UNIVERSITY OF CALIFORNIA
RIVERSIDE

A Study of Factors Controlling Transposition and How to Exploit Them

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Microbiology

by

Femila Lilly Manoj

June 2024

Dissertation Committee:
Prof. Thomas Kuhlman, Chairperson
Prof. Howard Judelson
Prof. Yujie Men

The Dissertation of Femila Lilly Manoj is approved:

_____

_____

_____
Committee Chairperson

University of California, Riverside

ACKNOWLEDGMENTS

The first chapter of this dissertation is based on “Targeted Insertion of Large Genetic Payload Using Cas Directed LINE-1 Reverse Transcriptase”, which was published in Scientific Reports Volume 11 Article Number 23625 on December 8, 2021. The authors of this paper are Femila Manoj, Laura W. Tai, Katelyn Sun Mi Wang, and Thomas E. Kuhlman who are affiliated with the Department of Physics at the University of California, Riverside. Katelyn was a student at Chaminade College Preparatory High School. Analysis and experiments are attributed to the authors who performed the work in their corresponding sections.

The second chapter of this dissertation is based on “Real-Time Quantification of the Effects of IS200/IS605 Family-Associated TnpB on Transposon Activity”, which was published in The Journal of Visualized Experiments (JoVE) Volume 191 Article Number 64825 on January 28, 2023. The authors of this paper are Michael Worcester, Femila Manoj, and Thomas E. Kuhlman who are affiliated with the Department of Physics at the University of California, Riverside. Analysis and experiments are attributed to the authors who performed the work in their corresponding sections.

The work in the final chapter of this dissertation is based on “Phylogenetic Distribution of Prokaryotic Non-homologous End Joining DNA Repair Systems in Bacteria and Archaea”, which was posted on bioRxiv on September 30, 2023. The authors of this paper are Femila Manoj and Thomas E. Kuhlman who are affiliated with the Department

of Physics at the University of California, Riverside. Analysis and experiments are attributed to the authors who performed the work in their corresponding sections.

DEDICATION

Dedicated to XX.

ABSTRACT OF THE DISSERTATION

A Study of Factors Controlling Transposition and How to Exploit Them

by

Femila Lilly Manoj

Doctor of Philosophy, Graduate Program in Microbiology
University of California, Riverside, June 2024
Prof. Thomas Kuhlman, Chairperson

Since the discovery of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), it has allowed for genome editing to be utilized in a way never seen before. Although it can create knockout and point mutations at ease, it can still be difficult to create knockin mutations. In order to combat this issue, GENEWRITE was created utilizing the reverse transcription function of the LINE-1 (long interspersed element-1) and combining it with the commonly used Cas (CRISPR associated proteins). This also worked around the need for CRISPR to silence NHEJ as it can be utilized in the protocol as well. This, however, made it difficult for GENEWRITE to function in organisms such as prokaryotes and archaea which did not already carry a NHEJ pathway. To identify prokaryotes and archaea the GENEWRITE protocol can function in, a bioinformatics analysis was done to identify the known forms of NHEJ in prokaryotes and archaea. Additionally, it was recently found that TnpB, a transposon associated gene from the IS*200*/IS*605* family found in bacteria, is a RNA-guided DNA endonuclease that could be programmed

similar to Cas proteins. To verify its potential function within the transposon, a quantitative microscopy study was carried out to observe how TnpB affected transposition in real time.

Table of Contents

LIST OF FIGURES

Introduction

Chapter 1: Targeted Insertion of Large Genetic Payload Using CAS Directed Line-1 Reverse Transcriptase

Chapter 2: Real-time Quantitification of the Effects of IS*200*/IS*605* Family-Associated TnpB on Transposon Activity

Chapter 3: Phylogenetic Distribution of Prokaryotic Non-homologous End Joining DNA Repair Systems in Bacteria and Archaea

LIST OF TABLES

Chapter 1: Targeted Insertion of Large Genetic Payload Using CAS Directed Line-1 Reverse Transcriptase

INTRODUCTION

"*Wooooooooooow! Science!" - Femila Manoj*

Background

Transposable elements (TES) are mobile genetic elements found in all the domains of life[1]. They were first discovered by Dr. Barbara McClintock in the mid 1900s when she observed corn kernels on a single ear of corn exhibited various colors instead of producing one uniform ear[2]. She discovered that this coloration was due to the presence or absence of DNA within the pigmentation gene. The DNA in the gene was found to be able to 'jump' in and out of the gene either allowing or preventing the pigment to be produced in the kernel. This 'jumping' DNA which was known as a 'jumping gene' for some time is what is now known as a TE. Many other scientists were skeptical of her findings at the time suppressing her career, but in 1983 she was finally awarded a Nobel Prize in Medicine for her findings.

It has since been found that TEs can reverse transcribe the sequence back into the genome by transcribing RNA originally transcribed from the TE back into DNA allowing it to reenter the genome[3]. This is done by a reverse transcriptase protein which can transcribe RNA into DNA. This means that TEs allow for self-editing of the genome and introduces diversity to an organism and species via these edits. While this also suggests that TEs could keep replicating and be found all over a genome, they are actually found in low copy numbers in prokaryotes as well as archaea, usually only

having a few TEs per organism[4]. In prokaryotes, this is because the primary mechanism of propagation is retrohoming, where an intron-encoded homing endonuclease allows TEs to insert into a specific target site rather than untargeted replicating all over the genome which is referred to as retrotransposition[5].

Insertion Sequences (ISs) are the simplest form of TEs which carry only the genes required for transposition allowing them to be less than 2.5 kbp[6]. ISs have become more relevant in recent years as it has been found that they aid in dispersing antibiotic resistance genes leading to the creation of antibiotic resistance strains of bacteria[7]. IS*200*/IS*605* is a family of TEs that are distributed widely in both prokaryotes and archaea[8]. TE members of this family carry imperfect palindromic (IP) sequences on their ends rather than inverted sequences which is more common for many transposons[9]. In some cases such as that found in family member IS*608*, a common experimental model system for TEs within the IS*200*/IS*600* family, the IPs are nearly identical allowing the formation of a stem loop structure in the DNA[10]. Two relevant genes from this IS*608* are tnpA and tnpB. TnpA is the transposase gene which does the actual transposing[11]. The tnpB gene is only found in the IS*600* family; the IS*200* family carries only tnpA[12]. Until recently, the function of tnpB was unknown, but it was recently discovered that it may regulate transcription levels and may even be able to be programmed to cut target sites in the genome similar to Cas (CRISPR [Clustered Regularly Interspaced Short Palindromic Repeats] associated) proteins[13].

In archaea, it is believed that TEs are found at low copy numbers because they are horizontally transferred from prokaryotes based on bioinformatics studies[4,14]. They

are found in eukaryotic organelles, including the chromosomes of mitochondria and chloroplasts at similar rates as in prokaryotes, but with their own unique genetic lineage. While they are found in eukaryotic organelles, they are absent in the eukaryotic nuclear genome[15]. However, it is also thought that these TEs are the evolutionary ancestors of spliceosomal introns as well as retrotransposons in eukaryotic organisms, such as the human retrotransposon long interspersed element-1 (LINE-1 or L1)[4].

The first copy of LINE-1 found was 6.4 kb, 4 times larger than any other transposon that had yet been discovered in 1980[16]. Since its discovery, LINE-1 has been the sole known active autonomous retrotransposon in humans[17]. On its own, LINE-1 sequences and their remnants make up 17% of the genome, but the vast majority of these sequences are mutated suppressing transposition from occurring[18,19]. LINE-1 transposes with the aid of its two open reading frames: ORF1 and ORF2[20]. The function of ORF2 was first discovered as it carries the endonuclease (EN) and reverse transcriptase (RT) domains[21,22]. The EN is required to recognize and cut genomic DNA at specific sites to allow for LINE-1 to reinsert into a new part of the genome. LINE-1 has been known to target A/T rich areas of the chromosome[23]. The ORF1 transcribes a RNA binding protein which aids retrotransposition with its chaperone activity[24]. The few functioning copies of LINE-1 can retrotranspose leading to diversification as well as the creation of new mutations. In many cases, non-homologous end-joining (NHEJ) has been known to affect the integration of LINE-1 during retrotransposition.

NHEJ is a form of repair for double-strand breaks (DSBs) in DNA that does not require homology between DNA ends[25]. Due to the potential use of incompatible

DNA ends, NHEJ is considered a less accurate form of DSB repair than homology-directed repair (HDR). For NHEJ to commence, Ku proteins must bind to the ends of the DSB forming a sort of scaffold[26]. A DNA ligase protein then works with the Ku scaffolding to rejoin the ends of the DSB[27]. While NHEJ can easily be found in eukaryotes, it is much less common in prokaryotes and only found in a single species of archaea[28,29].

Motivation (Review of Literature)

Despite their flexibility and ease of use, the repertoire of genome editing modalities that Cas proteins systems allow remains limited. Knockout or point mutants can be generated relatively easily by targeting Cas cleavage to coding or control regions of the genome. The cell must repair such cuts to survive, and errors introduced by the NHEJ repair machinery can lead to inactivation of control regions or introduction of missense or point mutations to coding sequences[30–33]. An additional editing modality is to introduce novel sequences to the genome through Homology Directed Repair (HDR), where a DNA fragment with ends homologous to the sequences flanking the cut site and containing the desired sequence to be inserted is introduced to the cell along with the Cas-sgRNA ribonucleoprotein (RNP) complexes. After cleavage, the fragment is then used to repair the cut by the cell's homologous recombination repair machinery, resulting in its integration. However, HDR remains inefficient and difficult to accomplish, particularly for gene-sized or larger [$\geq \sim 1$ kilobase pair (kbp)] fragments[34–37]. A primary reason for this difficulty is that for HDR to be successful, non-homologous end joining

(NHEJ) DNA repair, the primary repair mechanism for DNA repair in advanced eukaryotic cells, must be suppressed[38–44].

By creating a novel genome editing tool utilizing NHEJ rather than HDR, it becomes more viable in eukaryotic cells, however prokaryotes and archaea do not all carry the genes required for NHEJ. In fact, prokaryotes have been found to have at least three different possible types of NHEJ (Fig. 0.1). These NHEJ systems are based on the characterized pathways observed in *Bacillus subtilis*, *Streptomyces ambofaciens*, and *Sinorhizobium meliloti*. In the case of *B. subtilis* it has been determined that deletion of either the LigD or Ku protein hindered the functionality of NHEJ, suggesting both are necessary for the pathway[45]. This particular form of NHEJ that relies on only two proteins, is referred to as "minimal NHEJ". Further investigations conducted in *S. ambofaciens* revealed that mutations affecting other NHEJ-associated proteins, such as LigC, KuA, PolR, and PolK, had a significant impact on the overall efficiency of the NHEJ pathway[46]. The NHEJ pathway encompassing these four proteins will be designated as "core NHEJ". More recent findings in *S. meliloti* unveiled the existence of two additional distinct functional NHEJ pathways, each activated under different stress conditions[47]. The first pathway, operational in stress-free bacteria, necessitates the presence of LigD2 and Ku2 proteins and is referred to as the "main NHEJ" pathway. Conversely, the second pathway, induced exclusively under stress conditions, requires the participation of LigD4, Ku3, and Ku4 proteins and is referred to as the "secondary NHEJ" pathway. The combined occurrence of both main and secondary NHEJ pathways

in *S. meliloti* is referred to as "multiple NHEJ" in this. The distribution of minimal NHEJ in prokaryotes is known, but the distribution of core and secondary NHEJ remain unknown as they have only recently been discovered. This suggests that NHEJ may be in more prokaryotes or archaea than previously thought as well as the existence of more NHEJ pathways that have yet to be found.

Additionally, it was recently found that TnpB, a transposon associated gene from the IS*200/*IS*605* family found in bacteria, is a RNA-guided DNA endonuclease that could be programmed similar to Cas proteins[13]. TnpB had long been thought to be unessential for transposition of TEs in bacteria until bioinformatics predicted a potential RuvC gene within the TnpB gene[48]. RuvC is a part of the Cas protein that is responsible for cleavage[49]. All this information suggests that TnpB could be an ancestor of the Cas protein and that TnpB could be used similarly to Cas proteins in genome editing protocols.

Summary of Thesis Work

The work presented in this dissertation aims to better understand tools that could be used in genome editing as well as test a novel genome editing tool. This novel tool introduces a method for the active insertion of lengthy genetic sequences into host DNA we call GENEWRITE: Genome Engineering With RNA-Integrating Targetable Endonucleases. This is accomplished by coupling the targetable endonuclease activity of Cas enzymes to the reverse transcriptase activity of the human retrotransposon LINE-1 through translationally fusing Cas and LINE-1 reverse transcriptase proteins (Fig. 0.2).

We use *E. coli* expressing *B. subtilis* NHEJ enzymes as a simple experimental model to optimize the design and delivery of GENEWRITE for future application to more complex systems. The strategy used here for integration using GENEWRITE is shown in Fig. 2B–F. *E. coli* cells expressing GENEWRITE are transformed with two RNAs: a guide sgRNA to target Cas cleavage to the desired integration site, and a payload RNA carrying the coding sequence of the desired integration. The 3' end of the payload RNA is designed to be homologous to the bottom DNA strand downstream from the cut such that RNA–DNA hybridization occurs to prime reverse transcription. Host enzymes complete second strand synthesis and payload RNA removal, and the insert is sealed into the site[49]. We illustrate that GENEWRITE can be used to effectively target insertion of large, gene-sized payloads to specific locations, although not without off-target effects.

This dissertation also shows how real-time imaging of transposon activity can quantitatively reveal the impact of TnpB or any other accessory proteins on transpositional dynamics. By fusing TnpB to fluorescence protein mCherry, the individual transpositional events are identified by blue fluorescence and correlated with expression levels of TnpA (yellow fluorescence) and TnpB (red fluorescence). These results show that TnpB may aid the regulation and maintenance of transposition in bacteria.

Finally, a bioinformatic analysis was completed on available bacterial and archaeal genome sequences to characterize the phylogenetic distribution of the various NHEJ pathways found in prokaryotes. Prokaryotic species with all proteins required for

each NHEJ category were considered to have a complete NHEJ pathway. To identify archaea species that may have prokaryotic NHEJ systems, species with the four genes required for NHEJ in *Methanocella paludicola* were identified as it is the only form of prokaryotic NHEJ known to exist in archaea to date[4]. Through this analysis, it was found that NHEJ may be in more prokaryotes and archaea than previously expected although further investigation of prokaryotic NHEJ must be done to confirm complete pathways.

REFERENCES

1. Novikova, O. & Belfort, M. Mobile Group II Introns as Ancestral Eukaryotic Elements. *Trends Genet.* **33**, 773–783 (2017).

2. McCLINTOCK, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**, 344–355 (1950).

3. Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209–1211 (1970).

4. Mohr, G., Ghanem, E. & Lambowitz, A. M. Mechanisms used for genomic proliferation by thermophilic group II introns. *PLoS Biol.* **8**, e1000391 (2010).

5. Cousineau, B. *et al.* Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* **94**, 451–462 (1998).

6. Mahillon, J. & Chandler, M. Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).

7. Debets-Ossenkopp, Y. J. *et al.* Insertion of mini-IS605 and deletion of adjacent sequences in the nitroreductase (rdxA) gene cause metronidazole resistance in Helicobacter pylori NCTC11637. *Antimicrob. Agents Chemother.* **43**, 2657–2662 (1999).

8. He, S. *et al.* The IS200/IS605 Family and 'Peel and Paste' Single-strand Transposition Mechanism. *Microbiol Spectr* **3**, (2015).

9. Barabas, O. *et al.* Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. *Cell* **132**, 208–220 (2008).

10. Ronning, D. R. *et al.* Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Mol. Cell* **20**, 143–154 (2005).

11. de la Cruz, F. & Grinsted, J. Genetic and molecular characterization of Tn21, a multiple resistance transposon from R100.1. *J. Bacteriol.* **151**, 222–228 (1982).

12. Kersulyte, D., Mukhopadhyay, A. K., Shirai, M., Nakazawa, T. & Berg, D. E. Functional organization and insertion specificity of IS607, a chimeric element of Helicobacter pylori. *J. Bacteriol.* **182**, 5300–5308 (2000).

13. Karvelis, T. *et al.* Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021).

14. Rest, J. S. & Mindell, D. P. Retroids in Archaea: Phylogeny and Lateral Origins. *Mol. Biol. Evol.* **20**, 1134–1142 (2003).

15. Truong, D. M., Hewitt, F. C., Hanson, J. H., Cui, X. & Lambowitz, A. M. Retrohoming of a Mobile Group II Intron in Human Cells Suggests How Eukaryotes Limit Group II Intron Proliferation. *PLoS Genet.* **11**, e1005422 (2015).

16. Adams, J. W., Kaufman, R. E., Kretschmer, P. J., Harrison, M. & Nienhuis, A. W. A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. *Nucleic Acids Res.* **8**, 6113–6128 (1980).

17. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).

18. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

19. Ostertag, E. M. & Kazazian, H. H., Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).

20. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H., Jr. Isolation of an active human transposable element. *Science* **254**, 1805–1808 (1991).

21. Feng, Q., Moran, J. V., Kazazian, H. H., Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).

22. Mathias, S. L., Scott, A. F., Kazazian, H. H., Jr, Boeke, J. D. & Gabriel, A. Reverse transcriptase encoded by a human transposable element. *Science* **254**, 1808–1810 (1991).

23. Cost, G. J. & Boeke, J. D. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37**, 18081–18093 (1998).

24. Martin, S. L. & Bushman, F. D. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.* **21**, 467–475 (2001).

25. Benson, F. E., Baumann, P. & West, S. C. Synergistic actions of Rad51 and Rad52 in recombination and DNA repair. *Nature* **391**, 401–404 (1998).

26. Mari, P.-O. *et al.* Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 18597–18602 (2006).

27. Costantini, S., Woodbine, L., Andreoli, L., Jeggo, P. A. & Vindigni, A. Interaction of the Ku heterodimer with the DNA ligase IV/Xrcc4 complex and its regulation by DNA-PK. *DNA Repair* **6**, 712–722 (2007).

28. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55**, 709–742 (2001).

29. Bartlett, E. J., Brissett, N. C., Plocinski, P., Carlberg, T. & Doherty, A. J. Molecular basis for DNA strand displacement by NHEJ repair polymerases. *Nucleic Acids Res.* **44**, 2173–2186 (2016).

30. Guo, T. *et al.* Harnessing accurate non-homologous end joining for efficient precise deletion in CRISPR/Cas9-mediated genome editing. *Genome Biol.* **19**, 1–20 (2018).

31. Bothmer, A. *et al.* Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nat. Commun.* **8**, 1–12 (2017).

32. Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks. *Mol. Cell* **70**, 801–813.e6 (2018).

33. van Overbeek, M. *et al.* DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Mol. Cell* **63**, 633–646 (2016).

34. Nambiar, T. S. *et al.* Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant. *Nat. Commun.* **10**, 1–13 (2019).

35. Aird, E. J., Lovendahl, K. N., St. Martin, A., Harris, R. S. & Gordon, W. R. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Communications Biology* **1**, 1–6 (2018).

36. Liu, M. *et al.* Methodologies for Improving HDR Efficiency. *Front. Genet.* **9**, 422460 (2019).

37. Rozov, S. M., Permyakova, N. V. & Deineko, E. V. The Problem of the Low Rates of CRISPR/Cas9-Mediated Knock-ins in Plants: Approaches and Solutions. *Int. J. Mol. Sci.* **20**, 3371 (2019).

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

39. Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat. Rev. Mol. Cell Biol.* **4**, 712–720 (2003).

40. Davis, A. J. & Chen, D. J. DNA double strand break repair via non-homologous end-joining. *Transl. Cancer Res.* **2**, 130–143 (2013).

41. Iliakis, G. Backup pathways of NHEJ in cells of higher eukaryotes: Cell cycle dependence. *Radiother. Oncol.* **92**, 310–315 (2009).

42. Devkota, S. The road less traveled: strategies to enhance the frequency of homology-directed repair (HDR) for increased efficiency of CRISPR/Cas-mediated transgenesis. *BMB Rep.* **51**, 437–443 (2018).

43. Li, G. *et al.* Small molecules enhance CRISPR/Cas9-mediated homology-directed genome editing in primary cells. *Sci. Rep.* **7**, 1–11 (2017).

44. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).

45. Weller, G. R. *et al.* Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297**, 1686–1689 (2002).

46. Hoff, G. *et al.* Multiple and Variable NHEJ-Like Genes Are Involved in Resistance to DNA Damage in. *Front. Microbiol.* **7**, 1901 (2016).

47. Dupuy, P., Sauviac, L. & Bruand, C. Stress-inducible NHEJ in bacteria: function in DNA repair and acquisition of heterologous DNA. *Nucleic Acids Res.* **47**, 1335–1349 (2019).

48. Kapitonov, V. V., Makarova, K. S. & Koonin, E. V. ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs. *J. Bacteriol.* **198**, 797–807 (2015).

49. Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.* **39**, 9275–9282 (2011).

50. Lee, G. *et al.* Testing the retroelement invasion hypothesis for the emergence of the ancestral eukaryotic cell. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12465–12470 (2018).

FIGURES AND TABLES

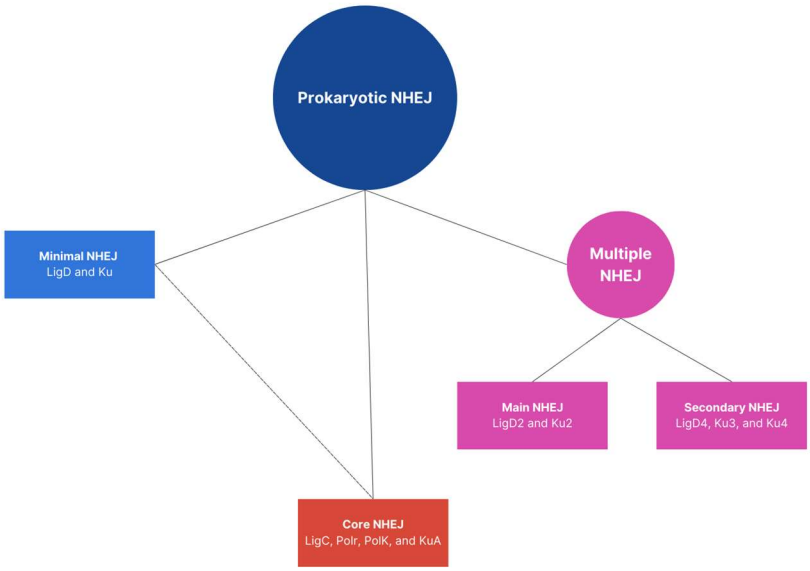

Fig 0.1. NHEJ pathways. A conceptual map illustrating the distinct categories of non-homologous end joining (NHEJ) observed in prokaryotes, accompanied by the corresponding protein constituents for each type. The prokaryotic NHEJ repertoire encompasses three major classifications: minimal, core, and multiple NHEJ. Minimal NHEJ requires LigD and Ku proteins. Core involvesLigC, PolR, PolK, and KuA proteins. Multiple NHEJ comprises two additional pathways: main and secondary NHEJ. Main NHEJ requires LigD2 and Ku2 proteins, whereas secondary NHEJ relies on LigD4, Ku3, and Ku4 proteins.

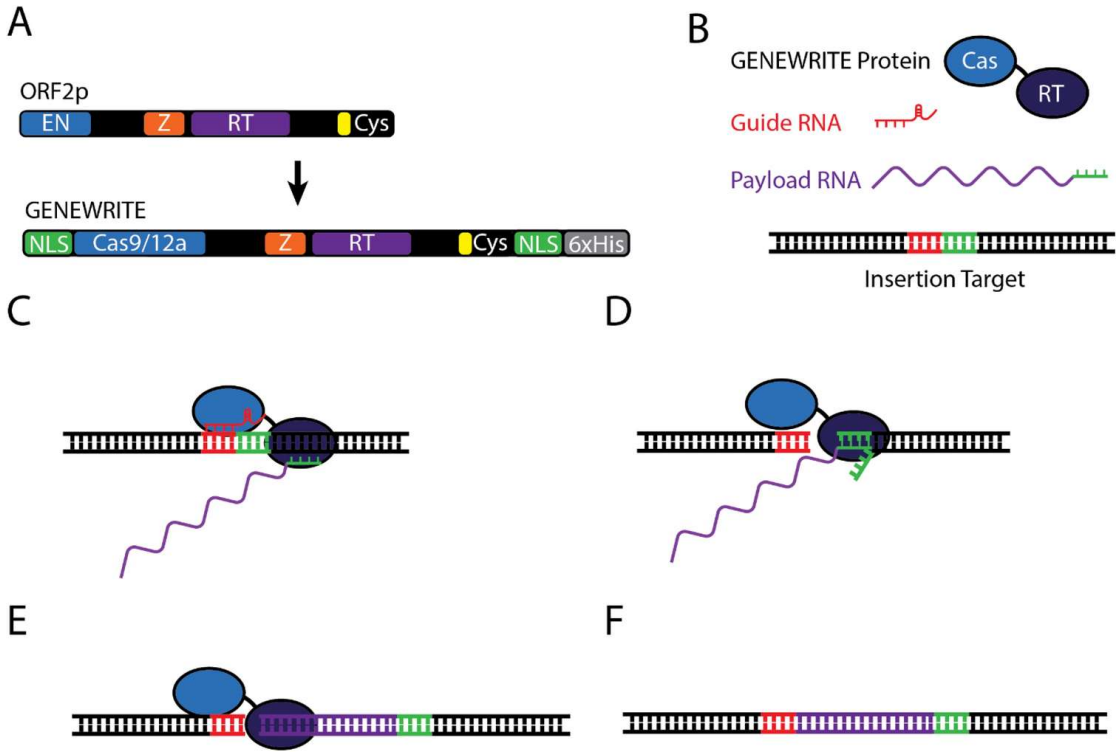

Figure 0.2. GENEWRITE components and strategy. (**A**) ORF2p and GENEWRITE domain structure. Wildtype ORF2p consists of endonuclease (EN, blue), Z (Z, orange), reverse transcriptase (RT), and cysteine-rich RNA binding domains (Cys, yellow). The GENEWRITE protein replaces the EN domain with a Cas protein (Cas9 or Cas12a/Cpf1, blue) and includes an N-terminal EGL13 nuclear localization signal (NLS, green), C-terminal c-Myc NLS (NLS, green), and 6xHis tag for in vitro purification (His, gray). (**B**) GENEWRITE components. The system consists of the GENEWRITE protein and a DNA target for insertion. A guide sgRNA complementary to the desired cut site (red) and a payload RNA encoding the desired insertion with a 3' end designed to hybridize to the insertion target (green). Optionally, as described in the text, NHEJ proteins, ORF1p protein, and 5' homology on the payload RNA to the target site can be included to increase insertion efficiency. (**C**) The sgRNA directs Cas cleavage to the integration site. (**D**) After Cas-induced cleavage, the 3' end of the payload RNA hybridizes with the cut site priming TPRT (**E**). After mRNA removal and second strand synthesis by host enzymes, the cut site is resolved (**F**).

