COMPUTATIONAL STUDY OF THE EVOLUTION OF BACTERIAL DNA REPAIR SYSTEMS

A THESIS TO BE SUBMITTED TO

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

FOR THE AWARD OF THE DEGREEOF
DOCTOR OF PHILOSOPHY

BY

MOHAK SHARDA

UNDER THE GUIDANCE OF

ASWIN SAI NARAIN SESHASAYEE

NATIONAL CENTRE FOR BIOLOGICAL SCIENCES,
TATA INSTITUTE OF FUNDAMENTAL RESEARCH,
BENGALURU, 560065

JUNE 2022

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013
BENGALURU - 560064

DECLARATION BY THE CANDIDATE

I declare that this thesis entitled "**Computational study of the evolution of bacterial DNA repair systems**" submitted for the award of Doctor of Philosophy to THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY, Bengaluru, is my original work, conducted under the supervision of my guide Dr. Aswin Sai Narain Seshasayee (and co-guide, Dr. Anjana Badrinarayanan). I also wish to inform that no part of the research has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged

I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Place: Bengaluru **Signature of the Candidate**

Date: JUNE 2022

Name of candidate: Mohak Sharda
Reg. No.: 20317030199
Month Year of Admission: March 2017

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013

BENGALURU - 560064

CERTIFICATE

This is to certify that the work incorporated in this thesis **"Computational study of the evolution of bacterial DNA repair systems"** submitted by **Mohak Sharda** was carried out under my supervision. No part of this thesis has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged. I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Research Supervisor:

JUNE 2022

Aswin Sai Narain Seshasayee

Associate Professor
National Centre for Biological Sciences
Tata Institute of Fundamental Research
GKVK, Bellary Road, Bengaluru, 560065

THE UNIVERSITY OF TRANS-DISCIPLINARY HEALTH SCIENCES AND TECHNOLOGY

Private University Established in Karnataka by ACT 35 of 2013

BENGALURU - 560064

CERTIFICATE

This is to certify that the work incorporated in this thesis **"Computational study of the evolution of bacterial DNA repair systems"** submitted by **Mohak Sharda** was carried out under my supervision. No part of this thesis has been submitted for a degree or examination at any university. References, help and material obtained from other sources have been duly acknowledged. I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.

Co-Supervisor:

JUNE 2022

Anjana Badrinaryanan

Assistant Professor
National Centre for Biological Sciences
Tata Institute of Fundamental Research
GKVK, Bellary Road, Bengaluru, 560065

Acknowledgements

I would like to thank Dr. Aswin Sai Narain Seshasayee for his support, encouragement and advice; for providing freedom to pursue the research directions and questions of my interest, all throughout my tenure in his lab. I would also like to thank Dr. Anjana Badrinarayanan for her constant guidance throughout my PhD. I have learnt a whole lot from both of them, something hard to summarize in just a line or two. I thank all other Thesis Committee Members – Dr. Sunil Laxman, Dr. Shivaprasad P. V. for their support and guidance. I also extend my thanks to Dr. Deepa Agashe and Dr. Dasaradhi Palakodeti for discussions and all the support during my PhD. I thank Prof. Sudhir Krishna for going all the way out and helping me in the most difficult phase of my PhD. I would like to thank my mentors from my Master's and Bachelor's degrees who have been instrumental throughout this journey – Dr. Vibha Chaudhary, Prof. Subha Srinivasan, Prof. N Yathindra, Dr. Simran Tandon and Prof. Rajinder Chauhan. I would also like to thank my present lab members– Akshara, Inder, Meghana, Nitish, Shweta, Terence and lab alumni – Aalap, Avantika, Farhan, Pabitra, Parul, Rajalakshmi, Reshma, Revathy, Savita and Supriya for their immense help and inputs. I would also like to extend my thanks all the Anjana lab members – Aditya, Akshaya, Afroze, Asha, Neha, Nitish, Suchita. I would like to thank NCBS IT department, especially Chakrapani, without whose support I would not have been able to carry out any analysis.

