenance as an episome if it is circularized. The T-DNA will also bear a short terminal microhomology just internal to the T-DNA left and right borders. Such substrates undergo efficient single-strand annealing (SSA)-mediated circularization, and thus stably transform yeast, in a manner that depends on 3' strands at both ends. Because T-strands introduced by Agrobacterium have a 3' strand at only one end they are incompetent for SSA, so that the extent of infection-mediated transformation of yeast to Ura+ will correlate with the efficiency of dsT-DNA formation (**Figure 6-1A**). qPCR flanking the circularization junction will also be applied to detect the T-DNA episome without the need for yeast outgrowth.

Our second readout will directly measure formation of the non-T-strand by physical monitoring using strand-specific qPCR (**Figure 6-1B**). Strand-specific qPCR has been used to

specifically amplify + or – strands of DNA and RNA molecules in many situations (Meyerhans *et al.*, 1992; Bingham *et al.*, 1996; Booth *et al.*, 2001; Lanford *et al.*, 1994). The common concept is that a single 5'-tailed primer is used in a first primer extension reaction to mark and linearly amplify a specific template strand. Subsequently, two new primers are added for qPCR amplification, with one primer annealing to the 5' tail of the primer extension primer. To prevent the primer extension primer from binding during qPCR, the T_m of the 1stand 2nd round primers are adjusted so that qPCR is performed at an annealing temperature too high for the primer extension primer to bind. In this way, only strands tagged in the linear primer extension cycles are exponentially amplified so that qPCR signal correlates to the starting amount of the strand being interrogated. Correlating the qPCR signals for the T-strand (present in all T-DNA forms using blue primers in **Figure 6-1**) and the non-T-strand (present only in dsT-DNA using red primers in **Figure 6-1**) will reveal the extent of dsT-DNA conversion. Importantly, the non-T-strand is also present in the binary vector present in *Agrobacterium* cells, so that contamination of yeast DNA with *Agrobacterium* DNA would confound results. To prevent this, *Agrobacterium* cells will be removed by differential centrifugation. Pilot experiments with this approach, scored by plating and colony counting, gave a 100-fold excess of yeast to bacterial cells in the final product, which can likely be improved by adding an SDS wash to lyse the bacterial but not the detergent-resistant yeast cells. For qPCR experiments, we will normalize the T-DNA strand-specific results to the extent of binary vector contamination determined using a separate strand-specific qPCR assay directed at the non-T-DNA portion of the vector backbone.

Using these readouts, we will first examine the baseline properties of T-strand to dsT-

DNA conversion. This will include the overall efficiency as well as kinetics of dsT-DNA formation relative to the appearance of T-strand in the cell, judged by using parallel control qPCR reactions directed against yeast genomic DNA. We will then look for differences in the efficiency of dsT-DNA formation in yeast mutants of the replication and recombination machineries. Genes to be studied include NHEJ proteins (*YKU70*, *DNL4*), HR proteins (*RAD51*, *RAD52*), and, most importantly, replication proteins (primase, *PRI1*; DNA polymerase α, *POL1*; DNA polymerase α, *POL3/CDC2*; DNA polymerase α, *POL2*). Because replication mutants are cell lethal, we will use well described temperature alleles that allow yeast to be grown at permissive temperatures with the target protein later inactivated during analysis at restrictive temperatures (*pri1*-1, *pol1*-1, *cdc2-3*, *pol2-11*) (Conrad and Newlon, 1983; Budd and Campbell, 1987; Francesconi *et al.*, 1991; Budd and Campbell, 1993). All are available to us in our strain set. A loss of dsT-DNA signal at the restrictive temperature will implicate that protein and pathway in dsT-DNA conversion.

Study the molecular mechanisms of break-end structure recognition in yeast

In **Chapter 5**, difference in repair kinetics and fidelity of rejoining of chromosomal DSBs with different overhang polarities is likely mediated by unknown mechanisms by which break-end structures are recognized. HO endonuclease is the native yeast enzyme initiates the mating-type switch (Haber, 1998) by creating a 3' overhanging DSB (3' DSB) in the *MAT* locus. This 3' DSB is repaired by the homologous recombination to complete the switch. Interestingly, our results suggest that chromosomal 5' overhanging DSBs (5' DSBs), at least in the studied *ILV1* promoter locus, are rejoined more rapidly with higher frequency of mutations than 3' DSBs, which may suggest a reason why the 3' DSBs is selected in evolution for yeast mating-type switch. Rejoining to a mutated *MAT* locus would abolish recleavage of the site and lead to failure of this important cell activity for yeast. Although the results of different assays are consistent and indicate that the overhang polarity of chromosomal DSBs can greatly impact the kinetics and fidelity of yeast NHEJ, we have not fully understood the mechanism(s) by which overhang polarity are recognized by NHEJ factors. We also do not fully understand the recognition mechanisms of other more complex break ends (Daley and Wilson, 2005), especially in the context of the highly complicated in-vivo environment.