CHAPTER 1: TARGETED INSERTION OF LARGE GENETIC PAYLOAD USING CAS DIRECTED LINE-1 REVERSE TRANSCRIPTASE

The work in this chapter is based on "Targeted Insertion of Large Genetic Payload Using Cas Directed LINE-1 Reverse Transcriptase", which was published in Scientific Reports Volume 11 Article Number 23625 on December 8, 2021. The authors of this paper are Femila Manoj, Laura W. Tai, Katelyn Sun Mi Wang, and Thomas E. Kuhlman who are affiliated with the Department of Physics at the University of California, Riverside. Katelyn was a student at Chaminade College Preparatory High School. Analysis and experiments are attributed to the authors who performed the work in their corresponding sections.

ABSTRACT

A difficult genome editing goal is the site-specific insertion of large genetic constructs. Here we describe the GENEWRITE system, where site-specific targetable activity of Cas endonucleases is coupled with the reverse transcriptase activity of the ORF2p protein of the human retrotransposon LINE-1. This is accomplished by providing two RNAs: a guide RNA targeting Cas endonuclease activity and an appropriately designed payload RNA encoding the desired insertion. Using *E. coli* as a simple platform for development and deployment, we show that with proper payload design and co-expression of helper proteins, GENEWRITE can enable insertion of large genetic payloads to precise locations, although with off-target effects, using the described approach. Based upon these results, we describe a potential strategy for implementation of GENEWRITE in more complex systems.

INTRODUCTION

"Cuz it's not like your PhD is on the line or anything." - Kira Stout

Discovered as a bacterial immune system against foreign genetic elements such as phages, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) Associated Proteins (Cas) are endonucleases that target and cleave DNA sequences based upon their homology with a "guide RNA"[1,2]. Consequently, by providing an engineered "single-guide" RNA (sgRNA), Cas enzymes can be targeted to cleave any desired sequence. This flexibility in gene editing by CRISPR-Cas endonucleases has revolutionized genome editing in a wide variety of organisms and in its application to the clinical therapeutics[3–24].

Despite their flexibility and ease of use, the repertoire of genome editing modalities that CRISPR/Cas systems allow remains limited. Knockout of point mutants can be generated relatively easily by targeting Cas cleavage to coding or control regions of the genome. The cell must repair such cuts to survive, and errors introduced by the nonhomologous end joining (NHEJ) repair machinery can lead to inactivation of control regions or introduction of missense or point mutations to coding sequences[25–28]. An additional editing modality is to introduce novel sequences to the genome through Homology Directed Repair (HDR), where a DNA fragment with ends homologous to the sequences flanking the cut site and containing the desired sequence to be inserted is introduced to the cell along with the Cas-sgRNA ribonucleoprotein (RNP) complexes. After cleavage, the fragment is then used to repair the cut by the cell's homologous

recombination repair machinery, resulting in its integration. However, HDR remains inefficient and difficult to accomplish, particularly for gene-sized or larger (≥ ~ 1 kilobase pair [kbp]) fragments[29–32]. A primary reason for this difficulty is that for HDR to be successful, non-homologous end joining (NHEJ) DNA repair, the primary repair mechanism for DNA repair in advanced eukaryotic cells, must be suppressed[33–39].

Here we introduce a method for the active insertion of lengthy genetic sequences into host DNA called GENEWRITE: Genome Engineering With RNA-Integrating Targetable Endonucleases. This is accomplished by coupling the targetable endonuclease activity of Cas enzymes to the reverse transcriptase activity of the human retrotransposon LINE-1 through translationally fusing Cas and LINE-1 reverse transcriptase proteins (Fig. 1.1A). Several recent reports have described approaches coupling the targetability of Cas enzymes with the activity of other transposons or reverse transcriptases. These include Tn7-like transposons whose genomic insertion is accomplished through an associated CRISPR-effector, from the cyanobacterium *Scytonema hofmanni* and *Vibrio cholerae*[40,41]. Insertion of these 2 -3 kbp bacterial transposons is programmable to the specific genomic locations in *Escherichia coli* through a guide RNA like other Cas enzymes. Another approach, prime editing, fuses a catalytically impaired Cas9 fused to an engineered Moloney Murine Leukemia Virus (M-MLV) reverse transcriptase, using a "prime editing guide RNA" (pegRNA) to target short insertions, deletions and all types of point mutations into human cells [42–46]. GENEWRITE offers functionality that is distinct from each of the examples. While prime editing similarly uses a reverse transcriptase to insert RNA-encoded sequences into the genome,

insertions performed with prime editing are typically limited to short 10 - 40 bp epitopes. Conversely, we illustrate the site-specific reverse transcription and insertion of ~ 1.5 kbp payload RNAs, larger than that offered by prime editing.

Previous studies have shown that reverse transcription by the LINE-1 protein ORF2P can be directed to pre-existing nicks and cuts in targeted DNA sequences in vitro, and we have previously shown that LINE-1 is functional in *E. coli*, particularly when complemented by expression of enzymes for NHEJ repair[47,48]. Here we use E. coli expressing B. subtilis NHEJ enzymes as a simple platform to optimize design and delivery of GENEWRITE for future application to more complex systems. The strategy used here for integration using GENEWRITE is shown in Fig. 1.1B–F. *E. coli* cells expressing GENEWRITE are transformed with two RNAs: a guide sgRNA to target Cas cleavage to the desired integration site, and a payload RNA carrying the coding sequence of the desired integration. The 3' end of the payload RNA is designed to be homologous to the bottom DNA strand downstream from the cut such that RNA–DNA hybridization occurs to prime reverse transcription. Host enzymes complete second strand synthesis and payload RNA removal, and the insert is sealed into the site[48]. We illustrate that GENEWRITE can be used to effectively target insertion of large, gene-sized payloads to specific locations, although not without off-target effects.

MATERIALS AND METHODS

Reagents

Primers used for PCR and RNA synthesis were synthesized by Integrated DNA Technologies (IDT, Coralville, IA). Kits used include QIAprep Spin Miniprep Kit (QIAGEN; Germantown, MD; Catalog Number 27106), QIAquick PCR Purification Kit (QIAGEN; Catalog Number 28106), DNeasy UltraClean Microbial Kit (QIAGEN; Catalog Number 12224-50), Megascript T7 Transcription Kit (ThermoFisher Scientific; Waltham, MA; Catalog Number AMB13345), TURBO DNase (ThermoFisher Scientific; Catalog Number AM2239), and NEBNext Ultra II Library Prep kit (New England Biosciences; Ipswich, MA; Catalog Number E7645S). PCR was performed with Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB; M0531L).

Biological resources and media

E. coli BL21-AI (ThermoFisher Scientific; Catalog Number C607003, GenBank accession number CP047231) was used for all experiments. Overnight seed cultures were grown in Super Optimal Broth with Catabolite Repression [SOC; SOB + 0.5% w/v glucose] medium with appropriate antibiotics. Electrocompetent cells were prepared by growth in Super Optimal Broth (SOB) with appropriate antibiotics.

Plasmid design and construction

All GENEWRITE proteins and variants were designed in Vector NTI software (Thermo Fisher Scientific) and synthesized de novo and cloned into pUC57-*kan* by GENEWIZ Gene synthesis (GENEWIZ); the exception is ORF2pZRT, which was cloned into pUC57-*amp* by GENEWIZ. A list of all constructs used in this study is found in Supplementary Table 1.

Bacillus subtilis NHEJ enzymes [Ku (encoded by the gene *ykoV*) and LigD (encoded by the gene *ykoU*)] were expressed from the anhydrotetracycline-inducible P_{LtetO1} promoter on the plasmid pZA31[48,49]. Cells not expressing NHEJ were transformed with empty pZA31 as a control.

sgRNA and payload RNA synthesis

DNA encoding sgRNAs were prepared using primers including a T7 promoter driving a 20 bp guide sequence. The 3' end of this primer was designed with a 14 bp overhang homologous to a 77 bp scaffold oligo containing sequence encoding the necessary sgRNA secondary structure and used to prime amplification of the sgRNA-encoding DNA. Sequences of oligos used in the study are available in Supplementary Table 2.

Payload RNAs were prepared using primers including a T7 promoter driving sequence encoding a strong, constitutive PlacIQ1 promoter, a Shine-Dalgarno ribosomal binding site, and 20 bp of sequence homologous to the spectinomycin resistance gene

aadA. Reverse primers were designed with 20 bp homology to the 3' end of *aadA* and included indicated lengths of sequence homologous to the intended integration site. Payload RNA-encoding DNA was amplified from the plasmid pTKRED using these primers[50].

RNAs were generated using the above DNA templates using T7 MEGAscript (Thermo Fisher Scientific), with incubation at 37 °C for 16 h. Samples were then digested with TURBO DNase (Thermo Fisher Scientific) for 1 h at 37 °C and purified by phenol–chloroform extraction and isopropanol precipitation.

Preparation of electrocompetent cells and transformation

Electrocompetent cells were prepared by preparation of a seed culture by overnight growth in SOC at 37 °C in a shaking water bath (New Brunswick C76). Seed cultures were diluted 10:1 in fresh SOB and grown at 37 °C in a shaking water bath until OD600 ~ 0.6, at which point 0.1% L-arabinose was added. Importantly, after L-arabinose was completely dissolved, cells were immediately harvested by centrifugation at 4 °C; extended induction of GENEWRITE with L-arabinose is lethal. This was followed by 3 × washing in ice-cold 10% v/v glycerol. 100 µl of cells thus prepared were mixed with an excess of payload RNA and sgRNA (5 µg and 10 µg, respectively); these quantities yielded success but have not been optimized. The mixture was electroporated (BIO-RAD Gene Pulser) using standard settings for *E. coli*. 1 ml of SOC + 1 mM IPTG and 100 ng/ml aTc were added, and cells were allowed to recover overnight. Transformants were

spread on LB agar plates containing 100 μg/ml spectinomycin, 0.5% w/v glucose, 1 mM IPTG, and 100 ng/ml aTc and then incubated overnight in a 37 °C air incubator.

Genome sequencing

Genomic DNA was obtained from cultures prepared in 2 ml Lysogeny Broth (LB) by purification using the QIAGEN DNeasy UltraClean Microbial Kit. Resulting samples were submitted to the UCR Genomic Core at the Institute for Integrative Genomic Biology for processing and sequencing. Samples were sheared using a Covaris S220 Ultrasonicator and libraries prepared using an NEBNext Ultra II Library Prep kit. After preparation, libraries were analyzed using qPCR and an Agilent 2100 Bioanalyzer. The resulting libraries were pooled and sequenced using an Illumina MiSeq sequencer with 150 bp paired end reads. Sequencing data were analyzed with Geneious Prime.

RESULTS

GENEWRITE rationale and design

Contributing Authors: TEK

The human retrotransposon LINE-1 (Long Interspersed Nuclear Element, or L1) encodes the two proteins ORF1p and ORF2p, and both proteins are required for efficient retrotransposition in humans. A primary function of ORF1p appears to be chaperone activity, while ORF2p includes endonuclease (EN) and reverse transcriptase (RT) domains[51]. To retrotranspose, ORF2p EN nicks TA-rich target DNA, and the 3' end of the LINE-1 mRNA hybridizes with DNA adjacent to the nick to initiate reverse

transcription through a process called target primed reverse transcription (TPRT)[52]. In most active L1 elements, this hybridization is facilitated through the presence of a ~ 100 bp long poly(A) tract, which is also thought to be the primary binding target of ORF2p to its encoding mRNA[53].

LINE-1 and its accessory proteins naturally exist in human cells, making it an appealing target for optimization as a genome editing tool. To attempt to further enhance specifically targeted reverse transcribed insertions by ORF2p in vivo, we removed the promiscuous ORF2p EN domain by deleting amino acids 1–347. The remaining fragment, from amino acids 348–1275, which includes the Z, RT, and cysteine-rich RNA-binding domains, we dub ORF2pZRT. Finally, the GENEWRITE protein consists of a translational fusion of ORF2pZRT to targetable Cas endonucleases (Cas9 or Cas12a/Cpf1) with a flexible linker. In addition, the GENEWRITE protein includes N and C-terminal nuclear localization signals (NLS) and a C-terminal 6xHis tag to enable purification (Fig. 1.1A). A previous similar attempt at replacing ORF2p EN with Cas9 and using *Alu*-like payload RNA to target ORF2p RT to specific loci in human cells proved unsuccessful[54]. As described below, we have made several refinements to the GENEWRITE system relative to this attempt, including the use of *Escherichia coli* as a simpler in vivo platform in which to test and optimize. We additionally show that the 10 base pair homologies between target and payload used in this previous study is likely inadequate for priming of TPRT.

High expression of GENEWRITE protein is lethal to *E. coli*

Contributing Authors: LWT, KSW, TEK

We designed and synthesized the GENEWRITE protein under control of a T7 promoter, which was cloned into the plasmid pUC57-*kan*. We transformed this plasmid into *E. coli* strain BL21-AI, along with either empty plasmid pZA31, or pZA31 carrying *ykoU* and *ykoV B. subtilis* NHEJ enzymes expressed from P_{LtetO1}[48,49]. In strain BL21-AI, GENEWRITE expression is inducible by the addition of L-arabinose. Curiously, while expression of Cas9/12a, ORF2pZRT, or both Cas9/12a and ORF2pZRT in individual *E. coli* cells does not affect growth, strong expression of the GENEWRITE Cas-ORF2pZRT fusion protein induced through the addition of arabinose is lethal to *E. coli*. This lethality is partially relieved by simultaneous expression of *B. subtilis* NHEJ enzymes. This suggests lethality may be a consequence of genomic breaks generated by GENEWRITE, perhaps driven by high affinity of ORF2p to arbitrary RNAs in vivo[55]. Consequently, the results described below rely upon low, leakage levels of expression of GENEWRITE without induction.

GENEWRITE is effective at insertions into high-copy number targets in *E. coli*

Contributing Authors: FLM, TEK

We expected the strategy outlined in Fig. 1.1B–F to be difficult to successfully execute for a number of reasons, including the expected difficulty of co-transforming individual cells with appropriate amounts of both sgRNA and payload RNA, as well as previously documented preference of ORF2p to act primarily upon its *cis*-encoding

RNA[56]. Hence, as an initial integration target, we chose the high copy number plasmid pUC57-*kan* [~ 500–1000 /cell] from which the GENEWRITE protein itself is expressed to maximize chances of success. For experiments described here, the ~ 1200 bp payload RNA consisted of an *aadA* spectinomycin resistance gene driven by a strong, constitutive *lacI*Q1 promoter and Shine-Dalgarno ribosomal binding site (RBS)[57]. Consequently, after the GENEWRITE protocol, cells were spread on plates containing spectinomycin to select potentially successful integrants.

Based upon our current understanding of TPRT, design of the payload RNA 3' hybridization region is critical. To determine the optimal length of the hybridization region, we generated an array of six identical payload RNAs with hybridization length variable from 0 to 50 bp in 10 bp increments. Based on prior reports of the essentiality of a 3' poly(A) tract for ORF2p binding and reverse transcription, we generated a second array of payload RNAs, identical to the first, but also including the 30 bp poly(A) tract found in the SINE element AluYA5[53,58].

We transformed the pUC57-targeting sgRNA along with each payload RNA into *E. coli* weakly expressing GENEWRITE-Cas9, either with or without simultaneous expression of *B. subtilis* NHEJ enzymes. The results are shown in Fig. 1.2. For those payload RNAs containing a poly(A) tract, we observed very few spectinomycin resistant colonies, for both with or without simultaneous co-expression of NHEJ. Conversely, without the poly(A) tract, we obtained hundreds of spectinomycin resistant colonies when complemented with co-expression of NHEJ (Fig. 1.2A). Site-specific integration was

verified by PCR using primers that amplified across the 5' and 3' integration junctions (Fig. 1.2B); 63 out of 96 colonies screened yielded a positive signal for a success rate of ~ 72% (Fig. 1.2C). Sequencing of eight purified plasmids revealed some small deletions at the 5' end of the insertion (Fig. 1.5). From these experiments, we conclude that optimal design of the payload RNA includes 40–50 bp of 3' homology to the intended target facilitated by NHEJ DNA repair, and with no poly(A) tract. This difference in essentiality of the poly(A) tract to TPRT between *E. coli* and humans may be the result of mRNA 3' poly(A) tails stabilizing RNAs in eukaryotes, while poly(A) tails designate mRNAs for degradation in bacteria[59–62].

We performed a series of controls and further investigations using the payload designed to target pUC57-*kan* with 40 bp 3' hybridization region (Fig. 1.2D): (1) as expected, the sgRNA is required for efficient targeting and integration; (2) NLS sequences at the N- and C-termini do not significantly interfere with function; (3) the Cas12a/Cpf1 GENEWRITE variant is functional, although with lower efficiency than the Cas9 variant, consistent with previous findings that blunt-end cuts fragments serve as better TPRT substrates than those with 3' or 5' overhangs; and (4) simultaneous co-expression of unfused Cas9 and ORF2pZRT, rather than the translationally-fused GENEWRITE protein, is not functional[47]. However, LINE-1 reverse transcriptase has been shown to function even when encoded and expressed separately from the endonuclease through association via the naturally occurring cryptic Z domain, raising the possibility of potentially using naturally expressed LINE-1 in the human genome as an editing tool.

GENEWRITE can insert payloads into low and single copy targets

Contributing Authors: FLM, TEK

For the next target, we attempted insertion into the much lower copy number pZA31 plasmid hosting the NHEJ genes (~ 20–30 copies/cell)[49]. Using 40 bp of 3' homology to target as described above, we obtained no colonies when transforming payload and sgRNAs into cells expressing GENEWRITE but deficient in NHEJ. However, we obtained ~ 50 colonies on average when transforming into cells expressing both GENEWRITE and NHEJ proteins. PCR screening of putative positive colonies generated a positive signal in 10 out of 50 colonies (Fig. 1.3A), yielding a success rate of 20%.

We finally attempted to use GENEWRITE to site-specifically insert a payload into single copy chromosomal loci. We attempted insertions at three loci we have previously shown to accept insertions at high efficiency using recombineering-like methods: the *nth* locus near the terminus of replication; the *atpI* locus near the origin of replication; and the *ybbD* locus midway on the right replichore (Fig. 1.3B)[63,64]. In these cases, repeated attempts at the GENEWRITE protocol as described above were unsuccessful. Prior reports and our own studies of retrotransposition of native LINE-1 in *E. coli* (Fig. 1.6) suggest that homology between the 5' end of the payload and insertion location may also aid in targeting[65]. Consequently, we attempted two strategies: (1) inclusion of 20 bp of 5' homology between the payload and the targeted insertion site; and (2) simultaneous co-expression of ORF1p. Each of these strategies alone was unsuccessful. Only when targeting *nth* by including 20 bp of 5' and 40 bp of 3' payload

homology to the target, along with simultaneous co-expression of ORF1p, did we obtain significant numbers of colonies after transformation (~ 20 colonies on average). Under these conditions, PCR screening of 50 positive colonies (Fig. 1.3C, Fig. 1.7) demonstrates a success rate of 60%. However, repeated attempts of insertion at the *atpI* and *ybbD* sites with ORF1p co-expression and 20 bp of 5' and 40 bp of 3' payload homology to target have so far proven unsuccessful. As with 3' homology, further optimization of the amount of 5' homology to the target included in the payload may improve the efficiency of insertion at low copy number targets. Inclusion of homology in the 5' end of the payload, with the same sequence as the sgRNA, suggests the possibility of using a single RNA as both guide and payload. However, attempting to include the necessary secondary structure and using the 5' end of the payload itself as the sgRNA for the Cas component proved unsuccessful, and we found it was necessary to co-transform two separate sgRNA and payload RNAs for successful targeted integration.

Off-target effects and application to complex organisms

Contributing Authors: FLM, TEK

Whole genome sequencing of cells subjected to the GENEWRITE expression exhibit larger numbers of high frequency mutations relative to a negative control, with mutations scattered throughout the genome (Fig. 1.4). Moreover, a large fraction of the plasmids purified and sequenced from GENEWRITE-exposed cells have curiously had the coding sequences of both GENEWRITE and NHEJ proteins excised from their host plasmids (Fig. 1.8), suggesting that GENEWRITE may also be effective in excising coding regions of inappropriately-highly expressed genes.

DISCUSSION

We have shown that a fusion of a Cas endonuclease and LINE-1 ORF2p reverse transcriptase, which we call GENEWRITE, is capable of integration of large genetic payloads in the *E. coli* genome through appropriate design of the homology regions of guide and payload RNAs. We also find that assistance from NHEJ DNA repair enzymes and LINE-1 ORF1p protein may help increase the efficiency and specificity of the insertion (results summarized in Table 1.3). We have not yet tested the limits on size of GENEWRITE payloads, but LINE-1 itself is ~ 5 kbp long and hence similarly sized payloads may be accessible.

The above-described results were obtained using a simplistic method where each component is delivered separately: the RNAs through electroporation, and the GENEWRITE protein through constitutive, low-level expression from a plasmid. We find GENEWRITE to be remarkably successful given this simplistic approach, despite ORF2p's *cis*-preference for its encoding RNA and propensity to produce inserts with 5' truncations[56,65–68]. However, using this method, we find significant off-target effects, including an increase in the rate of off-target mutations relative to a control, and the excision of highly expressed DNA segments from the genome. We speculate that these off-target effects and the lethality of the GENEWRITE protein to *E. coli* may be coupled: high affinity of ORF2p to arbitrary RNAs may force non-sgRNAs into the Cas component, serving as a guide for endonuclease activity and generating off-target DNA breaks. Hence, we suggest that rather than direct in vivo expression, deployment of the

GENEWRITE system to more complex mammalian cells may be better accomplished through in vitro assembly and transfection of RNP particles as is frequently performed with traditional CRISPR-Cas editing. Furthermore, it remains to be seen if GENEWRITE insertions will produce truncated insertions when applied to more advanced systems.

REFERENCES

1. Singh, D. *et al.* Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). *Proc. Natl. Acad. Sci. U. S. A.* **115**, 5444–5449 (2018).

2. Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat. Commun.* **7**, 12778 (2016).

3. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).

4. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

5. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).

6. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).

7. DiCarlo, J. E. *et al.* Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.* **41**, 4336–4343 (2013).

8. Zhang, G.-C. *et al.* Construction of a quadruple auxotrophic mutant of an industrial polyploid saccharomyces cerevisiae strain by using RNA-guided Cas9 nuclease. *Appl. Environ. Microbiol.* **80**, 7694–7701 (2014).

9. Liu, J.-J. *et al.* Metabolic Engineering of Probiotic Saccharomyces boulardii. *Appl. Environ. Microbiol.* **82**, 2280–2287 (2016).

10. Vyas, V. K., Barrasa, M. I. & Fink, G. R. A CRISPR system permits genetic engineering of essential genes and gene families. *Sci Adv* **1**, e1500248 (2015).

11. Ng, H. & Dean, N. Dramatic Improvement of CRISPR/Cas9 Editing in by Increased Single Guide RNA Expression. *mSphere* **2**, (2017).

12. Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).

13. Gratz, S. J. *et al.* Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics* **194**, 1029–1035 (2013).

14. Bassett, A. R., Tibbit, C., Ponting, C. P. & Liu, J.-L. Highly efficient targeted mutagenesis of Drosophila with the CRISPR/Cas9 system. *Cell Rep.* **4**, 220–228 (2013).

15. Yan, H. *et al.* An Engineered orco Mutation Produces Aberrant Social Behavior and Defective Neural Development in Ants. *Cell* **170**, 736–747.e9 (2017).

16. Trible, W. *et al.* orco Mutagenesis Causes Loss of Antennal Lobe Glomeruli and Impaired Social Behavior in Ants. *Cell* **170**, 727–735.e10 (2017).

17. Kistler, K. E., Vosshall, L. B. & Matthews, B. J. Genome engineering with CRISPR-Cas9 in the mosquito Aedes aegypti. *Cell Rep.* **11**, 51–60 (2015).