I am thankful to Dr. Awadhesh Pandit (NCBS Next Generation Genomics facility), Dr. Devashish Rath, (Bhabha Atomic Research Center, Mumbai), Dr. Varsha Singh (Indian Institute of Science, Bangalore), Dr. Rania Faouzi Zaarour (Gulf Medical University, UAE), Dr. Raefa Abou Khouzam (Gulf Medical University, UAE) and Prof. Salem Chouaib (INSERM, France), to have gotten the chance to collaborate with them on various projects and workshops. It helped me grow in different aspects of research experience. I also thank all the colleagues/ex-colleagues – Vivek Hari Sunder, Khyati Mehta, Sourav B, Noorul Mateen, Surbhi Balan and Inder Raj Singh; I got an opportunity to provide mentorship to, and in exchange got to learn even more in the process. This work would not have been possible without the funding agencies and

institutional help – a) Department of Biotechnology (DBT), Government of India, b) National Center for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, c) EMBO-EMBL fellowship, d) Marine Biological Laboratory, University of Chicago scholarship; I thank them all for supporting me.

I am thankful to a whole bunch of friends on and off campus who have been a great support and made this journey fun and joyful – Ankit, Bulbul, Calvin, Diwyanshu, Dolly, Gaurav, Harsha, Monika, Pragnya, Priya, Ritika, Sagar, Saurabh, Shashank, Shubham, Smit, Suchi, Sukanya, Surbhi, Rania, Vibhushit, Vineeta and Zeeshan. I would also like to thank all my friends from the Latin dance community in Bangalore and dance and improvisational theatre club 'Stage Fright' who have kept me sane throughout. Finally, last but definitely not the least I would like to thank my parents, brother, cousins and a whole bunch of relatives and friends for the huge support and encouragement they have provided me throughout my life and specially during my doctoral journey.

I dedicate this thesis to my grandparents - Dharam Pal Sharda and Sushila Sharda, my parents - Poonam Sharda and Man Mohan Sharda, and my brother, Archit Sharda.

Table of contents

LIST OF TABLES

LIST OF FIGURES

Table of contents

SYNOPSIS

Up until the early 1900s, DNA was thought to be a stable molecule, resistant to any damages and that if any mutations did arise in the macromolecule, it was thought to be due to the inherent errors made during cellular processes like DNA replication (1,2). The evidence for DNA damage and the subsequent repair was first established by the reports on UV radiation mediated damage and DNA photoreactivation (2–5). Since then a number of reports have catalogued different kinds of damages to DNA and related cellular responses, across all domains of life (1,2,6–12). Depending on the metabolic state of the cell, certain regions might be more prone to damage than others, making DNA damage a stochastic process. This stochasticity further dictates the kind of cellular responses that come at play at any given point in a cell(13–15).

Recent studies have suggested a difference in cellular responses to DNA damage across organisms (16,17). These differences occur both at an intra and interspecies level. A number of biological effects have been associated with the diversity of these responses found across all life forms, for example, hypermutator states (18), ability to withstand extreme environments (19,20), pathogenesis (21,22) and cancer rates (23,24) . Most of the studies proposing these associations, however, are one-off experimental studies. Therefore, we still do not know how generalizable these associations are and if at all they are adaptive in nature. Because these studies help us to understand the evolution of these cellular responses and in turn genome evolution, there is a need for research in this direction. Although there have been efforts to understand the evolution of DNA repair mechanisms in recent years, these studies are either based on a small number of genomes or they mostly are survey studies cataloguing the similarities and differences in repair repertoires across organisms at best (25–30). With the surge in genomic datasets, thanks to advancements in DNA sequencing technologies, and ever increasing improvements in algorithms used in the analyses of genomic data, it is now possible to go beyond just reporting the presence and absence of repair machineries and ask questions pertaining to how, when and why different DNA repair mechanisms evolved.