Inspired by the biochemistry findings of DNA-PKcs (the c-terminus of human Ku80) in recognition of the overhang polarity (Hammarsten *et al.*, 2000; Pawelczak *et al.*, 2005; Pawelczak and Turchi, 2008; Pawelczak *et al.*, 2011), we are testing the possibility that the c-terminus of yeast Yku80 in mediating the recognition of the overhang polarity, partially leading to our observed difference between ZFN-mediated and HO-mediated DSBs. Preliminary results show that repair of 5' DSBs has a larger defect (in NHEJ efficiency and DNL4 recruitment) than

3’ DSBs in the absence of the Yku80 c-terminus. Future work will be required to dissect the precise role of Yku80 c-terminus and presumably other end-joining components because of the nature of the end-joining catalysis with multiple interacting complexes.

Besides the Yku complex, exonucleases involved in NHEJ and 5’ resection might be able to recognize overhang polarity because of their directional activities at DNA ends. For example, Mre11, part of the MRX complex has 3’, but no 5’, exonuclease activity. Purified Mre11 and Mre11-Rad50 showed similar exonuclease activity in 5’ overhanging and blunt DNA ends, and much lower activity in 3’ overhanging ends (Paull and Gellert, 1998). This is consistent with our in vivo observations that the MRX complex is responsible for the robust end-processing in the 5’ DSBs, but not in the 3’ DSBs. As a putative model, degradation of the 3’ strands by MRX might help expose surrounding microhomologies of the 5’ DSBs, which facilitates their rejoining at the cost of more frequent mutations (**Figure 6-2A**). In addition, purified Mre11 forms stable complexes with 3’ overhangs which are resistant to Mre11 degradation (Paull and Gellert, 1998). It is possible that the MRX complex could protect 3’ DSBs from degradation by other nucleases during the initial end-joining phase before transition to 5’ resection in vivo (**Figure 6-2**). This may explain why the 3’ DSBs were less processed as compared to 5’ DSBs and why the loss of MRX and Yku complex showed a similar promoting effect on end processing of 3’ DSBs.

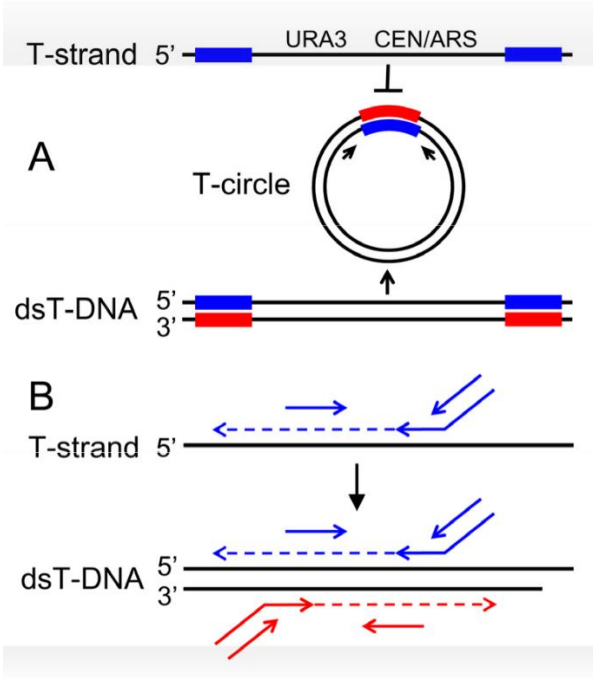


Figure 6-1. Assays for detecting dsT-DNA formation in yeast. (**A**) Terminal T-DNA repeats mediate circulation of the transfected T-DNA (T-circle) by SSA in yeast detectable by the T-DNA URA3 marker or qPCR flanking the circularization junction.(**B**) Parallel strand- specific qPCR assays detect the T-strand and nonT-strand (blue and red primer sets, respectively). Each assay is a two- step reaction: step 1 is linear amplification/primer extension and step 2 is qPCR where one primer binds to the 5' tail of the step 1 primer.