18. Friedland, A. E. *et al.* Heritable genome editing in C. elegans via a CRISPR-Cas9 system. *Nat. Methods* **10**, 741–743 (2013).

19. Jiang, W. *et al.* Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Res.* **41**, e188 (2013).

20. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).

21. Soni, D. *et al.* Deubiquitinase function of A20 maintains and repairs endothelial barrier after lung vascular injury. *Cell Death Discov* **4**, 60 (2018).

22. Guo, X. & Li, X.-J. Targeted genome editing in primate embryos. *Cell Res.* **25**, 767–768 (2015).

23. Baltimore, D. *et al.* Biotechnology. A prudent path forward for genomic engineering and germline gene modification. *Science* **348**, 36–38 (2015).

24. Xu, L. *et al.* CRISPR-Edited Stem Cells in a Patient with HIV and Acute Lymphocytic Leukemia. *N. Engl. J. Med.* **381**, 1240–1247 (2019).

25. Guo, T. *et al.* Harnessing accurate non-homologous end joining for efficient precise deletion in CRISPR/Cas9-mediated genome editing. *Genome Biol.* **19**, 170 (2018).

26. Bothmer, A. *et al.* Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nat. Commun.* **8**, 13905 (2017).

27. Brinkman, E. K. *et al.* Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks. *Mol. Cell* **70**, 801–813.e6 (2018).

28. van Overbeek, M. *et al.* DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Mol. Cell* **63**, 633–646 (2016).

29. Nambiar, T. S. *et al.* Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant. *Nat. Commun.* **10**, 3395 (2019).

30. Aird, E. J., Lovendahl, K. N., St Martin, A., Harris, R. S. & Gordon, W. R. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Commun Biol* **1**, 54 (2018).

31. Liu, M. *et al.* Methodologies for Improving HDR Efficiency. *Front. Genet.* **9**, 691 (2018).

32. Rozov, S. M., Permyakova, N. V. & Deineko, E. V. The Problem of the Low Rates of CRISPR/Cas9-Mediated Knock-ins in Plants: Approaches and Solutions. *Int. J. Mol. Sci.* **20**, (2019).

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

34. Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat. Rev. Mol. Cell Biol.* **4**, 712–720 (2003).

35. Davis, A. J. & Chen, D. J. DNA double strand break repair via non-homologous end-joining. *Transl. Cancer Res.* **2**, 130–143 (2013).

36. Iliakis, G. Backup pathways of NHEJ in cells of higher eukaryotes: cell cycle dependence. *Radiother. Oncol.* **92**, 310–315 (2009).

37. Devkota, S. The road less traveled: strategies to enhance the frequency of homology-directed repair (HDR) for increased efficiency of CRISPR/Cas-mediated transgenesis. *BMB Rep.* **51**, 437–443 (2018).

38. Li, G. *et al.* Small molecules enhance CRISPR/Cas9-mediated homology-directed genome editing in primary cells. *Sci. Rep.* **7**, 8943 (2017).

39. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–548 (2015).

40. Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).

41. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).

42. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

43. Baranauskas, A. *et al.* Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants. *Protein Eng. Des. Sel.* **25**, 657–668 (2012).

44. Gerard, G. F. *et al.* The role of template-primer in protection of reverse transcriptase from thermal inactivation. *Nucleic Acids Res.* **30**, 3118–3129 (2002).

45. Arezi, B. & Hogrefe, H. Novel mutations in Moloney Murine Leukemia Virus reverse transcriptase increase thermostability through tighter binding to template-primer. *Nucleic Acids Res.* **37**, 473–481 (2009).

46. Kotewicz, M. L., Sampson, C. M., D'Alessio, J. M. & Gerard, G. F. Isolation of cloned Moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. *Nucleic Acids Res.* **16**, 265–277 (1988).

47. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21**, 5899–5910 (2002).

48. Lee, G. *et al.* Testing the retroelement invasion hypothesis for the emergence of the ancestral eukaryotic cell. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12465–12470 (2018).

49. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).

50. Kuhlman, T. E. & Cox, E. C. Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Res.* **38**, e92 (2010).

51. Martin, S. L. The ORF1 protein encoded by LINE-1: structure and function during L1 retrotransposition. *J. Biomed. Biotechnol.* **2006**, 45621 (2006).

52. Lambowitz, A. M. & Belfort, M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* **3**, MDNA3–0050–2014 (2015).

53. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y. & Moran, J. V. A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol. Cell* **60**, 728–741 (2015).

54. Ade, C. M. *et al.* Evaluating different DNA binding domains to modulate L1 ORF2p-driven site-specific retrotransposition events in human cells. *Gene* **642**, 188–198 (2018).

55. Piskareva, O., Ernst, C., Higgins, N. & Schmatchenko, V. The carboxy-terminal segment of the human LINE-1 ORF2 protein is involved in RNA binding. *FEBS Open Bio* **3**, 433–437 (2013).

56. Wei, W. *et al.* Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* **21**, 1429–1439 (2001).

57. Calos, M. P. & Miller, J. H. The DNA sequence change resulting from the IQ1 mutation, which greatly increases promoter strength. *Mol. Gen. Genet.* **183**, 559–560 (1981).

58. Konkel, M. K. *et al.* Sequence Analysis and Characterization of Active Human Alu Subfamilies Based on the 1000 Genomes Pilot Project. *Genome Biol. Evol.* **7**, 2608–2622 (2015).

59. Dreyfus, M. & Régnier, P. The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. *Cell* **111**, 611–613 (2002).

60. Sarkar, N. Polyadenylation of mRNA in bacteria. *Microbiology* **142 (Pt 11)**, 3125–3133 (1996).

61. Hajnsdorf, E. & Kaberdin, V. R. RNA polyadenylation and its consequences in prokaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, (2018).

62. Mohanty, B. K. & Kushner, S. R. Bacterial/archaeal/organellar polyadenylation. *Wiley Interdiscip. Rev. RNA* **2**, 256–276 (2011).

63. Kuhlman, T. E. & Cox, E. C. Gene location and DNA density determine transcription factor distributions in Escherichia coli. *Mol. Syst. Biol.* **8**, 610 (2012).

64. Tas, H., Nguyen, C. T., Patel, R., Kim, N. H. & Kuhlman, T. E. An Integrated System for Precise Genome Modification in Escherichia coli. *PLoS One* **10**, e0136963 (2015).

65. Zingler, N. *et al.* Analysis of 5’ junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res.* **15**, 780–789 (2005).

66. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

67. Babushok, D. V., Ostertag, E. M., Courtney, C. E., Choi, J. M. & Kazazian, H. H., Jr. L1 integration in a transgenic mouse model. *Genome Res.* **16**, 240–250 (2006).

68. Chen, J.-M., Férec, C. & Cooper, D. N. Mechanism of Alu integration into the human genome. *Genomic Med.* **1**, 9–17 (2007).

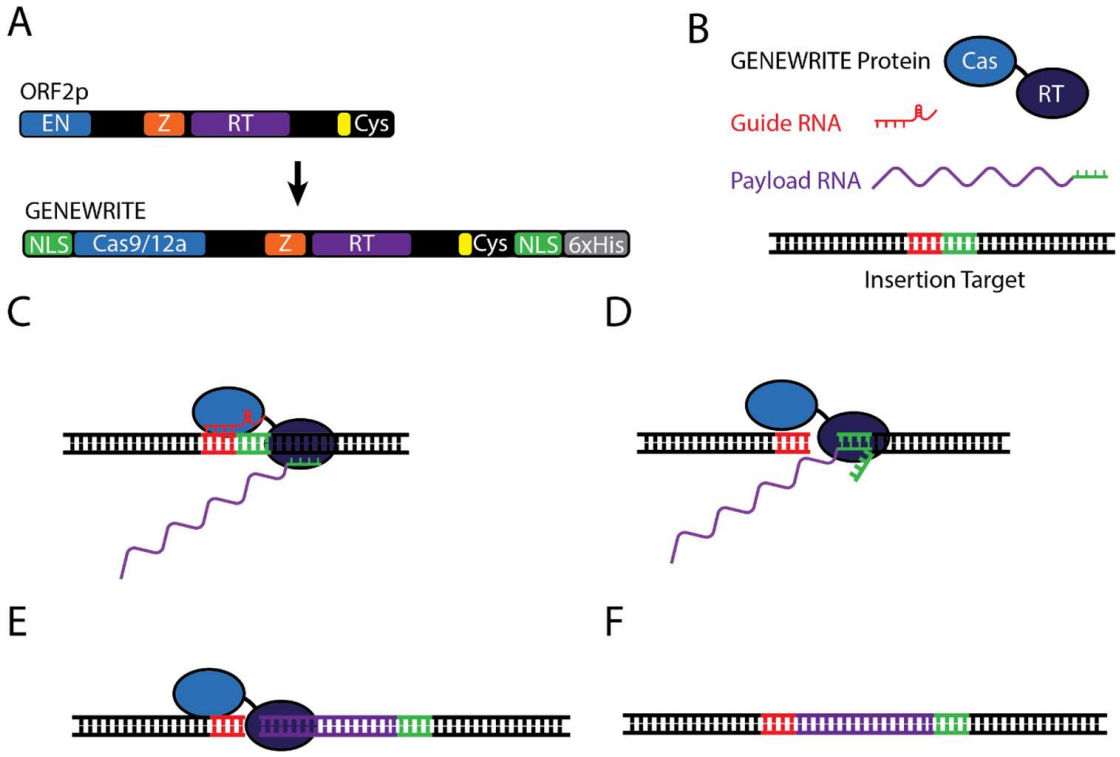

Figure 1.1. GENEWRITE components and strategy. (**A**) ORF2p and GENEWRITE domain structure. Wildtype ORF2p consists of endonuclease (EN, blue), Z (Z, orange), reverse transcriptase (RT), and cysteine-rich RNA binding domains (Cys, yellow). The GENEWRITE protein replaces the EN domain with a Cas protein (Cas9 or Cas12a/Cpf1, blue) and includes an N-terminal EGL13 nuclear localization signal (NLS, green), C-terminal c-Myc NLS (NLS, green), and 6xHis tag for in vitro purification (His, gray). (**B**) GENEWRITE components. The system consists of the GENEWRITE protein and a DNA target for insertion. A guide sgRNA complementary to the desired cut site (red) and a payload RNA encoding the desired insertion with a 3' end designed to hybridize to the insertion target (green). Optionally, as described in the text, NHEJ proteins, ORF1p protein, and 5' homology on the payload RNA to the target site can be included to increase insertion efficiency. (**C**) The sgRNA directs Cas cleavage to the integration site. (**D**) After Cas-induced cleavage, the 3' end of the payload RNA hybridizes with the cut site priming TPRT (**E**). After mRNA removal and second strand synthesis by host enzymes, the cut site is resolved (**F**).

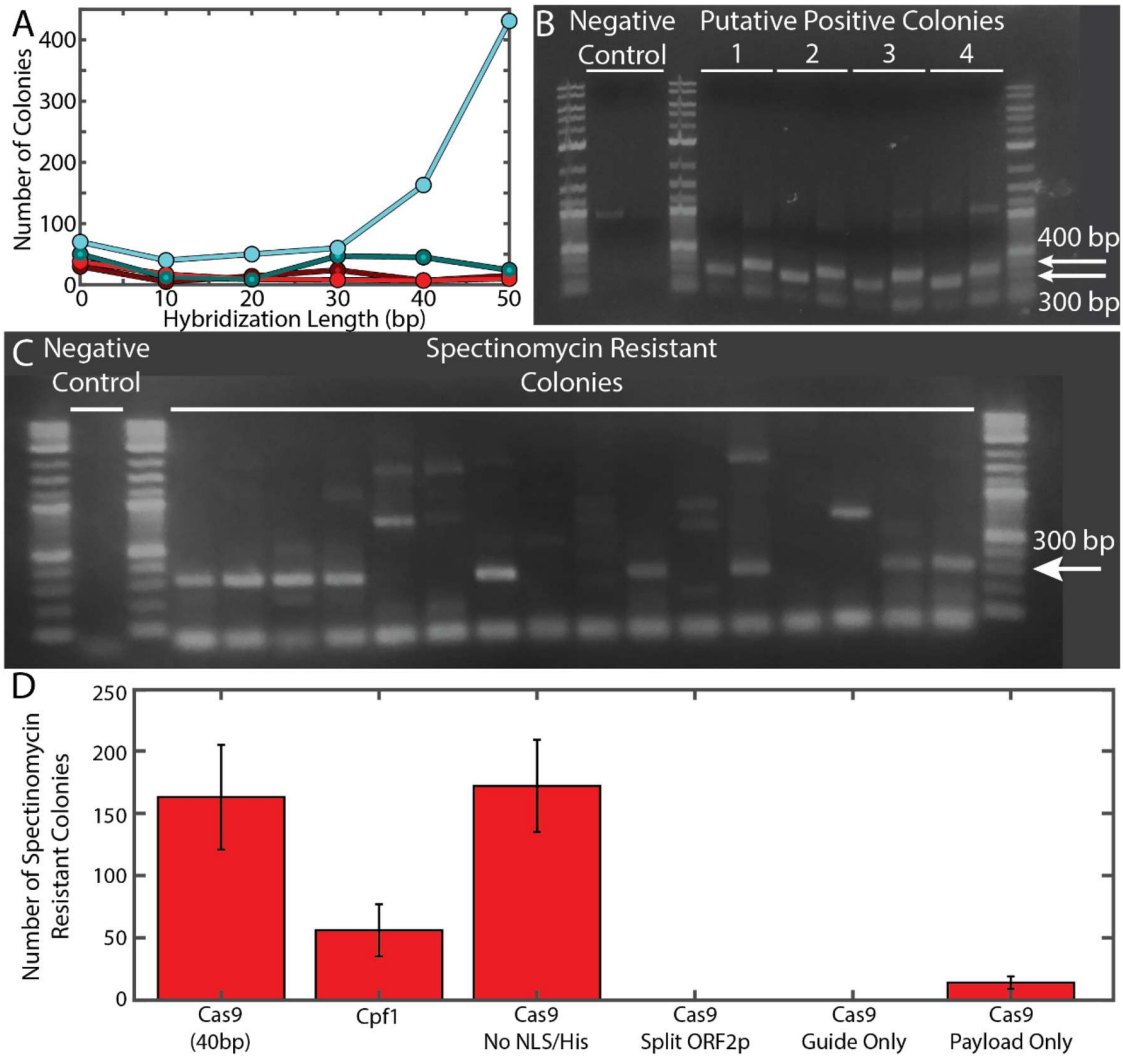
A
Number of Colonies
400
300
200
100
0
0
10
20
30
40
50
Hybridization Length (bp)
B
Negative Control
Putative Positive Colonies
1
2
3
4
400 bp
300 bp
C
Negative Control
Spectinomycin Resistant Colonies
300 bp
D
Number of Spectinomycin Resistant Colonies
250
200
150
100
50
0
Cas9 (40bp)
Cpf1
Cas9 No NLS/His
Cas9 Split ORF2p
Cas9 Guide Only
Cas9 Payload Only

Figure 1.2. GENEWRITE site-specific insertion in a high-copy number plasmid. sgRNA and payload RNAs were designed to integrate an *aadA* spectinomycin resistance gene into the plasmid pUC57-*kan*. (**A**) Number of spectinomycin resistant colonies as a function of payload RNA 3'hybridization length. Dark red: -NHEJ + poly(A); Dark cyan: -NHEJ -poly(A); Bright red: + NHEJ + poly(A); Bright cyan: + NHEJ -poly(A). Data points are the average of three replicates, error bars are SD. (**B**) PCR verification of integration. Lanes 1, 4, 13: NEB 1 kb Plus Ladder. 300 and 400 bp bands are indicated. Lanes 5–12: amplicons resulting from four spectinomycin resistant colonies. For each pair of lanes, the leftmost lane is PCR across the 5' junction (300 bp amplicon expected), rightmost is PCR across 3' junction (400 bp amplicon expected). (**C**) Representative screening of 16 randomly selected colonies by PCR across the 5' integration junction. (**D**) Controls and effect of various GENEWRITE components, relative to that of intact GENEWRITE-Cas9 with 40 bp homology payload RNA (first column). 2nd column: replacement of Cas9 with Cas12a/Cpf1; 3rd column: effect of removal of NLS and His tags; 4th column: simultaneous expression of unfused Cas9 and ORF2pZRT; 5th column: GENEWRITE-Cas9 transformed with only guide RNA but no payload; 6th column: GENEWRITE-Cas9 transformed with only payload RNA but no guide.

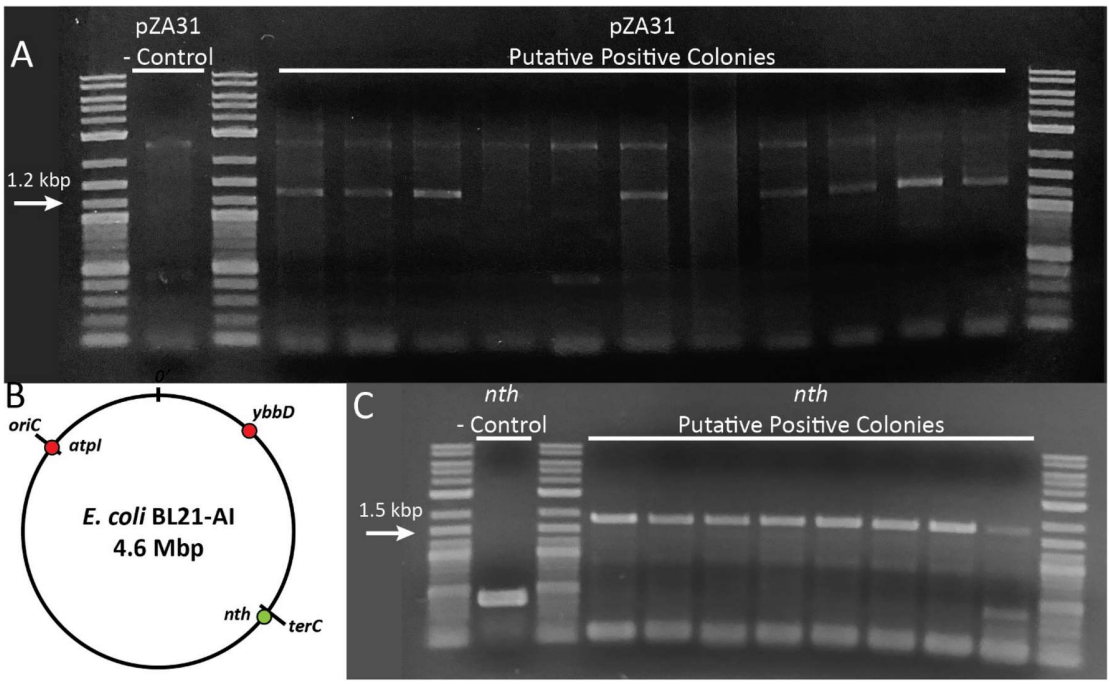

Figure 1.3. GENEWRITE site-specific insertion in low-copy number targets. (**A**) Insertion in low copy number (15–30 copies /cell) plasmid pZA31. Chosen primers bind at start of payload promoter and within the adjacent pZA31 sequence after 3' end of payload. Colony PCR was performed with 10% DMSO to eliminate extraneous non-specific amplification. Expected amplicon is ~ 1300 bp. (**B**) Attempted chromosomal insertion sites. Coordinates of sites are $x_{chromosome} = 0.0098$ (*atpI*), 0.5476 (*ybbD*), and -0.943 (*nth*), where $x = 0$ corresponds to *oriC* and $x = \pm 1$ corresponds to *terC*. (**C**) PCR amplification across *nth* integration location. Primers bind to chromosomal regions adjacent to targeted integration site. Amplicon expected from successful integration is ~ 1600 bp. We conservatively identify the last colony as negative for integration.

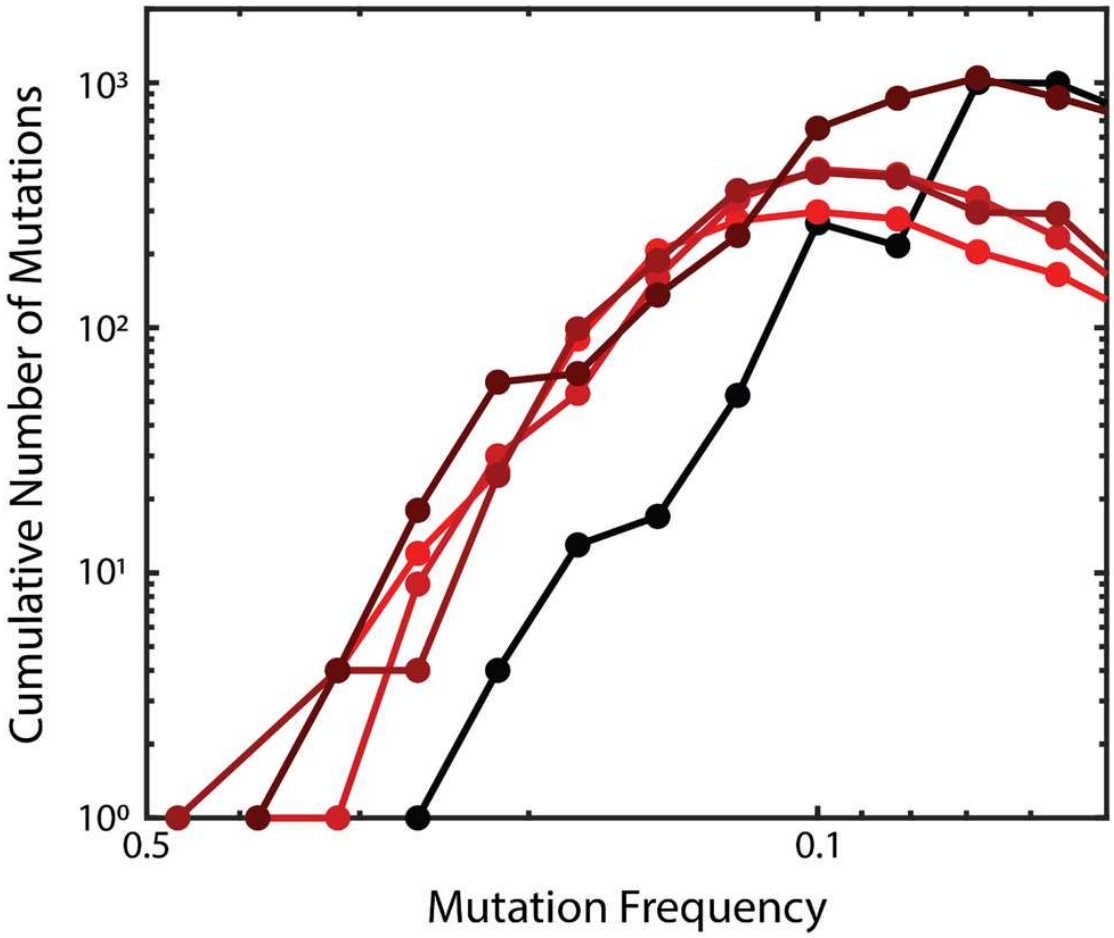

Figure 1.4. Cumulative number of mutations identified relative to the BL21-AI complete genome sequence (Accession NZ_CP047231.1, GI: 1797637028) for the BL21-AI negative control (black) and four replicates of BL21-AI subjected to the GENEWRITE expression.

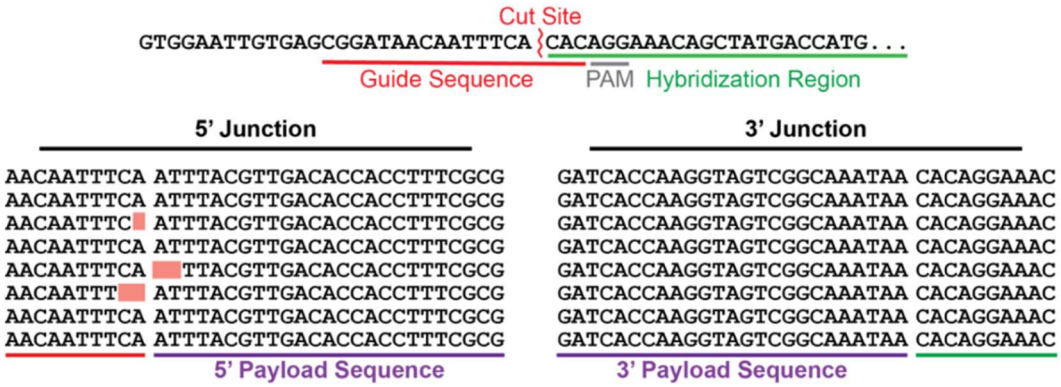

Figure 1.5. Sequencing of eight positive colonies with insertions in pUC57-kan. Top: Sequence of target site and design features. Note that guide sequence+PAM is destroyed upon successful integration. Bottom: Sequencing of eight positive clones, with mismatches highlighted in red.