Although we have started to have a good mechanistic understanding of how these repair pathways work, we do not understand how and why they came to be, the way we observe them in extant bacteria. In my PhD study, we have exploited publicly available bacterial genomic datasets, and used comparative, evolutionary and various statistical approaches to design pipelines and understand the evolution of two DNA repair mechanisms – a) Non-homologous end joining and b) Alkylation damage repair protein AlkB.

Chapter 1 reviews the existing knowledge on DNA damage and repair. I have opened the chapter with a historical perspective in the field of DNA repair, followed by different kinds of cellular responses to DNA damage that we know based on studies so far. Finally, I have identified the need and gaps in understanding the evolution of DNA repair mechanisms in bacteria.

Chapter 2 reviews the methodologies that have been employed in carrying out the presented PhD work. Here, I have presented an extensive understanding behind various comparative and evolutionary genomics algorithms. I also discuss the assumptions that the current methods make and how it might impact the results and interpretations.

Chapter 3 talks about the evolution of a bacterial double strand break repair pathway – Non-homologous end joining repair. Using a dataset of around 6000 complete bacterial genomes, I tried to understand the distribution of this repair pathway across the bacterial phylogeny. Further, to understand what processes could have contributed to the patchy distribution, I used ancestral reconstruction methods and various statistical approaches to understand the contribution of horizontal gene transfer. Finally, to understand the factors that could have dictated its evolution, I used various phylogenetic and statistical approaches to understand the association of NHEJ repair presence with three central genome characteristics – genome size, growth rate and GC content.

Chapter 4 talks about the evolution of AlkB, an oxidative demethylase, employed as a direct reversal alkylation damage repair protein in bacteria. Using the same dataset as above, I tried to understand its distribution across bacteria. To understand the selection pressures that could have dictated its evolution, I used phylogenetic, metagenomics and machine learning

approaches to find associations of *alkB* presence and different variables related to microbial habitat and lifestyle. To the best of our knowledge, I have incorporated a machine learning approach to the existing comparative genomics framework to understand the evolution of a DNA repair protein, for the first time. This approach is important, in the sense, it not only helps us test any existing hypotheses that have been proposed, but also helps us generate new hypotheses, based on a phylogenetically diverse set of organisms. This work would serve as a starting point for others interested in the field to experimentally validate the proposed findings.

Chapter 5 presents a general discussion on the findings reported in my PhD work. I talk about the caveats related to the results and the interpretations that follow. As mentioned earlier, there is a dearth of in-depth studies on DNA repair mechanisms to better understand how, when and why they could have evolved. Our study tries to bridge this gap by integrating various approaches in the form of a pipeline. The designed framework can be extended to understand the evolution of other cellular responses to DNA damage. Finally, this chapter also points out problems that remain unanswered and suggests directions for the future studies.

Chapter 1: Introduction

"...We totally missed the possible role of enzymes in repair although, due to Claud Rupert's early very elegant work on photoreactivation, I later came to realize that DNA is so precious that probably many distinct repair mechanisms would exist. Nowadays one could hardly discuss mutation without considering repair at the same time."

> - excerpt from "The double helix: a personal view" by Francis Crick, Nature 1974 (31)

The evidence for the damage of DNA and its repair started to gain traction about a decade before it was even established that DNA was the genetic material in 1953 (2,4,5). Research on DNA repair gained momentum with the seminal works of Whitkin, Radman and Lindahl among others (1,2,6–12). Their work contributed to refuting one of the biggest assumptions put forth by Francis and Crick at the time - *since the genetic material gets passed on to the next generations, the structure of DNA is inherently stable and it is not prone to any changes due to chemicals in the environment. And that if mutations in DNA arise, it is a result of errors made during the replication process instead* (1,2). Evolutionarily speaking, even though their assumption provided a parsimonious explanation of why DNA was selected as the genetic material over other candidates, the establishment of the reactive, damage-prone and repairable nature of DNA brought an intriguing question out in the open. Why was it less costly and easier to choose a multitude of DNA repair pathways along the course of evolution over choosing a better and more stable form of genetic material? Much of the research till date has been dedicated to trying to answer this question. One line of thought that has come out of research in recent years is the possibility of mutagenesis under stress, called stress-induced mutagenesis, supplying the raw material for evolution *i.e.* genetic variation (32). In the future, with increasing efforts in different areas of biology, it would be possible to understand the forces that could have shaped the evolution of DNA repair mechanisms.