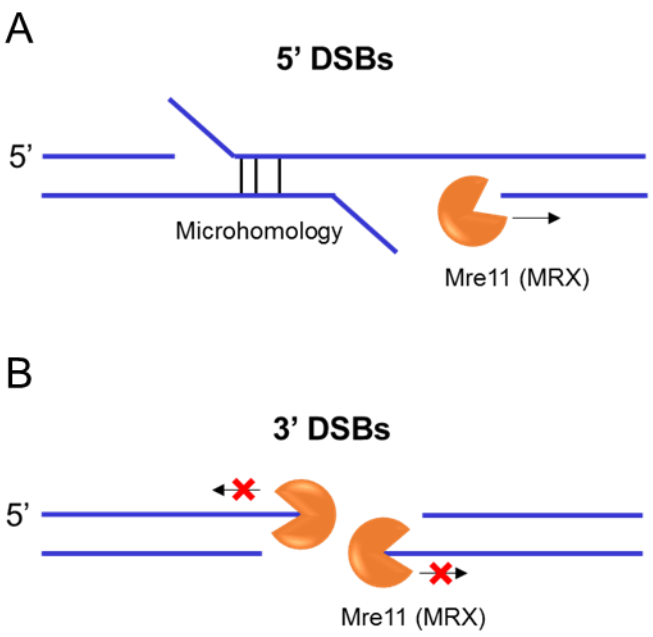


Figure 6-2. A model of Mre11 (MRX) exonuclease activity at 5' and 3' DSBs. (**A**) Mre11 (MRX) robustly degrades 3' strands of 5' DSBs allowing annealing of the 5' strands via microhomologies for rapid imprecise rejoining. (**B**) 3' DSBs are resistant to degradation of Mre11 (MRX). The DSB-bound MRX protects the break ends from processing by other proteins.

Reference

Bingham, J.P., Hartley, J.A., Souhami, R.L., and Grimaldi, K.A. (1996). Strand-specific measurement of cisplatin-induced DNA damage and repair using quantitative PCR. Nucleic Acids Res *24*, 987-989.

Booth, C., Griffith, E., Brady, G., and Lydall, D. (2001). Quantitative amplification of single-stranded DNA (QAOS) demonstrates that cdc13-1 mutants generate ssDNA in a telomere to centromere direction. Nucleic Acids Res *29*, 4414-4422.

Budd, M., and Campbell, J.L. (1987). Temperature-sensitive mutations in the yeast DNA polymerase I gene. Proceedings of the National Academy of Sciences of the United States of America *84*, 2838-2842.

Budd, M.E., and Campbell, J.L. (1993). DNA polymerases delta and epsilon are required for chromosomal replication in Saccharomyces cerevisiae. Mol Cell Biol *13*, 496-505.

Bundock, P., van Attikum, H., den Dulk-Ras, A., and Hooykaas, P.J. (2002). Insertional mutagenesis in yeasts using T-DNA from Agrobacterium tumefaciens. Yeast *19*, 529-536.

Conrad, M.N., and Newlon, C.S. (1983). Saccharomyces cerevisiae cdc2 mutants fail to replicate approximately one-third of their nuclear genome. Mol Cell Biol *3*, 1000-1012.

Daley, J.M., and Wilson, T.E. (2005). Rejoining of DNA double-strand breaks as a function of overhang length. Mol Cell Biol *25*, 896-906.

Francesconi, S., Longhese, M.P., Piseri, A., Santocanale, C., Lucchini, G., and Plevani, P. (1991). Mutations in conserved yeast DNA primase domains impair DNA replication in vivo. Proceedings of the National Academy of Sciences of the United States of America *88*, 3877-3881.

Haber, J.E. (1998). Mating-type gene switching in Saccharomyces cerevisiae. Annu Rev Genet *32*, 561-599.

Hammarsten, O., DeFazio, L.G., and Chu, G. (2000). Activation of DNA-dependent protein kinase by single-stranded DNA ends. J Biol Chem *275*, 1541-1550.

Lanford, R.E., Sureau, C., Jacob, J.R., White, R., and Fuerst, T.R. (1994). Demonstration of in vitro infection of chimpanzee hepatocytes with hepatitis C virus using strand-specific RT/PCR. Virology *202*, 606-614.

Meyerhans, A., Vartanian, J.P., and Wain-Hobson, S. (1992). Strand specific PCR amplification of low copy number DNA. Nucleic Acids Res *20*, 521-523.

Paull, T.T., and Gellert, M. (1998). The 3' to 5' exonuclease activity of Mre 11 facilitates repair of DNA double-strand breaks. Mol Cell 1, 969-979.

Pawelczak, K.S., Andrews, B.J., and Turchi, J.J. (2005). Differential activation of DNA-PK based on DNA strand orientation and sequence bias. Nucleic Acids Res 33, 152-161.

Pawelczak, K.S., Bennett, S.M., and Turchi, J.J. (2011). Coordination of DNA-PK activation and nuclease processing of DNA termini in NHEJ. Antioxid Redox Signal 14, 2531-2543.

Pawelczak, K.S., and Turchi, J.J. (2008). A mechanism for DNA-PK activation requiring unique contributions from each strand of a DNA terminus and implications for microhomology-mediated nonhomologous DNA end joining. Nucleic Acids Res 36, 4022-4031.

Piers, K.L., Heath, J.D., Liang, X., Stephens, K.M., and Nester, E.W. (1996). Agrobacterium tumefaciens-mediated transformation of yeast. Proceedings of the National Academy of Sciences of the United States of America *93*, 1613-1618.