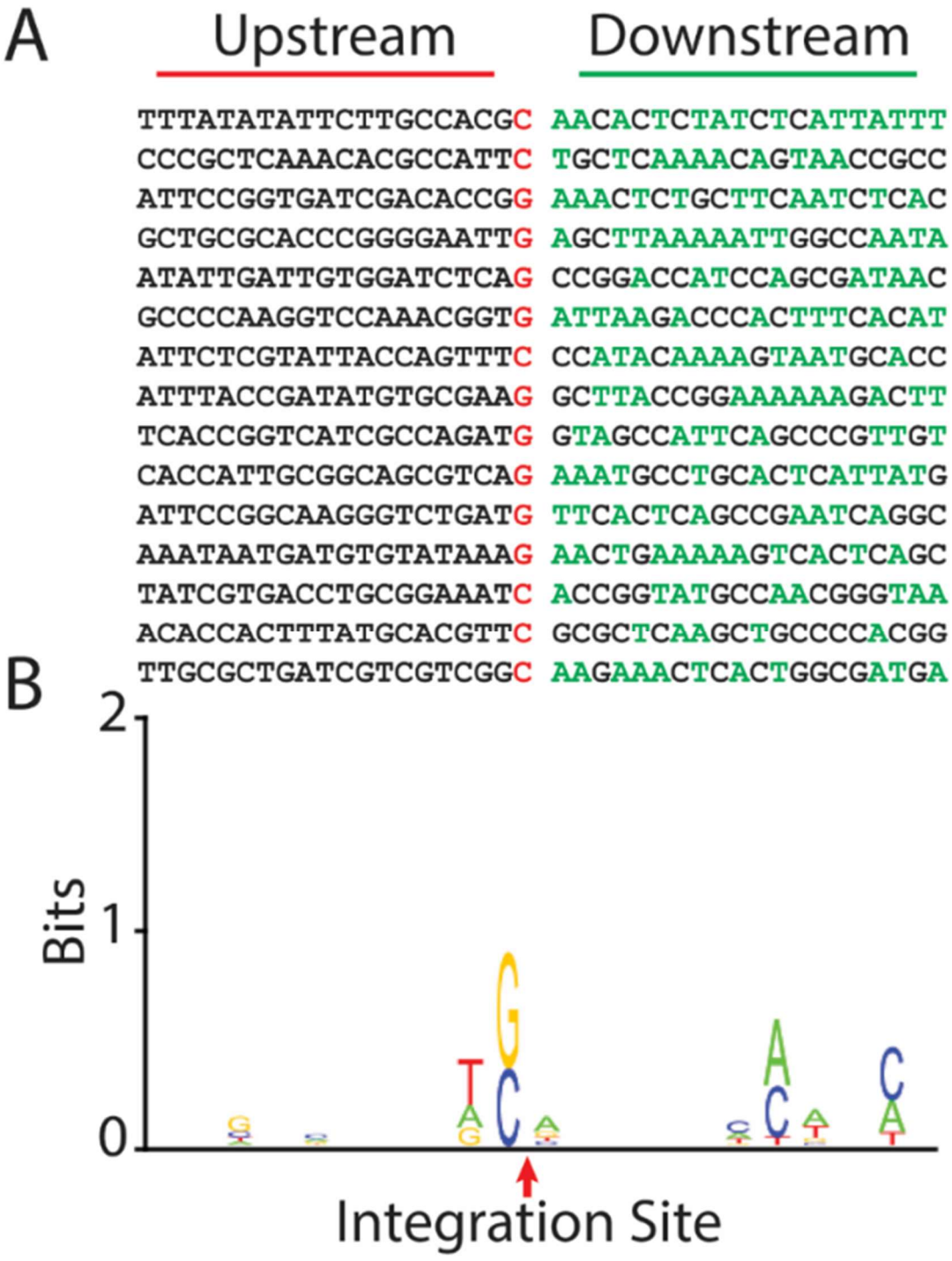
A
Upstream
Downstream
TTTATATATTCTTGCCACGC AACACTCTATCTCATTATTT
CCCGCTCAAACACGCCATTC TGCTCAAAACAGTAACCGCC
ATTCCGGTGATCGACACCGG AAACTCTGCTTCAATCTCAC
GCTGCGCACCCGGGGAATTG AGCTTAAAAATTGGCCAATA
ATATTGATTGTGGATCTCAG CCGGACCATCCAGCGATAAC
GCCCCAAGGTCCAAACGGTG ATTAAGACCCACTTTCACAT
ATTCTCGTATTACCAGTTTC CCATACAAAAGTAATGCACC
ATTTACCGATATGTGCGAAG GCTTACCGGAAAAAAGACTT
TCACCGGTCATCGCCAGATG GTAGCCATTCAGCCCGTTGT
CACCATTGCGGCAGCGTCAG AAATGCCTGCACTCATTATG
ATTCCGGCAAGGGTCTGATG TTCACTCAGCCGAATCAGGC
AAATAATGATGTGTATAAAG AACTGAAAAAGTCACTCAGC
TATCGTGACCTGCGGAAATC ACCGGTATGCCAACGGGTAA
ACACCACTTTATGCACGTTC GCGCTCAAGCTGCCCCACGG
TTGCGCTGATCGTCGTCGGC AAGAAACTCACTGGCGATGA
B
Bits
2
1
0
Integration Site

Figure 1.6. Insertion sites for LINE-1 retrotransposition in *E. coli*. 12 LINE-1 integration sites in *E. coli* identified by Illumina sequencing with 150 bp paired-end reads. (A) Sequences upstream and downstream of identified insertion locations. C/G immediately upstream of insertion is highlighted red, TA-rich regions downstream are highlighted green. (B) Logo plot of 20 bp surrounding integration site. Note G/C immediately upstream of the insertion site is the most prominent feature. In these experiment, LINE-1 was expressed in *E. coli* from a T7 promoter, and hence the first two basepairs at the 5' end of the transcript are GC.

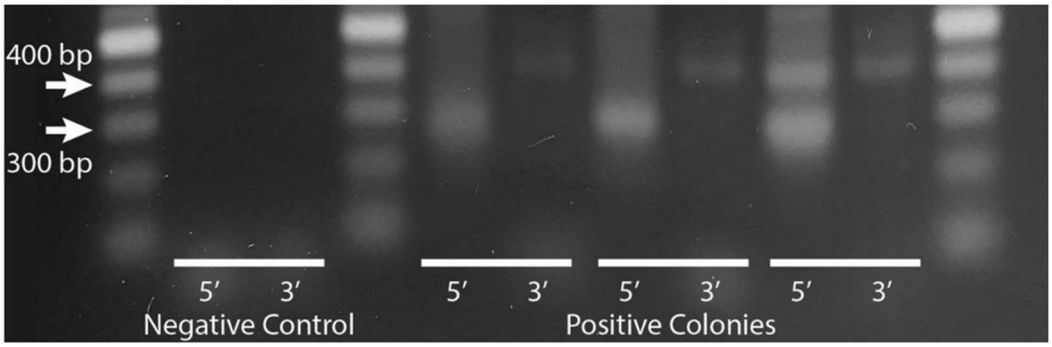

Figure 1.7. PCR amplification across junctions created by GENEWRITE insertion at the nth chromosomal locus with 2% agarose gel electrophoresis. Amplicons expected from amplification across the 5' junction is 290 bp, 3’ junction is 400 bp.

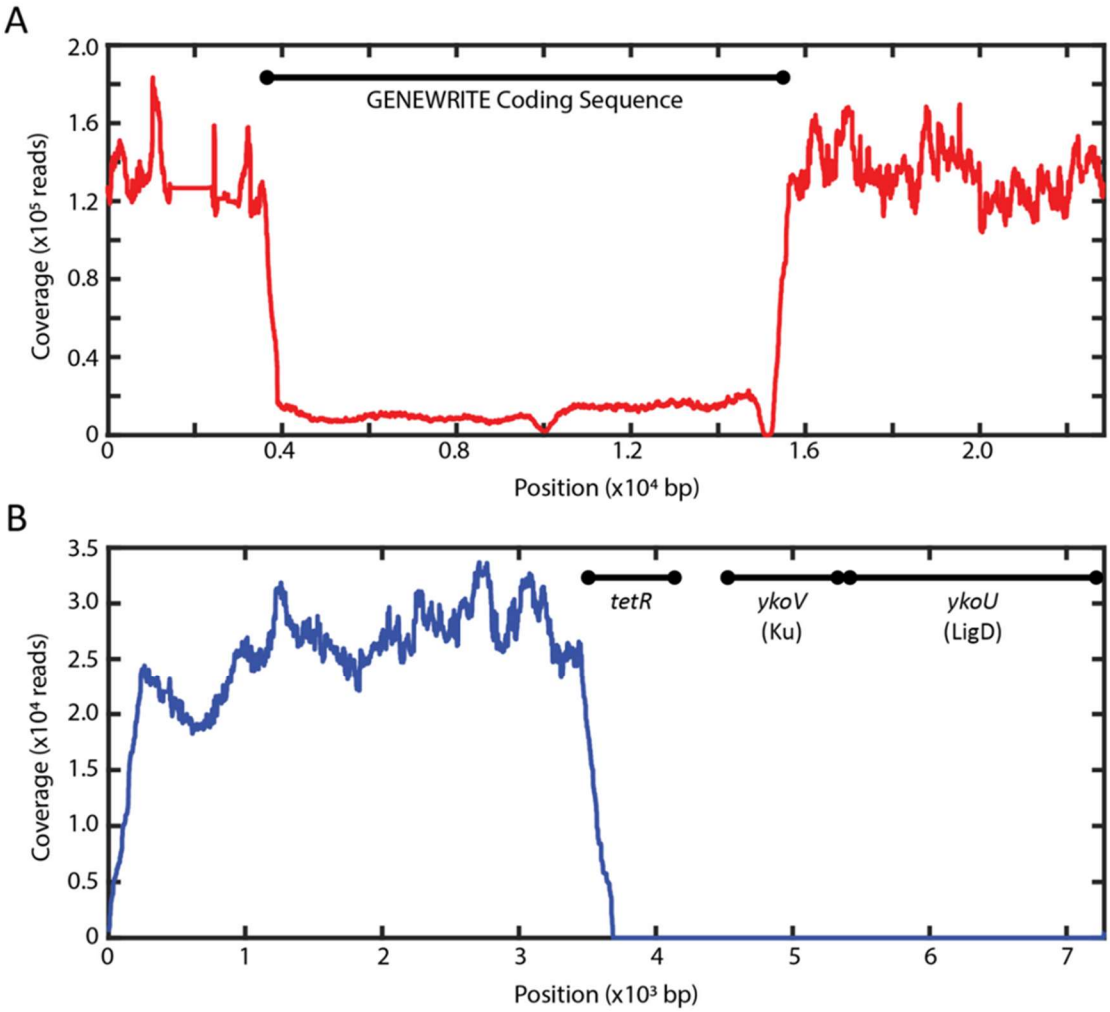

Figure 1.8. Sequencing coverage of (**A**) pUC57-kan-GENEWRITE and (**B**) pZA31-NHEJ, illustrating the excision of coding sequences of strongly expressed genes from these plasmids. Regions corresponding to each coding region are indicated by labeled lines.

Table 1.1. GENEWRITE Constructs. Variants of GENEWRITE constructs used in this study.

Plasmid	Antibiotic Resistance	Origin of Replication	Expressed Protein(s)	NLS	EN	Z	RT	NLS	6x His
pUC57-kan	Kanamycin (25 ug/ml)	pUC (~500 - 1000 copies/cell)	GENEWRITE (Cas9)	X	X (Cas9)	X	X	X	X
			GENEWRITE (Cas12a/Cpf1)	X	X (Cpf1)	X	X	X	X
			Cas9-ORF2pZRT		X (Cas9)	X	X		
			Cpf1-ORF2pZRT		X (Cpf1)	X	X		
			Cas9-ORF2pZRT + ORF1		X (Cas9)	X	X		
			Cas9Z		X (Cas9)	X			
pUC57-amp	Ampicillin (100 ug/ml)	pUC (~500 - 1000 copies/cell)	ORF2pZRT			X	X		
			ORF2pZRT -GAL4			X	X		
pZA31	Chloramphenicol (34 ug/ml)	p15A (~15 copies/cell)	NHEJ (ykoV + ykoU)						
			EMPTY						

Table 1.2. Oligos used in this study.

sgRNA Oligos	
scaffold	GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGG CTAGTCCGTTATCAACTTGAAAAAGTGGCACCGA GTCGGTGCT
scaffold R	AGCACCGACTCGGTGCCAC
T7 Cas9 pUC guide F	TAATACGACTCACTATAGGCGGATAACAATTTCAC ACGTTTTAGAGCTAGA
T7 Cpf1 pUC guide F	TAATACGACTCACTATAGCTGTGTGAAATTGTTAT CCGGTTTTAGAGCTAGA
T7 pZA31 guide F	TAATACGACTCACTATAGGGCGAAAATGAGACGT TGATGTTTTAGAGCTAGA
T7 atpI guide F	TAATACGACTCACTATAGAATATCAGTCTGCTAAA AATGTTTTAGAGCTAGA

Table 1.2 Continued. Oligos used in this study.

T7 atpI guide F	TAATACGACTCACTATAGAATATCAGTCTGCTAAAAATGTTTTAGAGCTAGA
T7 nth guide F	TAATACGACTCACTATAGTGTCAGTGTTAATAAGGCGAGTTTTAGAGCTAGA
T7 ybbD guide F	TAATACGACTCACTATAGCTGACTGAGAAAAGACATGTGTTTTAGAGCTAGA
Payload Oligos	
T7-PlacIQ1-RBS-aadA F	TAATACGACTCACTATAGATTTACGTTGACACCACCTTTCGCGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGAGGTAAATAATGCGCTCACGCAACTGGTCCAGAA
T7-pZA31-PlacIQ1-RBS-aadA F	TAATACGACTCACTATAGGGCGAAAATGAGACGTTGATATTTACGTTGACACCACCTTTCGCG
T7-atpI-PlacIQ1-RBS-aadA F	TAATACGACTCACTATAGAATATCAGTCTGCTAAAATTTACGTTGACACCACCTTTCGCG
T7-ybbD-PlacIQ1-RBS-aadA F	TAATACGACTCACTATAGCTGACTGAGAAAAGACATGTATTTACGTTGACACCACCTTTCGCG
T7-nth-PlacIQ1-RBS-aadA F	TAATACGACTCACTATAGTGTCAGTGTTAATAAGGCGAATTTACGTTGACACCACCTTTCGCG
pUC_lacZ0 R	TTATTTGCCGACTACCTTGGTGATC
pUC_lacZ10 R	GTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ20 R	GGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ30 R	CCTCGAGCATGGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ40 R	TTGGCTCGAGCCTCGAGCATGGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ50 R	CGCGCCGAGCTTGGCTCGAGCCTCGAGCATGGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC

Table 1.2 Continued. Oligos used in this study.

pUC_lacZ0pA R	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ10pA R	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ20pA R	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ30pA R	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCCTCGAGCATGGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ40pA R	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGCTCGAGCCTCGAGCATGGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
pUC_lacZ50pA R	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCGCGCCGAGCTTGGCTCGAGCCTCGAGCATGGTCATAGCTGTTTCCTGTGTTATTTGCCGACTACCTTGGTGATC
aadA pZA31 hybridization R	AGTGATCTTATTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCTTATTTGCCGACTACCTTGGTGATC
aadA atpI hybridization R	GACATTTTTAATAATGTTTTAACAGCCAATGATGGTTCTTAGCGCCGATTTTATTTGCCGACTACCTTGGTGATC
aadA nth hybridization R	TTCAAGCATCGCTGCAGGCGTATTCGCCACCGGGTAGAGTTTCGCCGTCGTTATTTGCCGACTACCTTGGTGATC
aadA ybbD hybridization R	CGAGTAGATATTCATCGTCTGAGCTATATGGCTTTACACAATAGCCGACATTATTTGCCGACTACCTTGGTGATC
aadA Cpf1 pUC hybridization R	CCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCTTATTTGCCGACTACCTTGGTGATC

Table 1.2 Continued. Oligos used in this study.

Verification and Sequencing Oligos	
aadA ver R	ACTGTACAAAAAAACAGTCATAAC
aadA ver F	CAGGCTTATCTTGGACAAGAAG
pUC ver F	AAACATCCCAATGGCGCGCCG
pUC ver R	GGCTTTACACTTTATGCTTC
pZA31 ver F	CGATAACTCAAAAAATACG
pZA31 ver R	GACGTCGATATCTGGCGAA
atpI ver F	CTTCGTCAGGTGCAACATGAGC
nth ver F	ACCACCGAGCTTAATTTCAGTTCGC
nth ver R	CCTGTTCGACGTTTTTCCCCGGCGC
ybbD ver F	ATTGGAGCTGGATTGCCTGATGCTTG
ybbD ver R	CTCACATTAAACACGTAACATTTTAATTAATG

Table 1.3. Summary of Results

Insertion Target	Target Copy Number (per cell)	3' Homology	5' Homology	ORF1p Coexpression	Efficiency
pUC57-*kan*	500 - 1000	x			72%
pZA31	20 - 30	x			20%
E. coli chromosome (*nth* locus)	1 - 2	x			0%
		x	x		0%
		x		x	0%
		x	x	x	60%

CHAPTER 2: REAL-TIME QUANTIFICATION OF THE EFFECTS OF IS*200*/IS*605* FAMILY-ASSOCIATED TNPB ON TRANSPOSON ACTIVITY

The work in this chapter is based on "Real-Time Quantification of the Effects of IS200/IS605 Family-Associated TnpB on Transposon Activity", which was published in The Journal of Visualized Experiments (JoVE) Volume 191 Article Number 64825 on January 28, 2023. The authors of this paper are Michael Worcester, Femila Manoj, and Thomas E. Kuhlman who are affiliated with the Department of Physics at the University of California, Riverside. Analysis and experiments are attributed to the authors who performed the work in their corresponding sections.

ABSTRACT

Here, a protocol is outlined to perform live, real-time imaging of transposable element activity in live bacterial cells using a suite of fluorescent reporters coupled to transposition. In particular, it demonstrates how real-time imaging can be used to assess the effects of the accessory protein TnpB on the activity of the transposable element IS*608*, a member of the IS*200*/IS*605* family of transposable elements. The IS*200*/IS*605* family of transposable elements are abundant mobile elements connected with one of the most innumerable genes found in nature, tnpB. Sequence homologies propose that the TnpB protein may be an evolutionary precursor to CRISPR/Cas9 systems. Additionally, TnpB has received renewed interest, having been shown to act as a Cas-like RNA-guided DNA endonuclease. The effects of TnpB on the transposition rates of IS*608* are quantified, and it is demonstrated that the expression of TnpB of IS*608* results in ~5x increased transposon activity compared to cells lacking TnpB expression.

INTRODUCTION

"That's almost as successful as my negative reading." - Noelle Reagen

Transposable elements (TEs) are genetic elements that mobilize within their host genomes by excision or catalyze copying followed by genomic reintegration. They were initially found by Dr. Barbara McClintock in the mid-1900s when she was observing corn kernels exhibiting multiple colors on a single ear of corn rather than growing a uniform ear of corn[1]. She found that DNA in the genes leading to kernel coloration were 'jumping' in and out leading to the variation of kernels. This led to the DNA initially being known as 'jumping genes' although we now know them as TEs.

TEs exist in all domains of life and the process of transposition restructures the host genome leading to mutations in coding and control regions[2]. This is done by a reverse transcriptase protein produced by the TE itself allowing for the reverse transcription of RNA back into DNA[3]. This generates mutations and diversity that play an important role in evolution, development, and several human diseases, including cancer[4–9].

Although it could be assumed that TEs would be able to transpose uncontrollably throughout the genome, they are actually found in low copy numbers in both prokaryotes and archaea[10]. In prokaryotes, it is known that TEs retrohome, meaning an intron-encoded homing endonuclease targets specific insert sites rather than replicating randomly throughout the genome[11].

Insertion Sequences (ISs) are the simplest form of TEs which only carry the genes required for transposition which allows them to be less than 2.5 kbp long[12]. In recent

years, it has been found that they aid in the distribution of antibiotic resistance genes in bacteria leading to creation of novel antibiotic resistance bacteria strains[13]. IS*200*/IS*605* is a family of TEs commonly found in both prokaryotes and archaea that can carry imperfect palindromic repeat (IP) sequences at the ends of the TEs[14]. This allows for the creation of unique structures which can then be recognized by the transposon [15]. One such example is the stem loop formed by IS*608*, a common experimental model from the IS*200/*IS*600* family.

Using novel genetic constructs that couple aspects of transpositional activity to fluorescent reporters, our previous work described the development of an experimental system based on the bacterial TE *IS608*, that allows for the real-time visualization of transposition in individual live cells(Figure 2.1)[16]. The TE system is displayed in Figure 1A. The TE comprises the transposase coding sequence, *tnpA*, flanked by Left End (LE) and Right End (RE) IPs, which are the excision sites for TnpA. *tnpA* is expressed using the promoter P_{LTetO1}, which is repressed by the *tet* repressor and is inducible with anhydrotetracycline (aTc)[17]. The TE splits the -10 and -35 sequences of a constitutive P_{lacIQ1} promoter for the blue reporter mCerulean3[18,19]. As shown in Figure 1C, when the production of *tnpA* is induced, the TE can be excised, leading to promoter reconstitution. The produced cell expresses mCerulean3 and fluoresces blue. The N-terminus of TnpA is fused to the yellow reporter Venus, allowing measurement of the TnpA levels by yellow fluorescence[20].

IS*608* and other members of the IS*200*/IS*605* family of transposons also typically encode a second gene of the thus far unknown function, *tnpB*[20,21]. The TnpB proteins are

a tremendously abundant but imperfectly characterized family of nucleases encoded by several bacterial and archaeal TEs, which often consist of only *tnpB*[22,23,24]. Furthermore, recent studies have renewed interest in TnpB due to the finding that TnpB functions as a CRISPR/Cas-like programmable RNA-guided endonuclease that will yield either dsDNA or ssDNA breaks under diverse conditions[25,26]. However, it remains unclear what role TnpB may play in regulating transposition. To perform real-time visualization of the effects of TnpB on *IS608* transposition, a version of the transposon, including the coding region of TnpB with an N-terminal fusion to the red fluorescent protein mCherry, was created.

Complementing more detailed bulk-level studies performed by the Kuhlman lab, it is shown here how real-time imaging of transposon activity can quantitatively reveal the impact of TnpB or any other accessory proteins on transpositional dynamics. By fusing TnpB to mCherry, the individual transpositional events are identified by blue fluorescence and correlated with expression levels of TnpA (yellow fluorescence) and TnpB (red fluorescence).

MATERIALS AND METHODS

Contributing Authors: FLM

Preparation of bacterial cultures

Grow *E. coli* strain MG1655 with plasmid transposon constructs (previously described in Kim et al.) overnight in LB with the appropriate antibiotics (25 μg/mL of

kanamycin) at 37 °C[8]. The sequences of the constructs used and the related sequences are available as GenBank[20] accession numbers OP581959, OP581957, OP581958, OP717084, and OP717085[19]. To achieve steady-state exponential growth, dilute cultures ≥100 fold into the M63 medium (100 mM KH_2PO_4, 1 mM $MgSO_4$, 1.8 µM $FeSO_4$, 15 mM $[NH_4]_2SO_4$, 0.5 µg/mL thiamine [vitamin B1]) supplemented with a carbon source (0.5% w/v glucose here) and appropriate antibiotics. Grow cultures at 37 °C until the optical density at 600 nm (OD600) reaches ~0.2. The cultures are ready for use.

Slide preparation

Contributing Authors: FLM, TEK

Prepare a slide by boiling M63 with 0.5% w/v glucose and 1.5% w/v agarose in the microwave to melt the agarose and ensure that it is completely molten and well mixed. Allow the mixture to cool to ~55 °C before adding antibiotics and inducers (25 µg/mL Kanamycin and 10 ng/µL anhydrotetracycline [aTc]). Place a microscope slide on the workbench. Stack two more slides perpendicular to the first and place another on top, parallel to the bottom slide. Ensure that there is a gap equal to one slide thickness between the bottom and top slides. Pipette ~1 mL of the M63 agarose mixture into this gap between the slides slowly to create a small gel square. Once the gel has solidified (~10-15 min), slide the top slide to remove it. Trim the agarose pad with a razor blade or knife. Then pipette 2.5 µL of the culture and put the coverslip on top. Seal the space between the slide and the coverslip with epoxy. Allow the epoxy to dry and the cells to settle onto the agarose pad for at least 1 h at 37 °C.

Timelapse fluorescence microscopy

Contributing Authors: MPW

Place the prepared sample on a fluorescence microscope in an environment heated and maintained at 37 °C. Set the exposure times appropriate for the camera used for image acquisition. Adjust the illumination intensity to minimize photobleaching. An exposure time of 2 s for each wavelength was used for the present study. For each wavelength, find a Field of View (FOV) containing minimal fluorescence. Acquire images to use during the analysis for background subtraction. Set up a protocol to acquire images in a grid at different wavelengths and at regular time intervals. Encode timelapse photography into the protocol. Set the acquisition frequency to the desired time interval (20 min here) and the total timelapse duration to the desired length (24 h). Encode appropriate wavelengths into the protocol (depending upon the construct used). The mCherry excitation peak is at 587 nm and the emission peak at 610 nm, mVenus is at 515 nm and 527 nm, while mCerulean3 is at 433 nm and 475 nm[11,12,20]. Set the grid size to capture between the desired number of FOVs. The representative data shown here used 8 x 8 FOVs.

Image analysis

Contributing Authors: MPW

Perform background subtraction on each color channel by using the respective background images acquired in step 3.1.2. For all the analysis steps, we use standard

modules in the open-source platform Fiji[21]. Approximate the total population at each point in time by thresholding the mCerulean channel and dividing the threshold area by the average cell area. To count the unique excision events, take the time derivative of the mCerulean3 channel. Perform this by subtracting successive images in the mCerulean3 channel. The excision events will be detected in the time derivative as a bright flash of fluorescence. Threshold the stack of excision events to eliminate unwanted fluorescence. Note that this process will threshold out parts of the excisions themselves. To fix this, dilate the images to restore the excisions to their original sizes. Analyses using similar thresholding and image analysis techniques can be performed on the other fluorescence channels too (e.g., correlate excision events with levels of transposase TnpA [yellow Venus fluorescence] and TnpB [red mCherry fluorescence]).