Today, an increasing number of studies are supporting the hypothesis that DNA might be under a constant threat of damage and that its repair could serve as an indispensable tool available across different domains of organisms, ensuring the continuity of life (1,2,33). Broadly, DNA damage occurs as a result of an

attack from two kinds of agents - a) *endogenous* or those arising within a cell and, b) *exogenous* or those arising outside the cell. Endogenous damage could arise due to a number of factors like spontaneous mutations as a result of oxidative and hydrolytic reactions and mismatches created by DNA replication errors among others. Exogenous damage could be attributed to ionizing radiation, ultraviolet radiation, chemicals like alkylating agents, cross-linking agents and enzymatic agents. It is to be noted, these distinctions are made for convenience. There can be overlaps in the different damages that could arise from both endogenous and exogenous sources of damage. For example, oxidative stress could arise due to intracellular processes associated with aerobic metabolism and extracellular agents like ionizing radiations.

Further elaborating on endogenous DNA damage, one can broadly divide it into two categories – a. spontaneous alterations in the chemistry of nitrogenous bases and, b. errors in replication created by mismatches.

a. **Spontaneous alterations in the chemistry of nitrogenous bases:** Spontaneous alterations in DNA are a result of its continuous reactivity with water and oxygen. The first kind of alteration is the deamination of the exocyclic amino groups from cytosine, 5-methylcytosine, adenine and guanine, in a temperature- and pH- dependent reaction, resulting in uracil, thymine, hypoxanthine and xanthine respectively. Another kind of alteration is depurination / depyrimidination. In this, bases are lost as a result of N-glycosyl bond cleavage in a pH dependent reaction, leaving the sugar phosphate backbone intact. Under extreme conditions, however, these rates could increase by a significant amount. For example, *Thermus thermophilus*, a thermophile, can optimally grow at 85 degrees Celsius. It could lose as many as 300 purines per genome per generation. Similarly, bacteria growing at 100 degrees Celsius could lose purines at a rate that is 1000 fold higher than those bacteria growing at 37 degrees Celsius. Third kind of alteration arises as a result of oxidative damage to DNA. Oxygen is an important component of energy production across most life forms. However,

paradoxically, reactive oxygen species or ROS are an inevitable by-product of aerobic metabolism. DNA is highly susceptible to attack by ROS. Oxidative damage agents could also react with other macromolecules like unsaturated fatty acids, through a process called as lipid peroxidation. This could give rise to alkylating agents that can in-turn damage DNA via alkylation.

b. **Errors in replication created by mismatches:** Two kinds of errors could occur during replication. First, incorrect bases could get incorporated during DNA replication. Even though the DNA replication polymerase and associated replication enzymes have a fidelity to disallow the incorporation of incorrect bases during the replication process, mismatches could still creep in and lead to mutations in the next generations, if left unrepaired. Second, due to activities associated to aerobic metabolism, nucleotide precursors could get damaged and get incorporated during the replication process. One commonly studied example is the presence of 7,8-dihydro-8-oxoguanine in parental strands after an oxidative damage event of DNA or incorporation of 8-oxo-dGTP instead of dGTP in the daughter strands during replication. This damaged base gets incorporated against an Adenine and could escape proofreading.