RESULTS

This method of visualizing transposon activity in live cells by fluorescence microscopy, while having lower throughput than bulk fluorescence measurements, allows direct visualization of transposon activity in individual live cells. Transposon excision events result in the reconstitution of the promoter for mCerulean3 (Fig. 2.1), allowing identification of cells undergoing transposon activity by bright blue fluorescence (Fig. 2.2, TnpB+).

It is found that cells expressing the accessory protein TnpB (Fig. 2.3, orange) experience 4-5 times higher levels of transposon activity compared to those that do not (Fig. 2.3, blue), consistent with the more detailed bulk-level studies. This is particularly

notable as the inclusion of the coding sequence of *mCherry-tnpB* increases the length of the transposon by ~2,000 bp, while previous studies have found that *IS608* transposon excision is an exponentially decreasing function of transposon length[22].

An advantage of real-time imaging is that once identified, cells undergoing transpositional events can be further tracked and analyzed to determine other characteristic parameters, such as growth rate, to determine the distribution of fitness effects or the expression level of accessory proteins to determine their impact on transpositional activity. For example, in TnpB+ cells, cells undergoing transposon excision events have higher expression levels of mCherry-TnpB than the general population (Fig 2.4A). Moreover, for cells undergoing excision events (Fig 2.4B, dark yellow), TnpB+ cells (Fig. 2.4B, bottom) express only marginally higher levels of Venus-TnpA transposase than TnpB- cells (Fig. 2.4B, top) (TnpB- 158.3 ± 68.2 AU, TnpB+: 193 ± 79.9 AU), which is higher than the yellow fluorescence of the general population (Fig 2.4B, light yellow). Taken together, these data suggest that TnpB protein is responsible for the observed higher levels of transpositional activity.

DISCUSSION

The unique method presented here for real-time imaging of transposable element activity in live cells is a sensitive assay that can directly detect transposition in live cells and in real-time and correlate this activity with the expression of accessory proteins. While the throughput is lower than can be accomplished by bulk methods, this method

achieves detailed measurements of TE activity and protein expression in individual living cells.

A variety of tools and techniques can be employed to grow cells directly on the microscope for real-time imaging. The method used here of cell growth on agarose pads has the advantage of being fast, cheap, and easy to perform. A possible disadvantage, depending on the cellular growth state of interest, is that resources available to support cell growth in the agarose pad are limited, and hence cells will naturally exhaust these resources and stop growing after a relatively short period of time (12-24 h). Consequently, care must be taken to prepare the cells in steady state growth and inoculate the pad at a low enough density to give ample time for measurement. Microfluidics can be employed to maintain cells in steady state exponential growth for extended periods of time, although these methods require additional expertise, equipment, and setup to be effective[23].

Complementing more detailed work from the Kuhlman lab, it is illustrated here that the IS*200*/IS*605* TE family-associated protein TnpB increases the rate of IS*608* excision by up to five-fold, and that increased excision is directly correlated with higher expression levels of TnpB. These methods are one example of improved assay techniques that may help shed light on transposon activity and its impact on mutational and evolutionary dynamics.

REFERENCES

1. McCLINTOCK, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U. S. A.* **36**, 344–355 (1950).

2. Cowley, M. & Oakey, R. J. Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet.* **9**, e1003234 (2013).

3. Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209–1211 (1970).

4. Schneider, D. & Lenski, R. E. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res. Microbiol.* **155**, 319–327 (2004).

5. Chao, L., Vargas, C., Spear, B. B. & Cox, E. C. Transposable elements as mutator genes in evolution. *Nature* **303**, 633–635 (1983).

6. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).

7. Kano, H. *et al.* L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev.* **23**, 1303–1312 (2009).

8. Belancio, V. P., Deininger, P. L. & Roy-Engel, A. M. LINE dancing in the human genome: transposable elements and disease. *Genome Med.* **1**, 97 (2009).

9. Goodier, J. L. Retrotransposition in tumors and brains. *Mob. DNA* **5**, 11 (2014).

10. Mohr, G., Ghanem, E. & Lambowitz, A. M. Mechanisms used for genomic proliferation by thermophilic group II introns. *PLoS Biol.* **8**, e1000391 (2010).

11. Cousineau, B. *et al.* Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* **94**, 451–462 (1998).

12. Mahillon, J. & Chandler, M. Insertion sequences. *Microbiol. Mol. Biol. Rev.* **62**, 725–774 (1998).

13. Debets-Ossenkopp, Y. J. *et al.* Insertion of mini-IS605 and deletion of adjacent sequences in the nitroreductase (rdxA) gene cause metronidazole resistance in Helicobacter pylori NCTC11637. *Antimicrob. Agents Chemother.* **43**, 2657–2662 (1999).

14. Ronning, D. R. *et al.* Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Mol. Cell* **20**, 143–154 (2005).

15. Barabas, O. *et al.* Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. *Cell* **132**, 208–220 (2008).

16. Kim, N. H. *et al.* Real-time transposable element activity in individual live cells. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 7278–7283 (2016).

17. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).

18. Calos, M. P. & Miller, J. H. The DNA sequence change resulting from the IQ1 mutation, which greatly increases promoter strength. *Mol. Gen. Genet.* **183**, 559–560 (1981).

19. Markwardt, M. L. *et al.* An improved cerulean fluorescent protein with enhanced brightness and reduced reversible photoswitching. *PLoS One* **6**, e17896 (2011).

20. Nagai, T. *et al.* A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat. Biotechnol.* **20**, 87–90 (2002).

21. Kersulyte, D. *et al.* Transposable element ISHp608 of Helicobacter pylori: nonrandom geographic distribution, functional organization, and insertion specificity. *J. Bacteriol.* **184**, 992–1002 (2002).

22. Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR-Cas systems. *Nat. Rev. Microbiol.* **15**, 169–182 (2017).

23. Bao, W. & Jurka, J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mob. DNA* **4**, 12 (2013).

24. Li, X. *et al.* Genetic Characterization and Passage Instability of a Hybrid Plasmid Co-Harboring and Reveal the Contribution of Insertion Sequences During Plasmid Formation and Evolution. *Microbiol Spectr* **9**, e0157721 (2021).

25. Altae-Tran, H. *et al.* The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57–65 (2021).

26. Karvelis, T. *et al.* Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature* **599**, 692–696 (2021).

27. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).

28. Shaner, N. C. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nat. Biotechnol.* **22**, 1567–1572 (2004).

29. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

30. Ton-Hoang, B. *et al.* Single-stranded DNA transposition is coupled to host replication. *Cell* **142**, 398–408 (2010).

31. Wang, P. *et al.* Robust growth of Escherichia coli. *Curr. Biol.* **20**, 1099–1103 (2010).

FIGURES AND TABLES

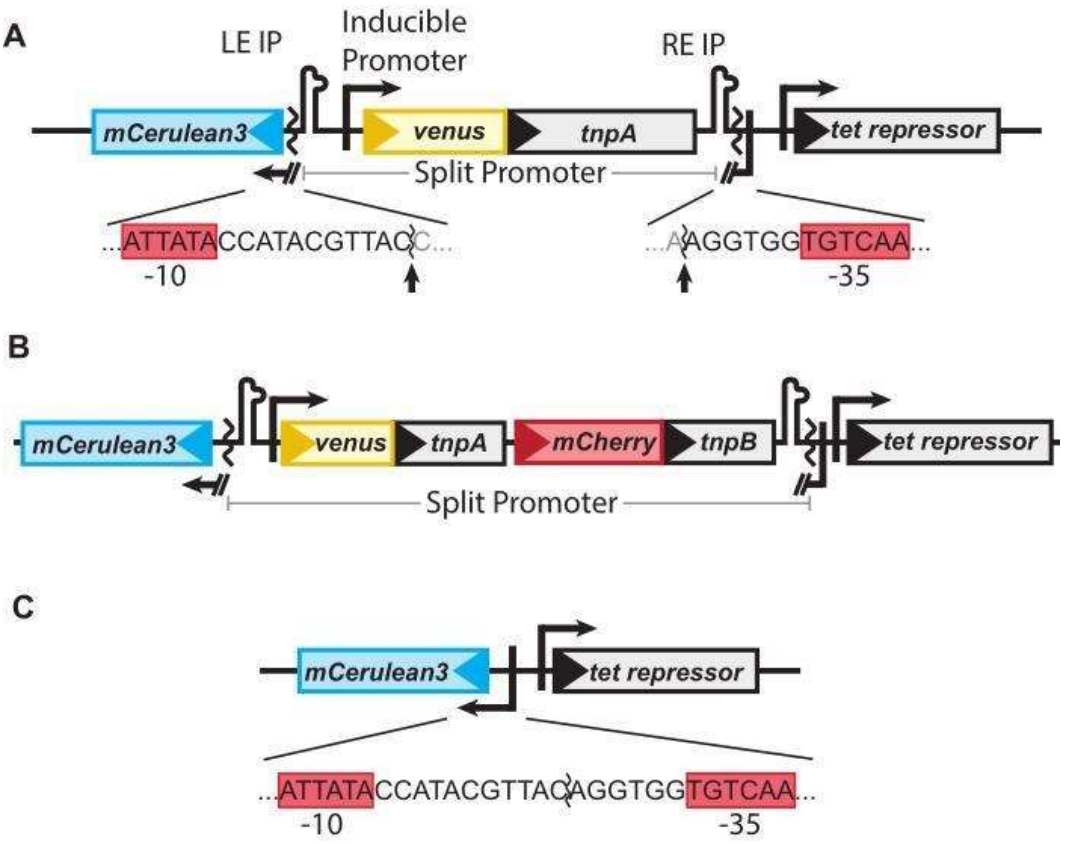

Figure 2.1. Genetic constructs for imaging of real-time transposon dynamics. (**A**) The mCerulean3 promoter is disrupted by the TE, the ends of which are flanked by the left end and right end faulty palindromic sequences (LE IP and RE IP). The transposase, *tnpA* (gray), is expressed from P_{LtetO1}, which is regulated by the *tet* repressor (gray) and is inducible with anhydrotetracycline (aTc). The sequences of the Promoter/TE junction and promoter -10 and -35 sequences (red boxes), and TnpA cleavage sites are shown by arrows. (**B**) The TnpB+ construct is where mCherry-*tnpB* has been transcriptionally fused to venus-*tnpA* such that both are transcribed as a polycistronic mRNA, mimicking the natural configuration of *IS608*. (**C**) Upon excision, the mCerulean3 promoter is repaired, and the cell exhibits blue fluorescence. The reconstituted promoter sequence is displayed below the diagram.

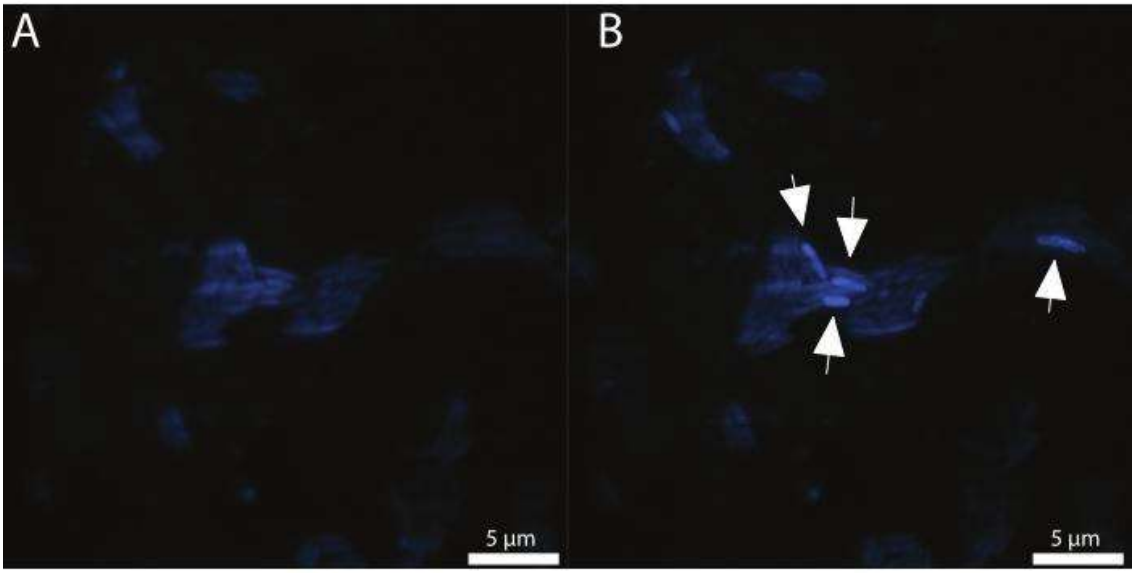

Figure 2.2. Visualization of transposon excision events. The example field of view of TnpB+ with cells (**A**) immediately before and (**B**) after detecting transposon excision events by blue fluorescence. White arrows indicate excision events. The time difference between the two frames is 20 min. Scale bar = 5 µm.

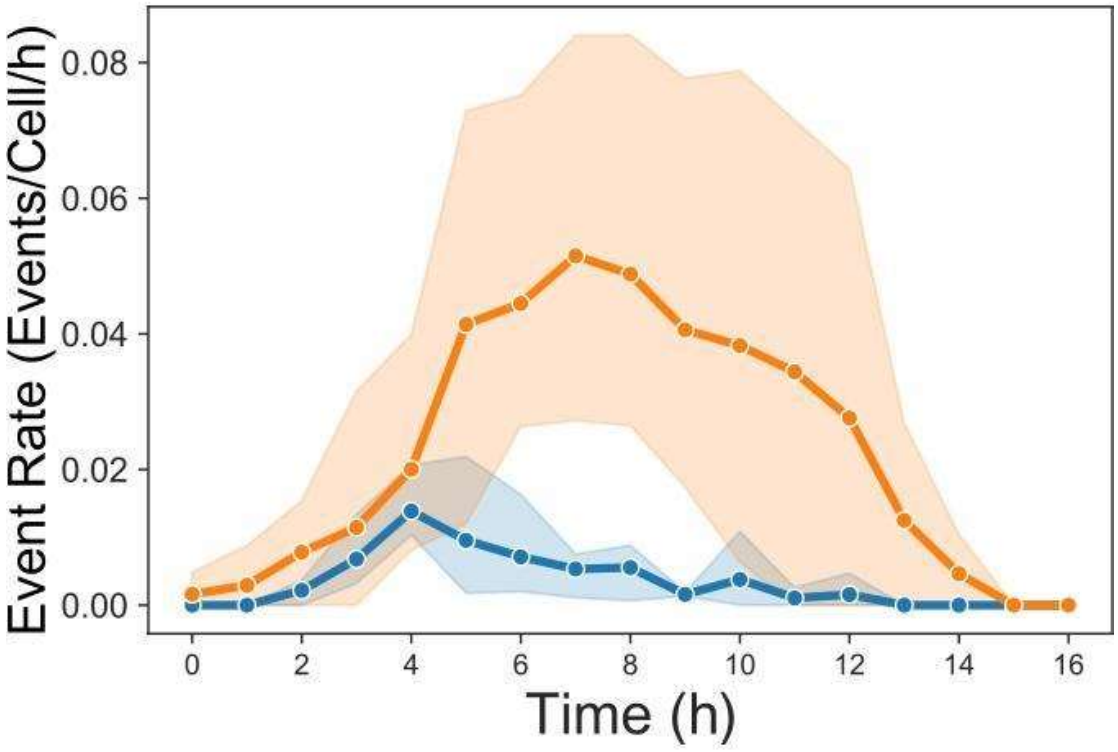

Figure 2.3. TnpB enhances transposon excision rate. The excision rate for TnpB+ cells (orange) and TnpB- (blue) cells. The mean rate from three replicates is shown as points with shaded regions with a 95% confidence interval. The data are aligned so that cells begin excising at t = 0. The maximum measured rate for TnpB+ cells was $5.1 \pm 2.4 \times 10^{-2}$ events per cell per hour, while for TnpB- was $1.4 \pm 0.48 \times 10^{-2}$ events per cell per hour. The average rate over the whole interval shown was $2.6 \pm 1.8 \times 10^{-2}$ events per cell per hour for TnpB+ cells and $5.3 \pm 2.9 \times 10^{-3}$ events per cell per hour for TnpB- cells.

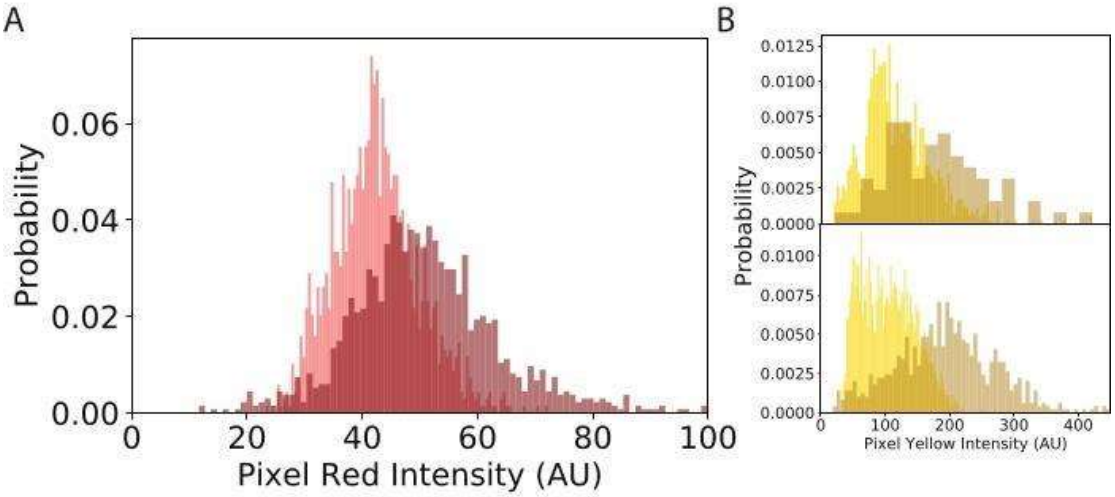

Figure 2.4. Protein expression statistics for excising cells versus total cell population. Each frame was divided into 64 equal blocks, and fluorescence was measured for cells excising within the block and for all cells contained within the block regardless of the excision activity. Probability, as plotted on the y-axis, is measured as the number of pixels of the indicated intensity in each cell type divided by the total number of pixels. The block size for each frame was set to 445 x 445 pixels. (**A**) The cells that undergo excision events (dark red) express more TnpB than the general population (light red). The average red fluorescence for excising cells was 51.3 ± 15.4 AU (dark red), while that for all cells was 42.5 ± 7.4 AU (light red). (**B**) Venus-TnpA transposase levels are similar in TnpB- (top) and TnpB+ (bottom) cells. The data sets are normalized so that the mean yellow fluorescences of the total cell population (light yellow) are equal for TnpB+/- at 105.7 AU. The cells that have been identified as undergoing transposon excision events (dark yellow) exhibit higher yellow fluorescence than the general population, with similar distributions for TnpB- (top) and TnpB+ (bottom). Mean yellow fluorescences of the excising populations are TnpB-: 158.3 ± 68.2 AU and TnpB+: 193 ± 79.9 AU.

CHAPTER 3: PHYLOGENETIC DISTRIBUTION OF PROKARYOTIC NON-HOMOLOGOUS END JOINING DNA REPAIR SYSTEMS IN BACTERIA AND ARCHAEA

The work in this chapter is based on "Phylogenetic Distribution of Prokaryotic Non-homologous End Joining DNA Repair Systems in Bacteria and Archaea", which was posted on bioRxiv on September 30, 2023. The authors of this paper are Femila Manoj and Thomas E. Kuhlman who are affiliated with the Department of Physics at the University of California, Riverside. Analysis and experiments are attributed to the authors who performed the work in their corresponding sections.

ABSTRACT

Non-homologous end-joining (NHEJ) is a repair mechanism for double strand breaks (DSBs) of DNA. This mechanism is ubiquitously observed within the eukaryotic domain; however, its presence is not as pervasive among prokaryotes and archaea. Notably, in prokaryotes, it has been discerned that multiple distinct NHEJ pathways have evolved in contrast to the singular NHEJ pathway prevalent in eukaryotes. We performed phylogenetic analysis to gain deeper insights into the distribution of these prokaryotic NHEJ pathways. Concurrently, components of the prokaryotic NHEJ pathways were used to find if any archaea carry the genes required and may be able to carry out NHEJ. The results show that few prokaryotes carry the components required for NHEJ, but multiple pathways may be active in a single species. In the context of Archaea, the analysis revealed that a substantial number of species contain fragments or segments of prokaryotic NHEJ elements. Nevertheless, the presence of all the necessary components for the complete execution of the NHEJ pathway remains relatively rare within the archaeal domain.

INTRODUCTION

"Delulu is the solulu." - Leticia Perez

Non-homologous end joining (NHEJ) is a crucial repair pathway that safeguards genome integrity by resolving double-strand breaks (DSBs) when a suitable DNA template for homologous recombination (HR) is lacking[1]. While the basic tenets of NHEJ have been explained, further investigations have uncovered numerous NHEJ-like genes in bacterial genomes, suggesting the existence of other DNA repair components[2]. Additionally, prokaryotic NHEJ is relatively scarce in Archaea, primarily due to the absence of key NHEJ proteins, such as Ku, in most Archaea[3]. While all the genes for proteins required for prokaryotic NHEJ can be found in very few Archaea, the genes are not next together in the genome and do not form a single NHEJ operon. Instead, the individual parts of NHEJ are found spread out in different places in the genome[4].

Prokaryotic NHEJ was first discovered in *Bacillus subtilis* 21 years ago[5]. This involved targeted deletion of the genes responsible for LigD and Ku proteins, followed by assessing the survival rate after exposure to ionizing radiation (IR), which induces DSBs. Prior to this discovery, it was widely believed that prokaryotes exclusively relied on HR for DNA repair. Subsequently, it was demonstrated that *Mycobacterium tuberculosis* and *Mycobacterium smegmatis* possess an operational NHEJ pathway capable of repairing plasmid DNA following transformation[6]. This further solidified the notion that numerous prokaryotic species possess the ability to employ NHEJ for DNA repair processes. However, many prokaryotes, such as *Escherichia coli*, have been found to lack any form of NHEJ[7].

This study focuses on elucidating the three distinct types of prokaryotic NHEJ pathways as outlined by Bertrand *et al*. (Fig 3.1)[8]. These NHEJ classifications are based on the characterized pathways observed in *B. subtilis*, *Streptomyces ambofaciens*, and *Sinorhizobium meliloti*. In the case of *B. subtilis* it was determined that deletion of either the LigD or Ku protein hindered the functionality of NHEJ, suggesting both are necessary for the pathway[5]. To denote this form of NHEJ that relies on only two proteins, it will henceforth be referred to as "minimal NHEJ". Further investigations conducted in *S. ambofaciens* revealed that mutations affecting other NHEJ-associated proteins, such as LigC, KuA, PolR, and PolK, had a significant impact on the overall efficiency of the NHEJ pathway[9]. Here, the NHEJ pathway encompassing these four proteins will be designated as "core NHEJ". More recent findings in *S. meliloti* unveiled the existence of two additional distinct functional NHEJ pathways; each activated under different stress conditions[10]. The first pathway, operational in stress-free bacteria, necessitates the presence of LigD2 and Ku2 proteins and will be referred to as the "main NHEJ" pathway. Conversely, the second pathway, induced exclusively under stress conditions, requires the participation of LigD4, Ku3, and Ku4 proteins and will be designated as the "secondary NHEJ" pathway. The combined occurrence of both main and secondary NHEJ pathways in *S. meliloti* will be referred to as "multiple NHEJ".

Here, we perform a bioinformatic analysis on available bacterial and archaeal genome sequences to characterize the phylogenetic distribution of the various NHEJ pathways. We consider prokaryotic species with all proteins required for an NHEJ type to have a complete NHEJ pathway. To identify archaea species that may have prokaryotic

NHEJ systems, we identified species with the four genes required for NHEJ in *Methanocella paludicola*[4].

MATERIALS AND METHODS

Contributing Authors: FLM

Data

Computations were performed using the computer clusters and data storage resources of the HPCC, which were funded by grants from NSF (MRI-2215705, MRI-1429826) and NIH (1S10OD016290-01A1).

Identification of Prokaryotes with a NHEJ Pathway

The UniProtKB database was used to identify the protein sequences of proteins necessary for functional prokaryotic NHEJ (LigD and Ku for minimal and core NHEJ, LigC, PolR, PolK, and KuA, for core NHEJ and LigD2, Ku2, LigD4, Ku3, and Ku4 for multiple NHEJ) using BLAST+ with the *E*-value cutoff of 0.0001 on the HPCC[8,11,12]. Query sequences were pulled from the NHEJ pathways of *Bacillus subtilis*, *Mycobacterium tuberculosis*, *Streptomyces ambofaciens*, and *Sinorhizobium meliloti* respectively. Results were downloaded and RegEx from Python was used to split the organism names from the results. Python was then used to remove duplicate organism names from the results and find duplicates in multiple results to sort organisms based on which NHEJ pathway they had.

Identification of Archaea with a NHEJ Pathway

The UniProtKB database was again used to BLAST the sequences of proteins necessary for functional NHEJ using prokaryotic elements in *Methanocella paludicola* (Ku, Pol, PE, and Lig) using BLAST+ with the *E*-value cutoff of 0.0001 on the HPCC[11–13]. Query sequences were pulled from *Methanocella paludicola.* Results were downloaded and RegEx from Python was used to split the organism names from the results. Python was then used to remove duplicate organism names from the results and find duplicates in multiple results to sort organisms based on if they had all the required protein sequences.

Prokaryotic Phylogenetic Tree Construction

The 16S ribosomal ribonucleic acid (rRNA) sequences of identified prokaryotes were downloaded from the National Center for Biotechnology Information (NCBI)[12]. Single 16S rRNA sequences were chosen from prokaryotes with multiple sequences by selecting the sequence with the fewest Ns and longest sequence length. Multiple sequence alignments (msa) were generated using muscle v5 on default settings[14]. IQTREE v2.2.0 was used to generate a maximum-likelihood (ML) phylogenetic tree[15]. The best-fit model was found to be TIM3+F+R10 (LogL = -278,567.605, BIC = 582,595.148) using Model Finder[16]. UFBoot2 (-bnni option) was used to assess branch supports[17]. Finally iTOL was used to visualize the trees and create colored ranges to signify NHEJ pathways[18].

Archaic Phylogenetic Tree Construction

The 16S ribosomal ribonucleic acid (rRNA) sequences of identified archaea were downloaded from the National Center for Biotechnology Information (NCBI)[12]. Single 16S rRNA sequences were chosen from archaea with multiple sequences by selecting the sequence with the fewest Ns and longest sequence length. Multiple sequence alignments (msa) were generated using muscle v5 on default settings[14]. IQTREE v2.2.0 was used to generate a maximum-likelihood (ML) phylogenetic tree[15]. The best-fit model was found to be GTR+F+R4 (LogL = -18,316, BIC = 38,109.029) using Model Finder[16]. UFBoot2 (-bnni option) was used to assess branch supports[17]. Finally iTOL was used to visualize the trees[18].

RESULTS

Few Prokaryotes have Proteins Required for NHEJ

The protein sequences essential for NHEJ were subjected to Basic Local Alignment Search Tool (BLAST) analysis against the UniProtKB database[11]. The obtained results were utilized to extract species names associated with individual proteins as well as those encompassing the entire complement of proteins necessary for each NHEJ type. Subsequently, the 16S ribosomal ribonucleic acid (rRNA) sequences of species possessing the complete set of NHEJ proteins were retrieved from the National Center for Biotechnology Information (NCBI)[12]. These sequences were employed to generate an alignment and construct a phylogenetic tree.

Initially, this study focused on prokaryotes due to the existence of known (NHEJ) pathways as outlined by Bertrand et al. in their review published in 2019 (Fig 3.1)[8]. These pathways include minimal NHEJ, core NHEJ, and multiple NHEJ. Minimal NHEJ requires only two proteins: Ku and LigD. Core NHEJ requires more proteins: LigC, PolR, PolK, and KuA. Multiple NHEJ is further broken down into two subpathways: main NHEJ and secondary NHEJ. Main NHEJ is activating during times of stability and requires LigD2 and Ku2. Secondary NHEJ is activating during times of stress and requires LigD4, Ku3, and Ku4. Utilizing the information of these established pathways, an investigation was conducted to ascertain the number of prokaryotic species carrying all the essential genes necessary for at least one of the known functional NHEJ pathways.

It was found that prevalence of NHEJ in prokaryotes is relatively limited, with only 2624 out of 12,948 species available in the UniProtKB database exhibiting a complete form of NHEJ (Fig 3.2A)[12].

Minimal NHEJ is the Most Commonly Found NHEJ Pathway in Prokaryotes

Following the determination of prokaryotic species with known NHEJ pathways, further investigation was directed towards each established NHEJ pathway. Among the previously mentioned 2624 species possessing a complete set of genes required for NHEJ, an overwhelming majority of 2384 were found to be associated with the minimal NHEJ pathway, thus consolidating its prominence as the predominant form of NHEJ within the prokaryotic domain, as depicted in Figure 3.2C. A total of 20 prokaryotic species carrying the minimal NHEJ pathway were identified to also possess all the essential components necessary for multiple NHEJ.

The major phylums of prokaryotes carrying minimal NHEJ were Actinomycetota, Bacilliota, Bacteroidota, and Pseudomonadata with the Actinomycetota phylum having 400 species with NHEJ (Fig 3.3 - 3.6). Within the Actinomycetota phylum, major orders with minimal NHEJ were Micrococcales, Actinomycetales, Geodermatophilales, Pseudonocardinales, Mycobacteriales, Corynebacteriales, Propionibacteriales, and Streptosporangiales (Fig 3.3). Within the order of Micrococcales, major genera carrying minimal NHEJ were *Brevibacterium*, *Promicromonospora*, *Terrabacter*, *Pedococcus*, *Curtobacterium*, *Agreia*, *Agromyces*, *Mycetocola*, and *Cryobacterium* (Fig 3.7). Within the order of Actinomycetales, the *Microbacterium* genus carried the most species with minimal NHEJ followed by *Gordonia*, *Cellulomonas*, *Tessaracoccus*, *Mycobacteroides*, *Frankia*, *Flavimobilis*, and *Motilibacter* (Fig 3.8). The Geodermatophilales order only has three genera with minimal NHEJ: *Blastococcus*, *Modestobacter*, and *Geodermatophilus* (Fig 3.9). Within the order of Pseudonocardinales, *Amycolatopsis* was the most prominent genus followed by *Pseudonocardia*, *Saccharomonospora*, *Prauserella*, and *Actinoalloteichus* (Fig 3.10). Only four genera in the Mycobacteriales order carry the genes for minimal NHEJ: *Nocardia*, *Williamsia*, *Mycolicibacterium*, and *Mycobacterium* (Fig 3.11). Within the order of Corynebacteriales, *Rhodococcus* is the only genus to carry the genes for minimal NHEJ (Fig 3.12A). In the order of Propionibacteriales, the only major genus to carry minimal NHEJ genes is *Microlunatus* and in the order of Streptosporangiales, *Nonomuraea* is the major genus carrying NHEJ (Fig3.12).

The other phylums carry fewer species with minimal NHEJ. Major genera carrying NHEJ within the Bacilliota phylum are *Brevibaccillus*, *Alkalihalobacillus*, and *Bacillus* although many other genera also carry a few species with minimal NHEJ (Fig 3.4). Within the Bacteroidota phylum, the genus with the most species carrying minimal NHEJ is *Chryseobacterium*, but other major genera within the phylum include *Chitinophaga*, *Dyadobacter*, and *Ephilithonimonas* (Fig. 3.6). Many genera within the phylum Pseudomonadota carry minimal NHEJ, but major genera are *Brevundimonas*, *Novoshingobacterium*, *Paracoccus*, *Devosia*, *Bordetella*, *Achromobacter*, *Cupriavidus*, *Caballeronia*, *Paraburkholderia*, and *Burdholderia* (Fig 3.5). Genera from phylums with fewer species are shown in Figure 3.13.

Core and Multiple NHEJ are Less Popular in Pathways

Core NHEJ and multiple NHEJ genes are observed in only 2% of prokaryotic species with multiple NHEJ being the least common with it found in only 0.4% of prokaryotes (Fig 3.2B). Notably, it was observed that all prokaryotic organisms with 16S rRNA available on NCBI hosting the core NHEJ pathway genes also possess the multiple NHEJ pathway genes, indicating a consistent correlation between these two NHEJ pathways (Fig 3.14 and 3.15). Phylogenetic analysis showed the two pathways sporadically inherited, but notably, the genus *Streptomyces* was found to consistently carry core NHEJ (Fig 3.14). Other major genera carrying multiple NHEJ include *Rhizobium*, *Bradyrhizobium*, *Saccharopolyspora*, and *Mycobacterium* (Fig 3.15).

Archaea Mostly Carry Parts of NHEJ

Since there is a prevalence of prokaryotic NHEJ components among Archaic species, a comparative analysis akin to the prokaryotic investigation was undertaken with archaea[4]. The components of NHEJ found in *Methanocella paludicola* were considered required for a NHEJ pathway akin to that of Prokaryotes to be carried out. The NHEJ proteins used in *M. paludicola* are Ku, Pol, PE, and Lig. The protein sequences of these proteins were used for a BLAST analysis against the UniProtKB database[11]. Species names of carriers of all the proteins required for prokaryotic NHEJ were compiled. The 16S rRNA sequences of these were then retrieved from the NCBI database and utilized to create an alignment and construct a phylogenetic tree to showcase the evolutionary relationship of archaea carrying prokaryotic NHEJ[12].

Among the 567 Archaea species documented in the NCBI database, only 16 do not have any prokaryotic NHEJ genes (Fig 3.16). However, of the remaining 545 species, only a total of 97 species from six different classes possess all the requisite parts for NHEJ indicating that while many Archaea carry parts of NHEJ genes, few could carry out the NHEJ pathway. The major genera carrying prokaryotic NHEJ in Archaea are *Methobacterium*, *Archaeoglobus*, *Methanococcoides*, *Nitrosopumilus*, and *Thermococcus* (Fig 3.17).

DISCUSSION

To identify prokaryotic species with a higher likelihood of possessing a functional NHEJ pathway, a collection of prokaryotic species encompassing all the requisite

proteins for NHEJ pathways was compiled. The NHEJ pathways that were used were minimal NHEJ, core NHEJ, and multiple NHEJ, as described in Bertrand et al. 2019[8]. For minimal NHEJ, the essential proteins were LigD and Ku, whereas LigC, KuA, PolR, and PolK were required for core NHEJ. Multiple NHEJ comprises two subpathways, namely main NHEJ and secondary NHEJ, necessitating LigD2 and Ku2 for the former and LigD4, Ku3, and Ku4 for the latter. Species possessing all the proteins required for a particular pathway were classified as having "complete NHEJ" and thus are potentially capable of executing the pathway. For archaea, species with PE, Lig, Pol, and Ku were considered ‘complete’ as those are the only NHEJ proteins that have been found in archaea thus far[4].

Unlike Archaea, prokaryotic genomes tend to possess NHEJ protein coding sequences in closer proximity, making it less common to encounter isolated components of the NHEJ machinery within prokaryotes. This study revealed that while 80% of Prokaryotes do not have all the proteins required for functioning NHEJ, minimal NHEJ is the most common form found (Fig 3.2A). Notably, minimal NHEJ only relies on two distinct proteins, whereas the other prokaryotic NHEJ forms involve a greater number of proteins, suggesting that the simplicity of minimal NHEJ may facilitate its widespread occurrence. Minimal NHEJ was predominantly found in the Actinomycetota phylum with over 400 species, but they were spread over multiple orders and genera showcasing the capabilities of horizontal gene transfer (HGT) (Fig 3.3). Some genera within Actinomycetota where vertical gene transfer is more likely are *Cellulomonas*, *Microbacterium*, *Gordonia*, *Rhodococcus*, *Geodermatophilus*, *Mycolicibacterium*,

Mycobacterium, *Nocardia*, *Amycolatopsis*, *Pseudonocardia*. Notably, a popular genus of Actinomycetota that did not carry any minimal NHEJ was *Streptomyces*. Within the Bacteroidota phylum, *Cryseobacterium* is a genus that may have also evolved with minimal NHEJ rather than receiving it from HGT (Fig 3.6).

The least commonly observed NHEJ form in prokaryotes is multiple NHEJ, which comprises main and secondary NHEJ subpathways. Intriguingly, species exhibiting multiple NHEJ could also possess core NHEJ (Fig 3.14 and 3.15). Core NHEJ was also exclusively identified in the *Streptomyces* genus, implying a potential vertical transfer of core NHEJ in that genus (Fig 3.14). Furthermore, the occurrence of both multiple and core NHEJ in similar species across different clades suggests a combination of horizontal and vertical gene transfers of these NHEJ forms. Reliance of core NHEJ on multiple NHEJ could also be implied by this. It is possible that core NHEJ is also an accessory part of multiple NHEJ or that the two share similar mechanisms and thus genetics as well. Since core and minimal NHEJ were found to never occur together, it is possible that one inhibits the other leading to the organism choosing one pathway over the other.

Another prominent genus that carries multiple NHEJ is *Mycobacterium* (Fig 3.15). This is relevant because it was also found to carry minimal NHEJ, but no other genera from the same order were found to also carry both minimal and multiple NHEJ (3.11). Due to this it is more likely that multiple NHEJ was horizontally transferred at some point during the evolution of the genera itself allowing many species to carry the genes. It is possible that in this genera, minimal NHEJ is a sub pathway of multiple NHEJ

that has not yet been seen, however, without *in vivo* experiments it is impossible to know if both pathways are active and function within the genus.

Based on our results, it is plausible that many Archaea actually carry at least one protein involved in prokaryotic NHEJ (Fig 3.16). However, only approximately 18% of Archaea harboring an NHEJ protein possess all the proteins necessary for functional prokaryotic NHEJ. Due to the number of species within the genera *Methanobacterium* and *Thermococcus*, it is possible that NHEJ genes have been vertically transferred through evolution, but note that the only archaeal species NHEJ has been found to be active *in vivo* is *M. paludicola*[4] (Fig 3.17). Although the other Archaea possess all the requisite proteins, they may not be able to actually perform DNA repair through the NHEJ pathway. Their inability to execute the pathway may be attributed to the lack of operon fusion, as observed in Prokaryotes. This could potentially be attributed to horizontal gene transfer events involving individual protein sequences from prokaryotes.

Interestingly, while Asgard lineages exhibit partial prokaryotic NHEJ components, none were found to possess all the required protein sequences. This observation suggests that horizontal gene transfer between prokaryotes or archaea and Asgard lineages may not be as straightforward. Alternatively, it raises the possibility that Asgard lineages may rely upon eukaryotic or eukaryotic-like NHEJ systems rather than prokaryotic.

REFERENCES

1. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA Repair* **7**, 1765–1771 (2008).

2. McGovern, S. *et al.* C-terminal region of bacterial Ku controls DNA bridging, DNA threading and recruitment of DNA ligase D for double strand breaks repair. *Nucleic Acids Res.* **44**, 4785–4806 (2016).

3. White, M. F. & Allers, T. DNA repair in the archaea-an emerging picture. *FEMS Microbiol. Rev.* **42**, 514–526 (2018).

4. Bartlett, E. J., Brissett, N. C. & Doherty, A. J. Ribonucleolytic resection is required for repair of strand displaced nonhomologous end-joining intermediates. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E1984–91 (2013).

5. Weller, G. R. *et al.* Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297**, 1686–1689 (2002).

6. Stephanou, N. C. *et al.* Mycobacterial nonhomologous end joining mediates mutagenic repair of chromosomal double-strand DNA breaks. *J. Bacteriol.* **189**, 5237–5246 (2007).

7. Bowater, R. & Doherty, A. J. Making Ends Meet: Repairing Breaks in Bacterial DNA by Non-Homologous End-Joining. *PLoS Genet.* **2**, e8 (2006).

8. Bertrand, C., Thibessard, A., Bruand, C., Lecointe, F. & Leblond, P. Bacterial NHEJ: a never ending story. *Mol. Microbiol.* **111**, 1139–1151 (2019).

9. Hoff, G. *et al.* Multiple and Variable NHEJ-Like Genes Are Involved in Resistance to DNA Damage in. *Front. Microbiol.* **7**, 1901 (2016).

10. Dupuy, P., Sauviac, L. & Bruand, C. Stress-inducible NHEJ in bacteria: function in DNA repair and acquisition of heterologous DNA. *Nucleic Acids Res.* **47**, 1335–1349 (2019).

11. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).

12. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).

13. Simon, D. M. *et al.* Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA* **14**, 1704–1713 (2008).

14. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

15. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

16. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

17. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

18. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

FIGURES AND TABLES

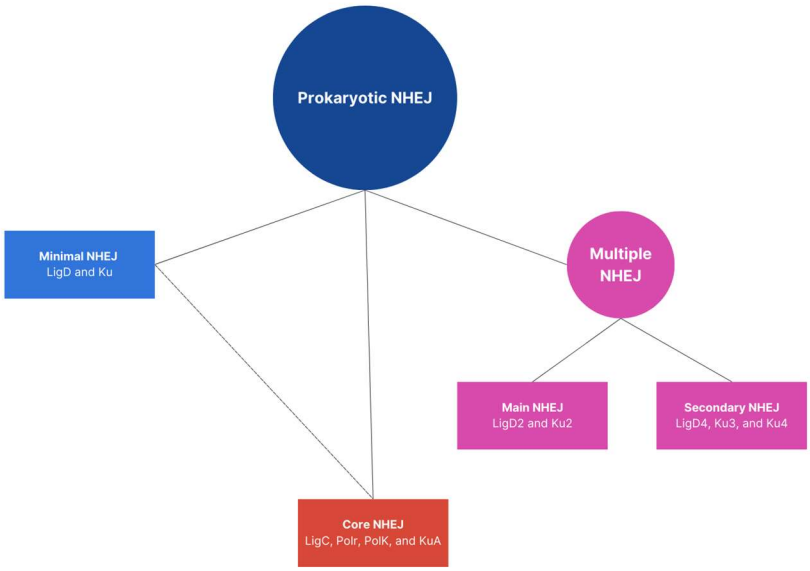

Figure 3.1. NHEJ pathways. A conceptual map illustrating the distinct categories of non-homologous end joining (NHEJ) observed in prokaryotes, accompanied by the corresponding protein constituents for each type. The prokaryotic NHEJ repertoire encompasses three major classifications: minimal, core, and multiple NHEJ. Minimal NHEJ requires LigD and Ku proteins. Core involvesLigC, PolR, PolK, and KuA proteins. Multiple NHEJ comprises two additional pathways: main and secondary NHEJ. Main NHEJ requires LigD2 and Ku2 proteins, whereas secondary NHEJ relies on LigD4, Ku3, and Ku4 proteins.

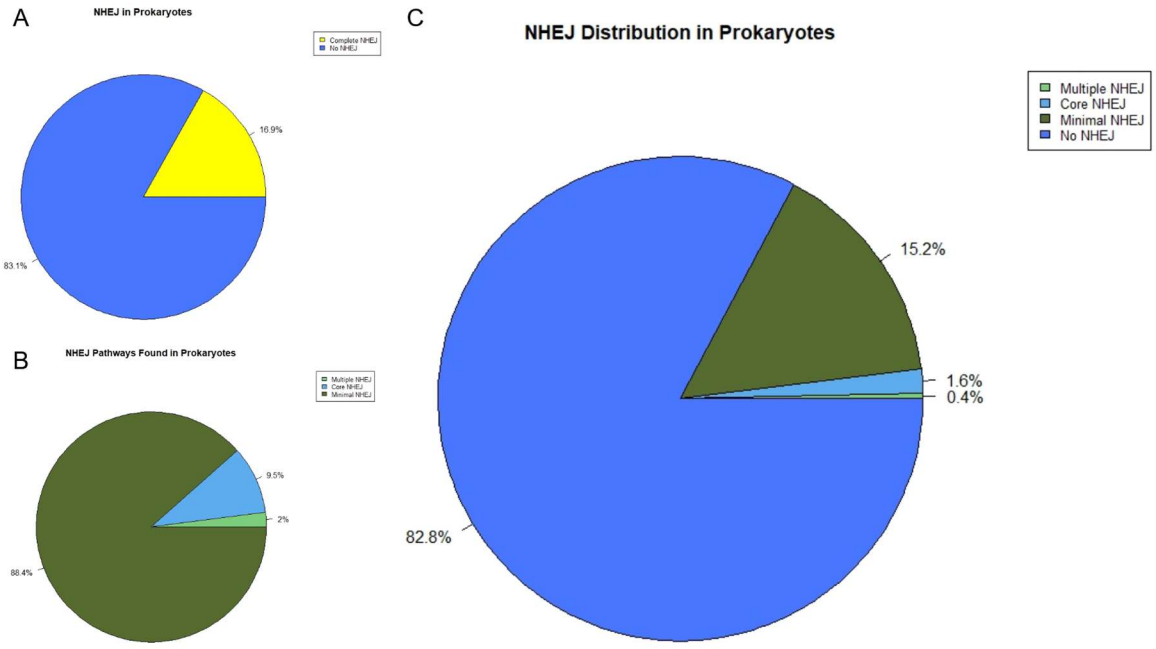

Figure 3.2. Prokaryotes and NHEJ. (A) The distribution of prokaryotic species possessing the complete set of proteins necessary for NHEJ is represented by yellow, whereas blue corresponds to species lacking one or more of the essential proteins. (B) Distribution of prokaryotes possessing distinct forms of NHEJ. The dark green color represents the prokaryotes with minimal NHEJ, while the light blue color indicates the prokaryotes with core NHEJ. The light green color represents the prokaryotes with multiple NHEJ. (C) Distribution of prokaryotes based on their NHEJ type. The blue color represents prokaryotes lacking the NHEJ genes, while the dark green color indicates prokaryotes with minimal NHEJ. Prokaryotes exhibiting core NHEJ are depicted by the light blue color, while the light green color represents prokaryotes with multiple NHEJ.

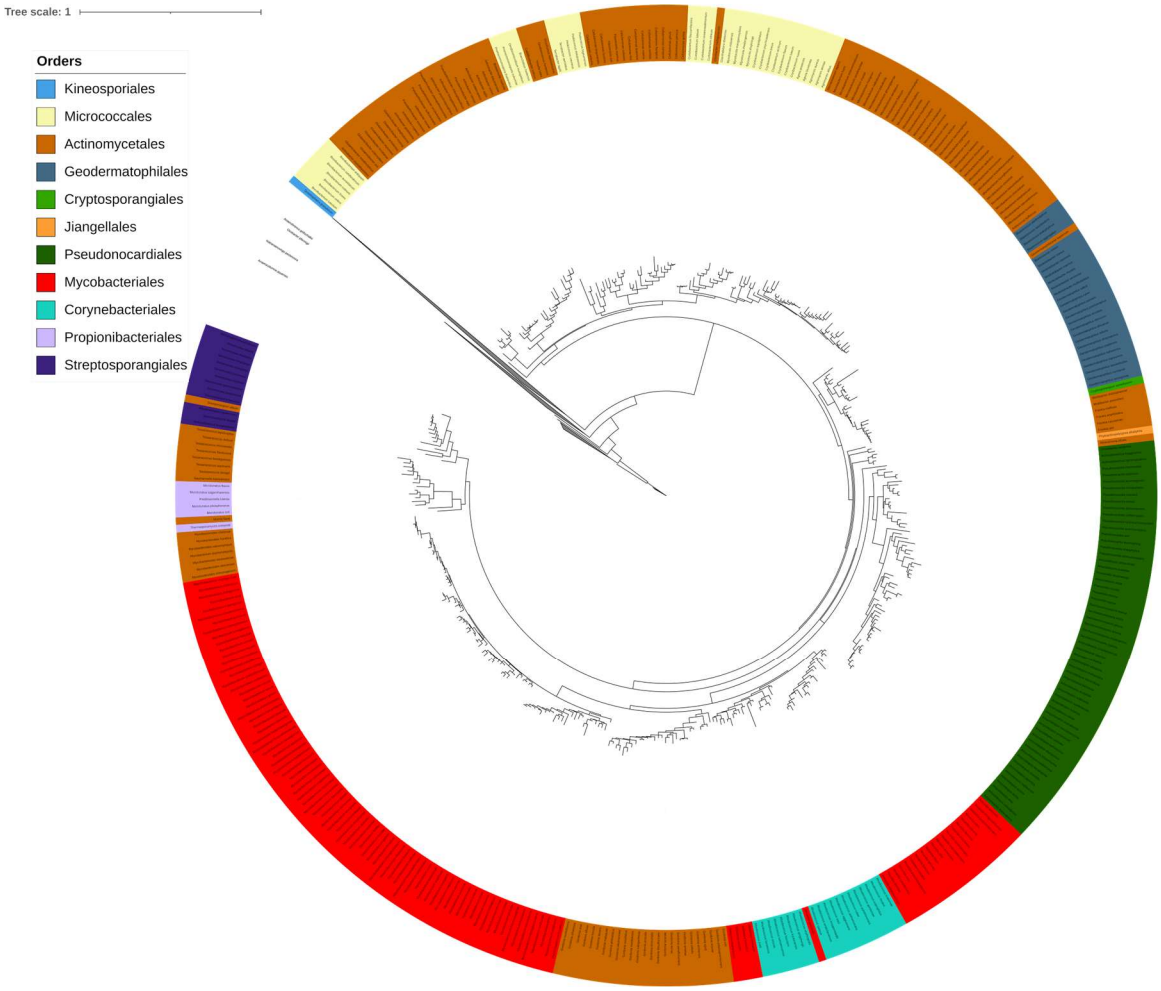

Figure 3.3. Minimal NHEJ pathway in Actinomycetes. A phylogenetic tree of Actinomycetes exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the order the species is from.

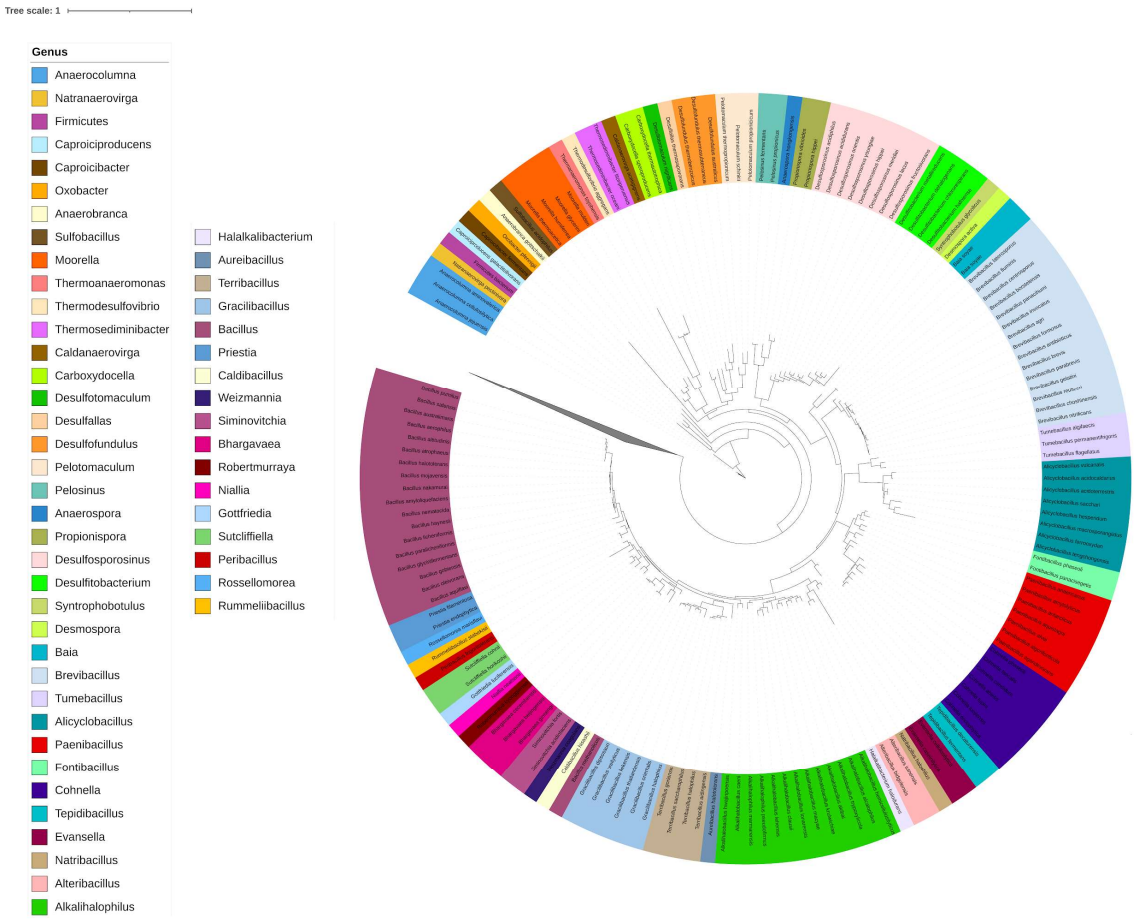

Figure 3.4. Minimal NHEJ pathway in Bacillota. A phylogenetic tree of Bacillota exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

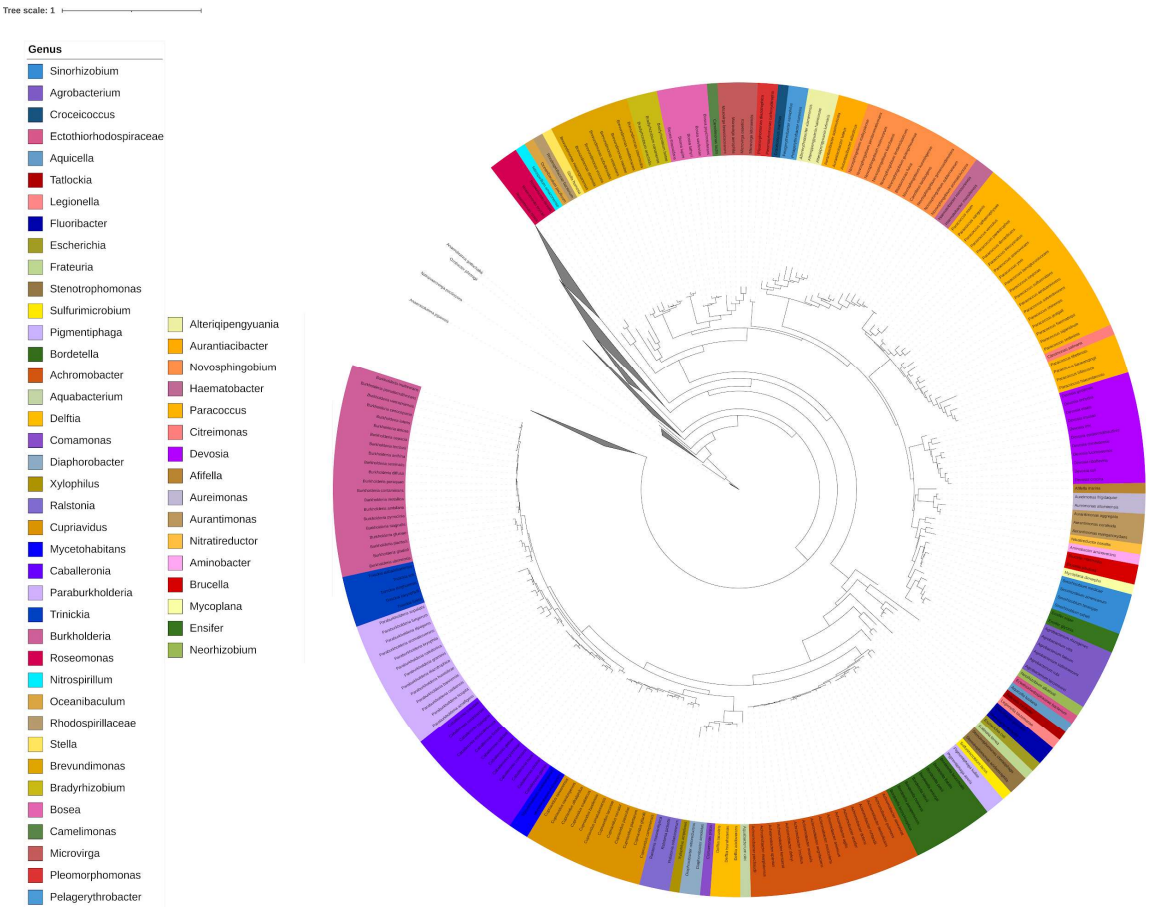

Figure 3.5. Minimal NHEJ pathway in Pseudomondota. A phylogenetic tree of Pseudomondota exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

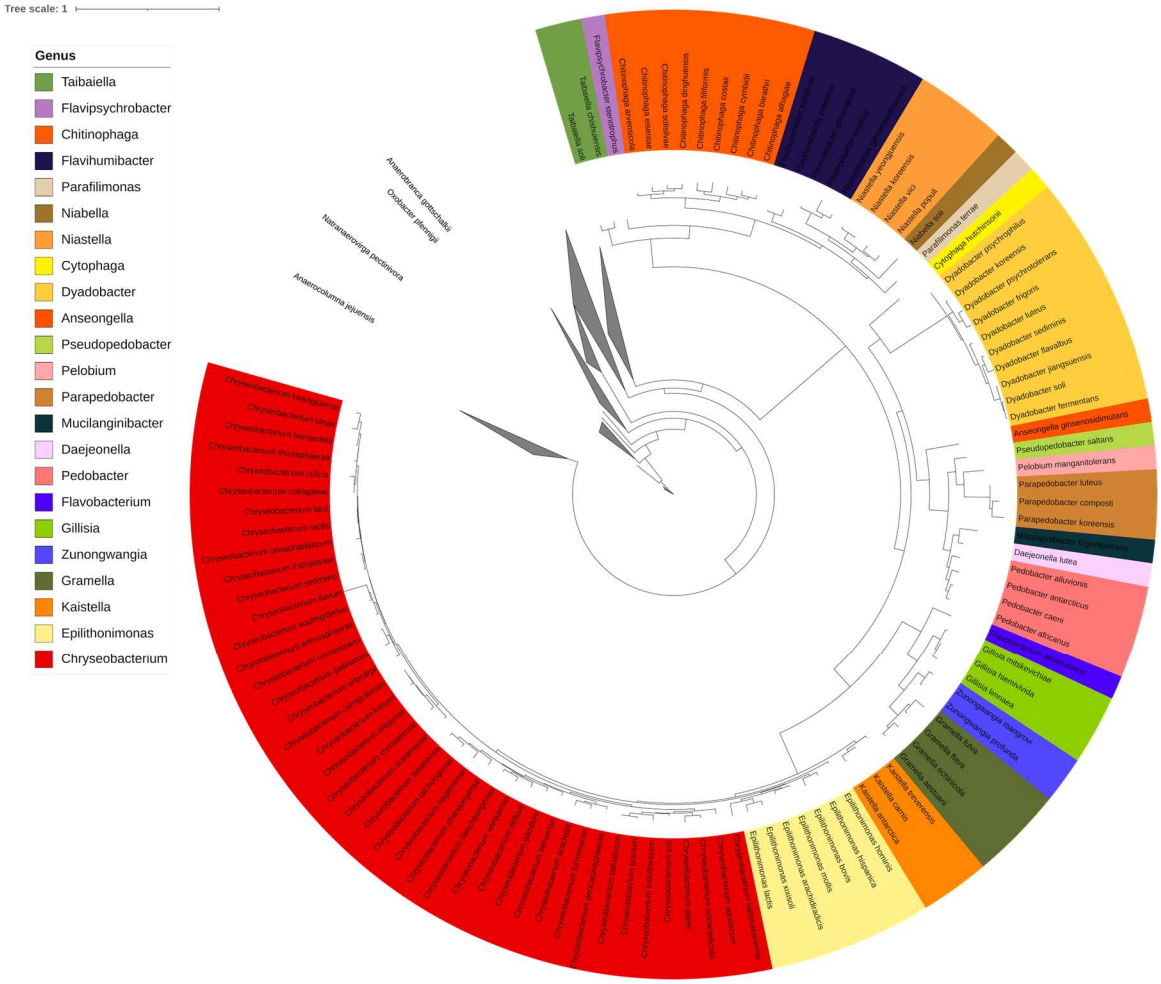

Figure 3.6. Minimal NHEJ pathway in Bacteroidota. A phylogenetic tree of Bacteroidota exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

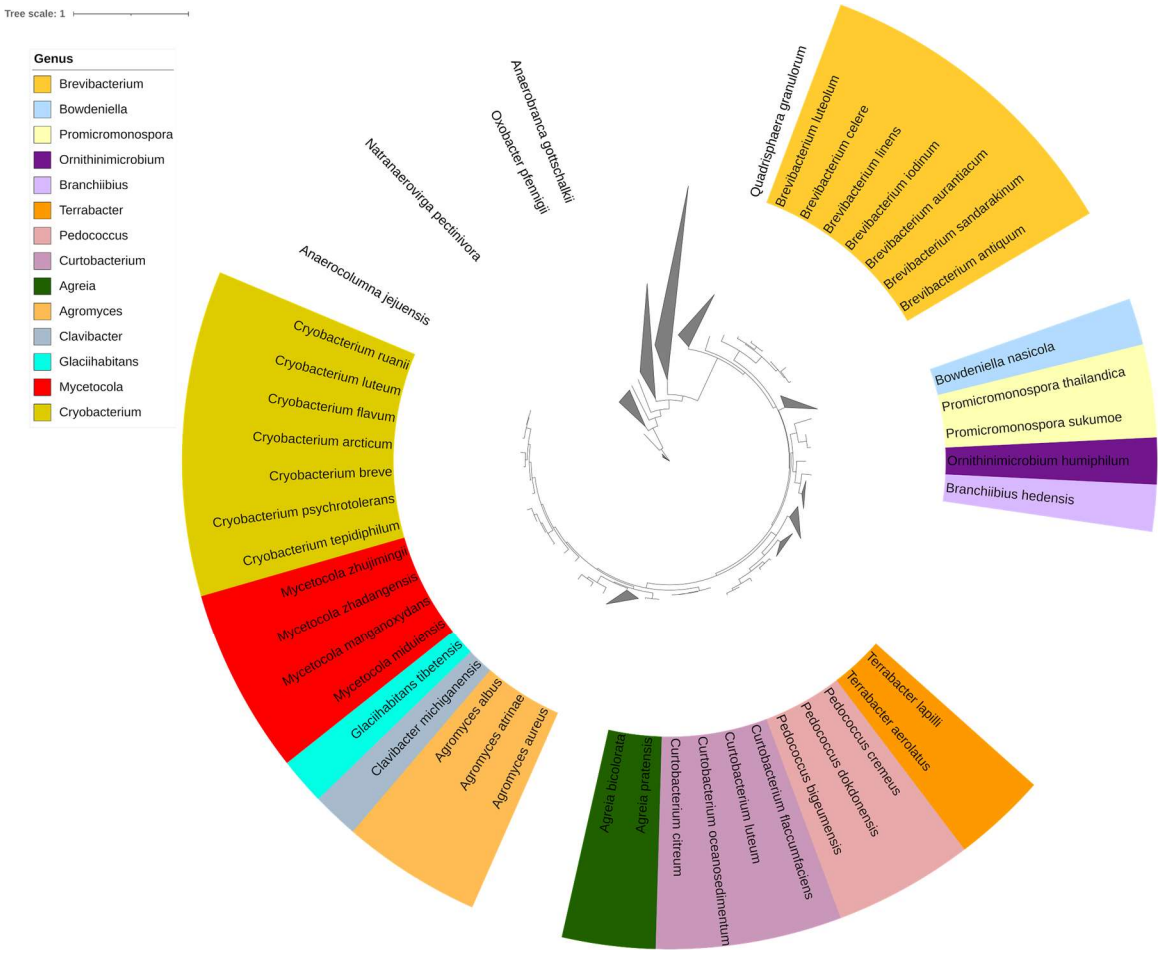

Figure 3.7. Minimal NHEJ pathway in Micrococcales. A phylogenetic tree of Micrococcales exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

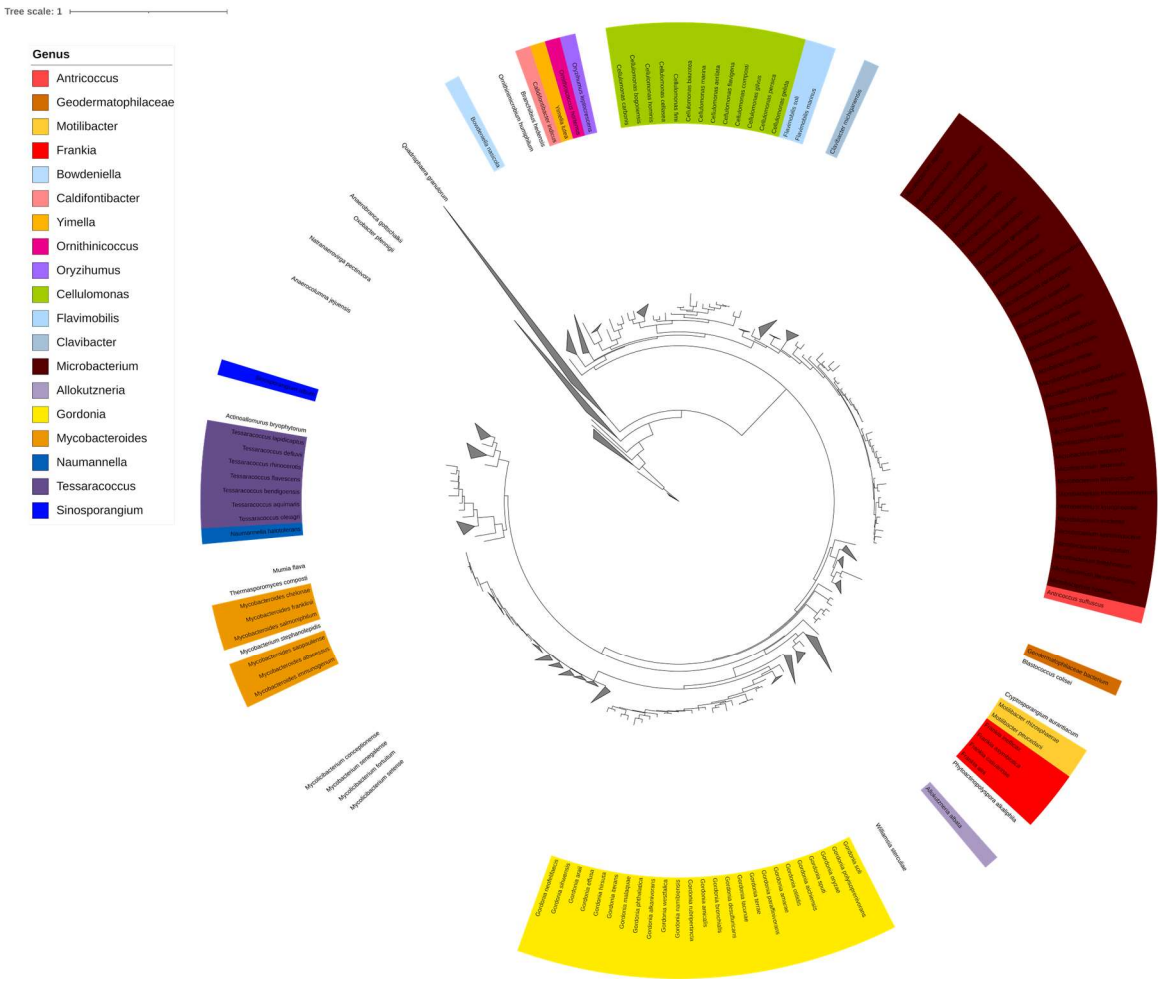

Figure 3.8. Minimal NHEJ pathway in Actinomycetales. A phylogenetic tree of Actinomycetales exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

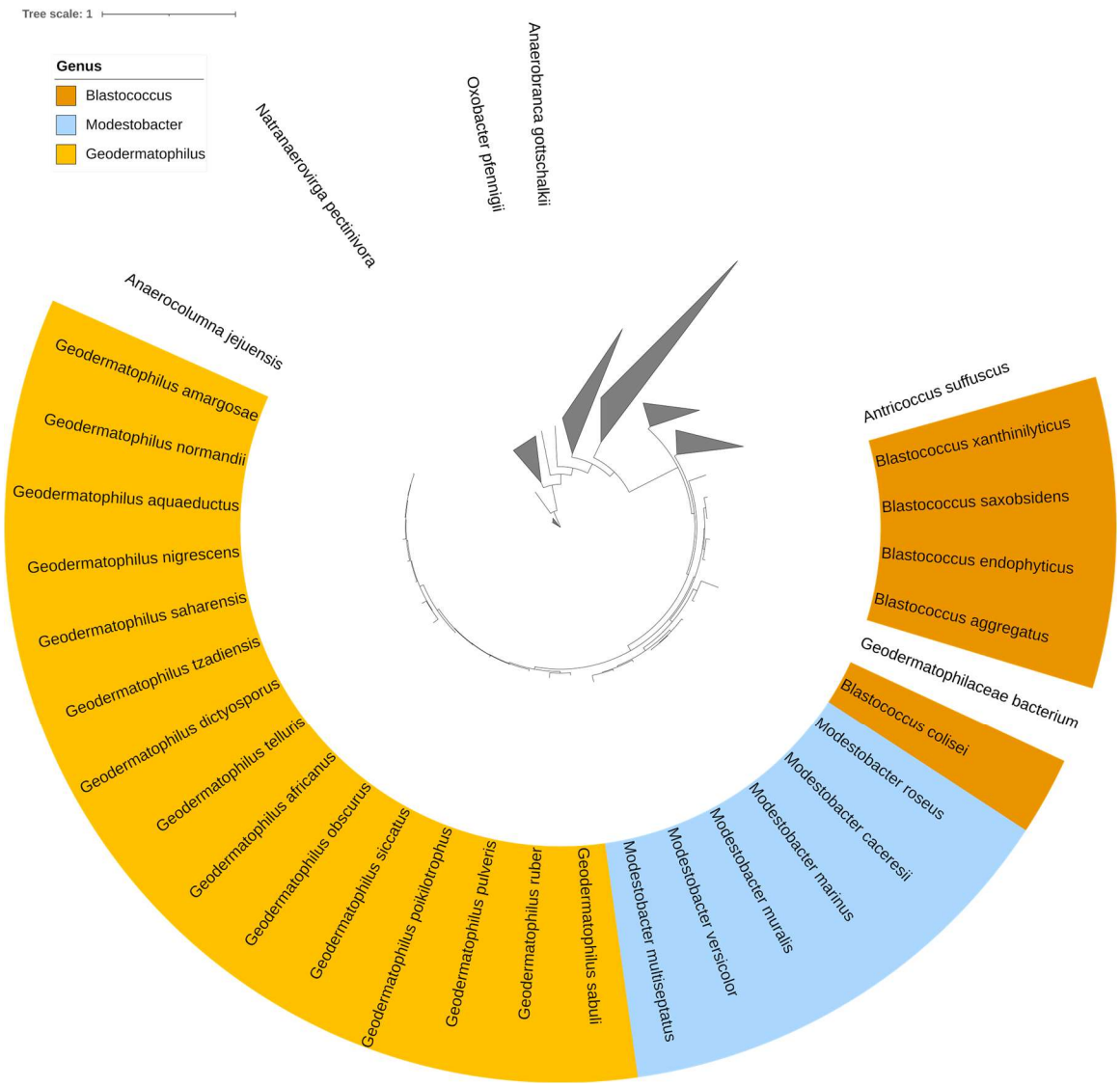

Figure 3.9. Minimal NHEJ pathway in Geodermatophilales. A phylogenetic tree of Geodermatophilales exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

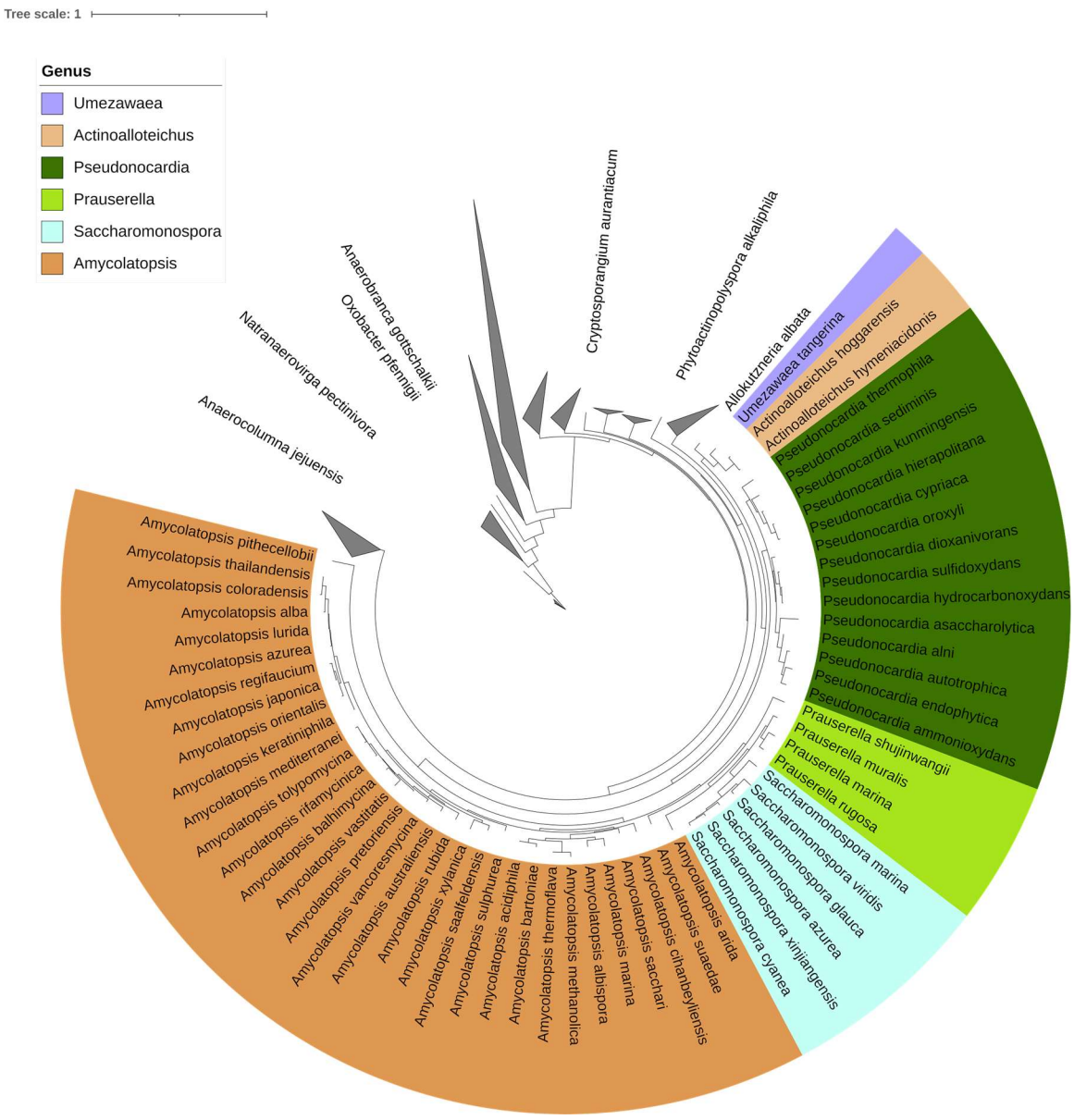

Figure 3.10. Minimal NHEJ pathway in Pseudonocardinales. A phylogenetic tree of Pseudonocardinales exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

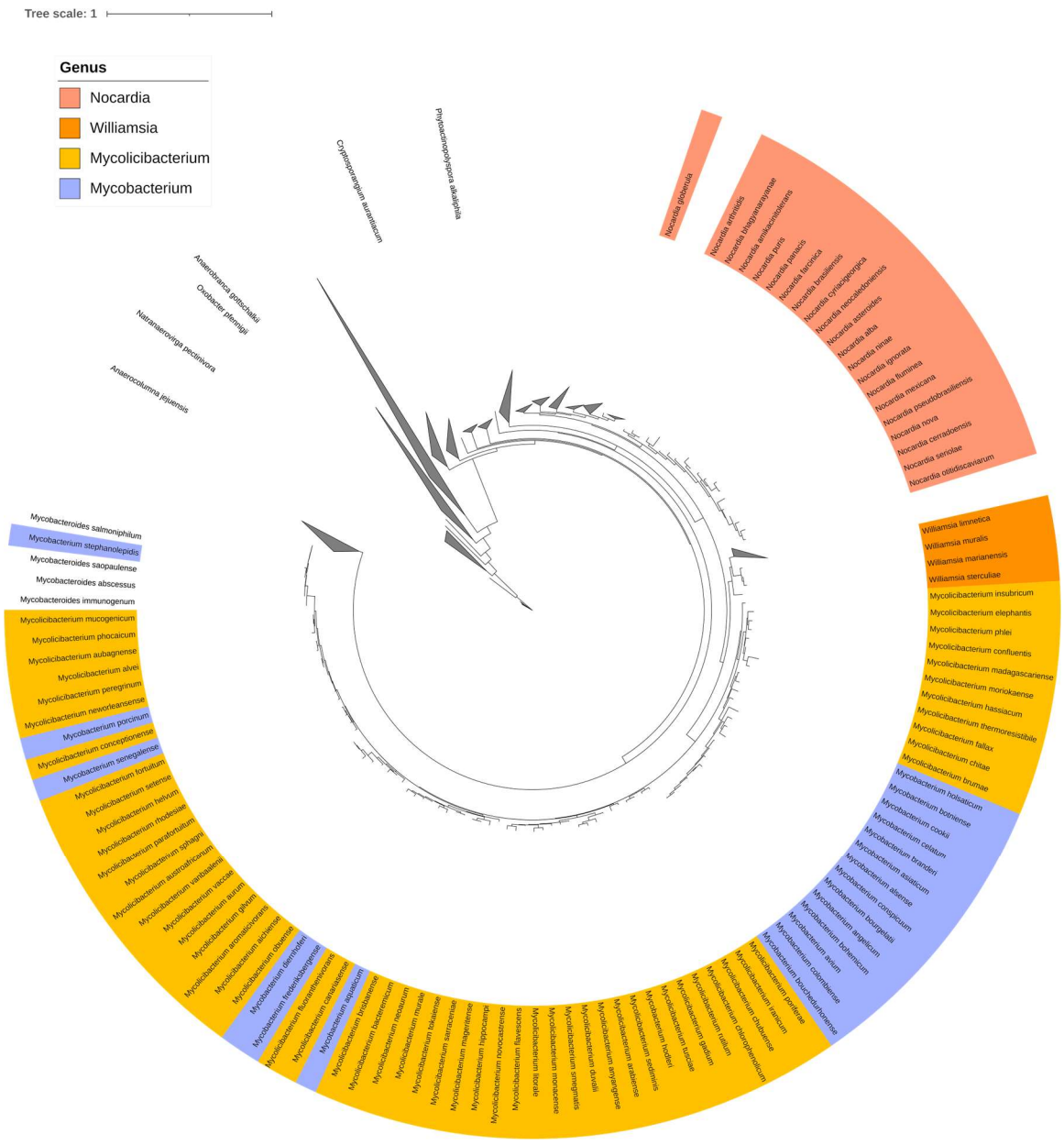

Figure 3.11. Minimal NHEJ pathway in Mycobacteriales. A phylogenetic tree of Mycobacteriales exhibiting minimal NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

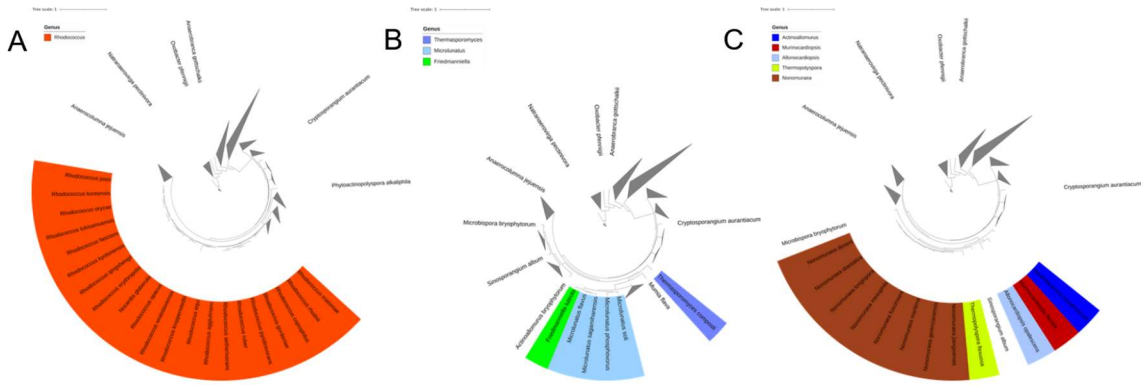

Figure 3.12. Minimal NHEJ pathway in Corynebacteriales, Propionibacteriales, and Streptosprangiales. **(**A) Minimal NHEJ pathway in Corynebacteriales. A phylogenetic tree of Corynebacteriales exhibiting minimal NHEJ. (B) Minimal NHEJ pathway in Propionibacteriales. A phylogenetic tree of Propionibacteriales exhibiting minimal NHEJ. (C) Minimal NHEJ pathway in Streptosprangiales. A phylogenetic tree of Streptosprangiales exhibiting minimal NHEJ. The tree scale bars signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

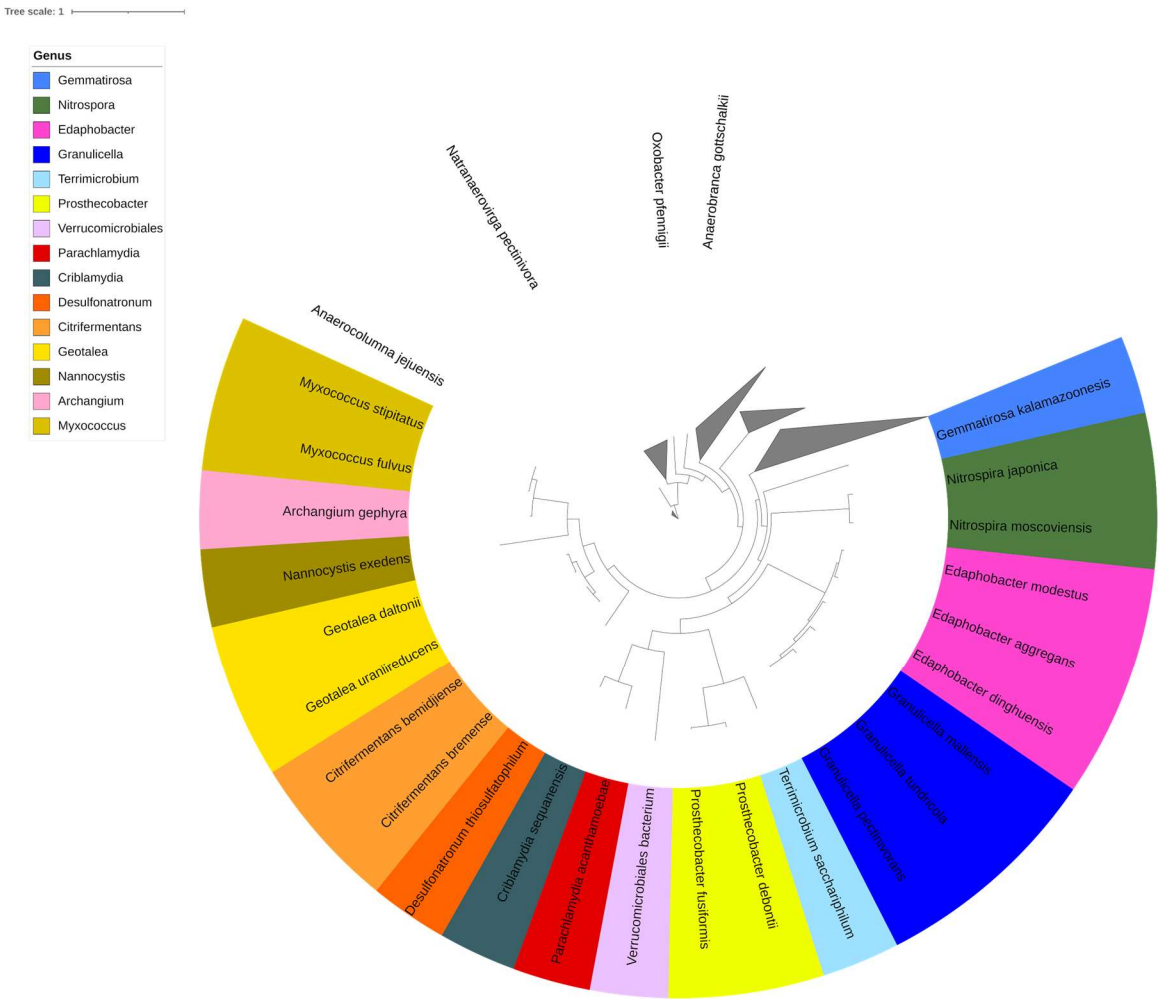

Figure 3.13. Minimal NHEJ pathway in phylums with few species. A phylogenetic tree of species exhibiting minimal NHEJ from phylums with fewer minimal NHEJ occurances. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

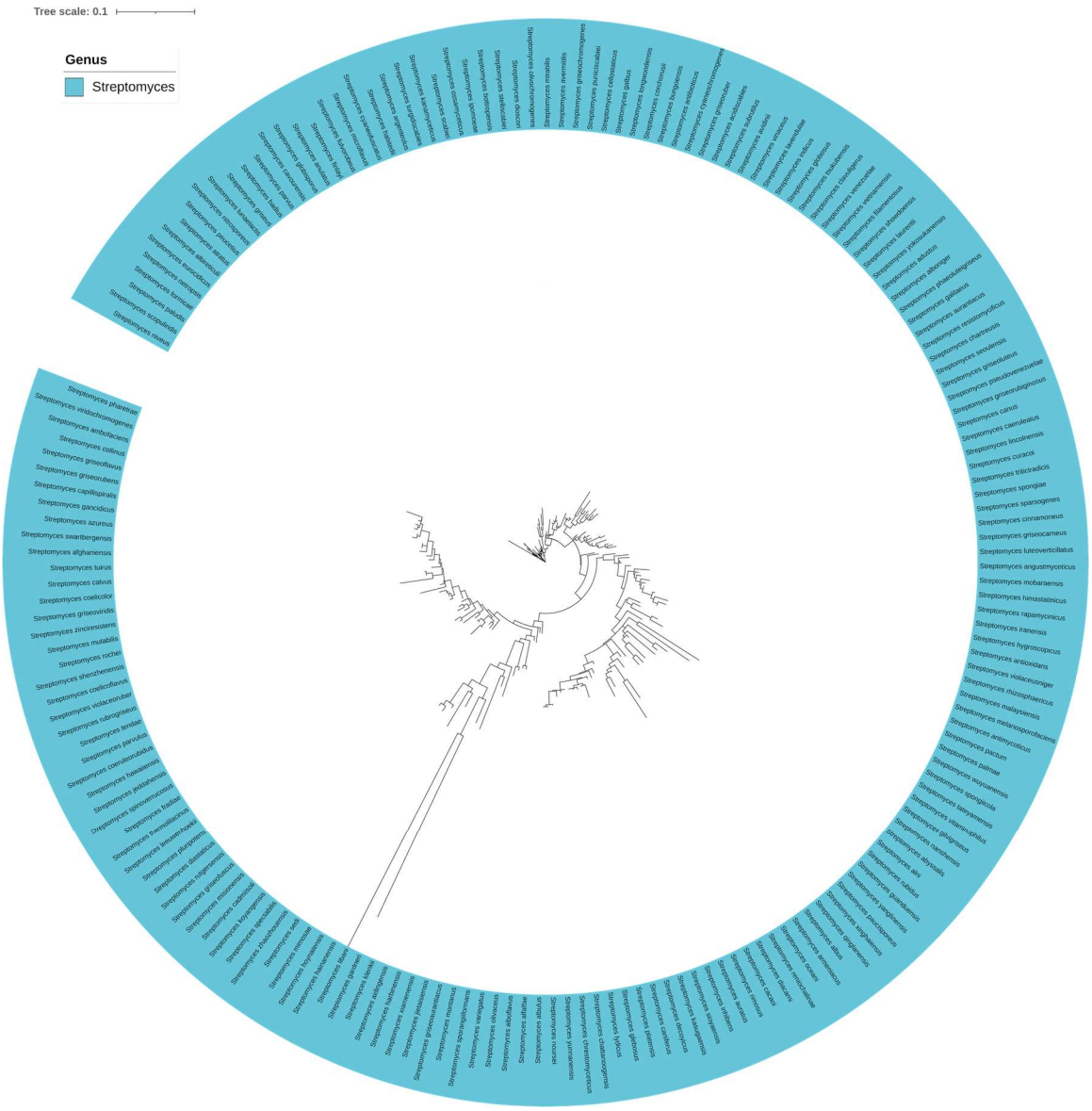

Figure 3.14. Core NHEJ pathway in Prokaryotes. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

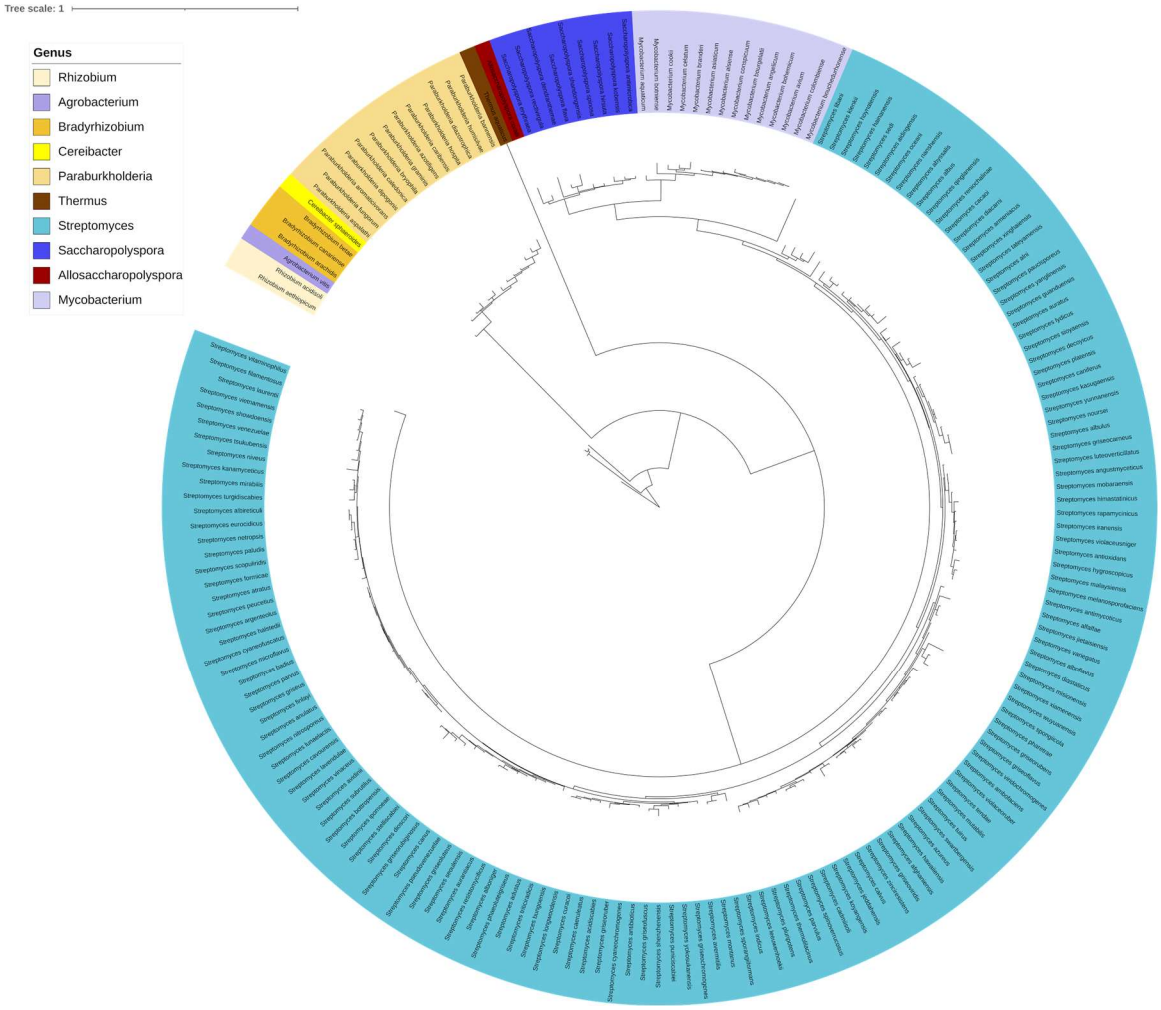

Figure 3.15. Multiple NHEJ pathway in Prokaryotes. The tree scale bar signifies the percent difference between the 16S rRNA sequences. Leaves are highlighted based on the genus the species is from.

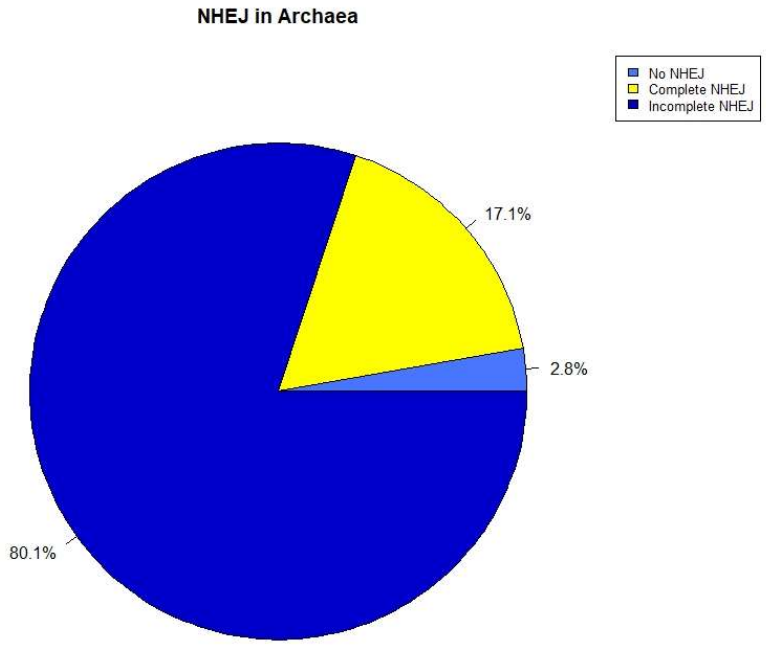

Figure 3.16. Archaea and prokaryotic NHEJ. The distribution of archaeal species exhibiting various levels of prokaryotic NHEJ. The dark blue shade represents the number of archaeal species possessing at least one prokaryotic NHEJ protein. The yellow shade corresponds to the number of archaeal species encompassing all the proteins essential for prokaryotic NHEJ. Finally, the light blue shade represents the number of archaeal species lacking any prokaryotic NHEJ proteins.

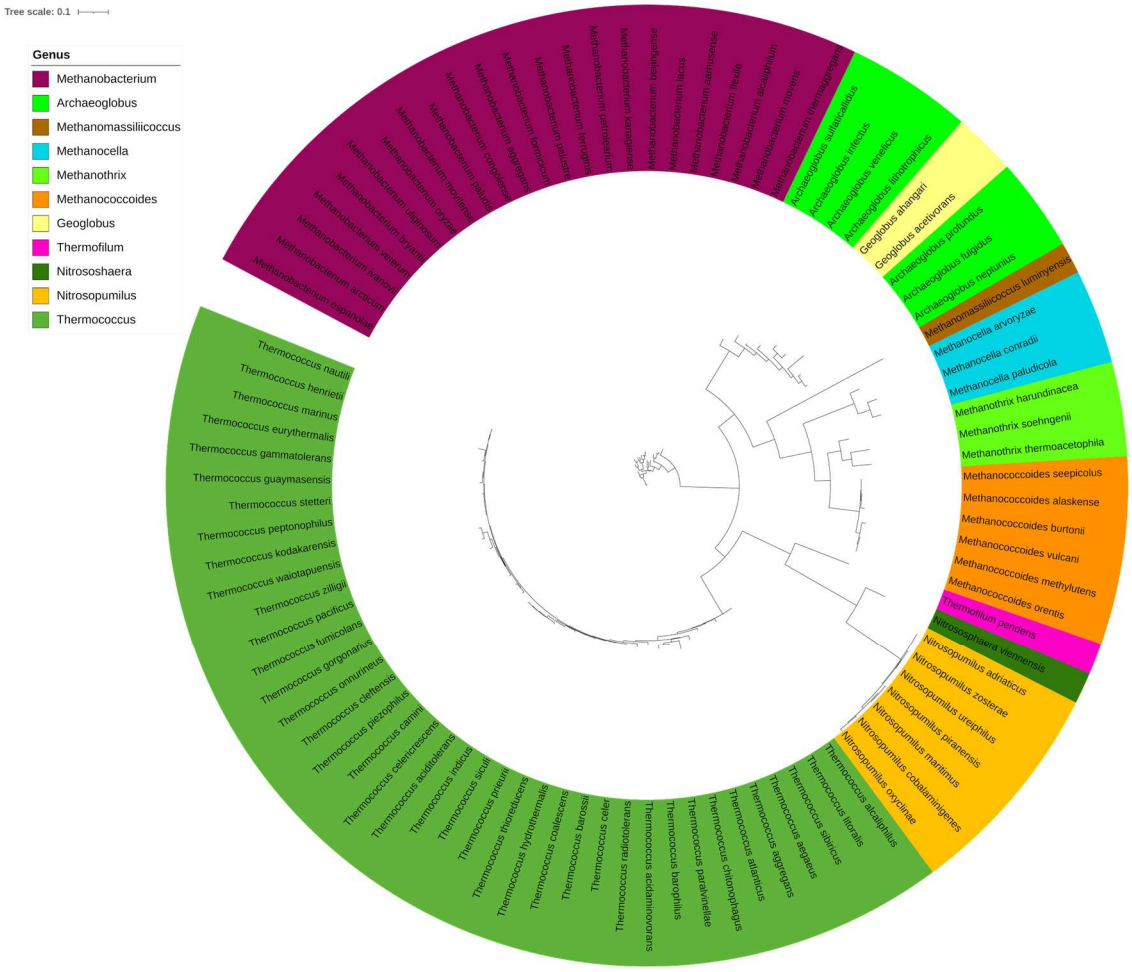

Figure 3.17. Prokaryotic NHEJ in Archaea. Phylogenetic tree that illustrates the evolutionary relationships among archaeal lineages with NHEJ. The tree scale bar signifies the percent difference between the 16S rRNA sequences. The color highlighting each leaf is based on the genus of the species.

CONCLUSION

"I overhyped how good of an idea this is." - Michael Worcester

This dissertation aimed to learn more about potential utilities for genome editing as well as test a novel genome editing tool. Based on results and analysis from this work, it was found that GENEWRITE is able to create edits in plasmids and chromosomes *in vivo* in *E.coli* with the aid of ORF1 of the LINE-1 retrotransposon and NHEJ. While this work is limited to *E.coli*, it creates a foundation for broader applications in eukaryotes or archaea. It also establishes the potential for LINE-1 and transposons as genome editing tools that can be controlled in a laboratory setting rather than simply creating random insertions. Further work is required to determine if GENEWRITE will function in other organisms and evaluate if it can be applied in the biotechnology industry. The application of GENEWRITE in eukaryotes could allow for simpler manipulation of the genome without silencing NHEJ.

Microscopic analysis on TnpB proteins found that transposon activity was 4 - 5 times higher when the transposon was co-expressed with TnpB suggesting its utility during transposition. This supports previous findings that TnpB aids transposition, but if it can be modified to target specific insertion sites similar to Cas proteins remains to be seen. Further work should therefore be completed to evaluate what modifications can make TnpB programmable to cut specific target sites. This could allow for an alternative endonuclease for genome editing to Cas proteins which do not necessarily function in all species as well as provide a better understanding of TnpB and its relation to Cas proteins.

Finally, bioinformatics analysis of NHEJ distribution in archaea and prokaryotes found that while minimal NHEJ is most common in prokaryotes, other forms of NHEJ are also distributed throughout the domain and many archaea carry parts of the NHEJ although they do not express it. This broadens the possibility of more forms of NHEJ existing in all the domains of life and brings up the question of if genome editing tools requiring NHEJ would function with all forms of the DSB repair mechanism. Moving forward, the alternative forms of NHEJ should be considered in prokaryotic work and further work should be done *in vivo* to verify if species carrying NHEJ are able to carry out the pathway. This will give us a better understanding of what species GENEWRITE could function in without the addition of NHEJ to the protocol as well as provide better understanding of the various forms of NHEJ.