Similarly, exogenous DNA damage could be divided into two broad categories – a. Radiation induced DNA damages and, b. Chemical agents induced DNA damages.

a. Radiation induced DNA damages: Broadly, there are two sources of radiation induced DNA damage – 1. Ionizing radiation and, 2. Ultra-violet radiation. These two have been a source of physical damage to DNA since the beginning of organismal evolution. The exposure of Ionizing radiation is estimated to be on average more at higher altitudes as compared to habitats at the sea level. Natural radioactivity from the soil also depends on the local geography. IR induced damages could either directly affect DNA or indirectly affect DNA via production of ROS. Ionizing radiations could additionally damage the sugar-phosphate backbone causing strand breaks. It is estimated that 20%

of the hydroxyl groups that react with DNA affect the sugars in the sugar-phosphate backbone. Second, UV radiation spectrum can be divided into three types, depending on the wavelength: a. UV-A, b. UV-B and, c. UV-C. UV-C damage (100 – 295 nm) is specific to DNA. The most common photoproduct is the cyclobutane pyrimidine dimer. Strand breaks from UV-C have not been observed. Such breaks are only possible, as has been shown experimentally, only at longer wavelengths of UV. Other than the direct absorption of photons by DNA bases, UV radiation induced damage can also be caused indirectly via sensitizer molecules like ketones (acetophenone) and oxygen (giving rise to reactive oxygen species). These photosensitizers could be endogenous (aromatic amino acids, riboflavin etc.) or exogenous (certain drugs) in nature.

b. **Chemical agents induced DNA damage:** Broadly, they can be classified as follows: 1. Alkylating agents and, 2. Cross-linking agents. Alkylating agents are electrophilic compounds that attack the nucleophilic centers in different nitrogenous bases. In general, the ring nitrogen of the bases are more nucleophilic than the oxygen, where the N7 position of guanine and N3 position of adenine being the most reactive. Direct-acting methylating compounds exist in the environment. For example, MeCl or Methyl chloride is one of the most abundant environmental mutagen and carcinogen, produced as a result of biomass burning and biosynthesis by microorganisms and marine algae. Another naturally occurring antibiotic, Streptozotocin, is produced by the soil bacterium *Streptomyces achromogenes,* Methylating agents have been shown to be produced *in vitro* by nitrosation of endogenous metabolites catalyzed by bacterial enzymes. For example, catabolite methylamine reacts with carbamyl phosphate, a precursor of pyrimidines, to give methylurea, which in turn can be nitrosated to yield MNU (methylnitrosourea), a potent alkylating agent. Another common source of alkylating agents is the reactive alkali radicals and nonradical products generated through lipid peroxidation chain reactions. Cross linking agents could either cause inter-strand DNA cross-links or intra-strand adducts. Endogenous sources of inter-strand DNA cross-link causing agents include one of the product of normal cellular glycolysis, acetaldehyde. Aldehydes formed as products of lipid peroxidation could also cause these

cross-links. A classic example of exogenous cross linking agent is mitomycin C, naturally produced by *Streptomyces lavendulae.*

All life forms have thrived in the presence of different kinds of agents including oxygen and water - some that have existed long enough and some only in the recent timescales, giving rise to different types of damages across environments - providing selection pressures for the evolution of appropriate responses within a cell.

Different cellular responses get activated across different organisms once the DNA gets damaged:

1) **DNA damage repair (DDR)**: DNA repair is strictly defined as a mechanism that restores the original DNA sequence and structure. The damage could be divided into two kinds, targeting either the nitrogenous bases or sugar-phosphate backbone.

a) *Damage to bases* could be either reversible or irreversible in nature. Reversible damages, like the addition of alkyl moieties, could be repaired by an ensemble of proteins that form a part of the adaptive response pathway (ARP). Irreversible base damages could be repaired by excision of nucleotides or free bases. Three common repair systems employed during irreversible damage are mismatch repair (MMR), nucleotide excision repair (NER) and base excision repair (BER).

b) *Damage to the sugar-phosphate backbone* could give rise to strand breaks, either single strand breaks (SSBs) or double strand breaks (DSBs) or both. Organisms take advantage of the homologous copy of the DNA to repair strand breaks preferentially by a process called recombination based repair (RBR). Under an event of a double strand break, in situations where the second copy is unavailable, a relatively error-prone repair is carried out, called Non-homologous end joining (NHEJ) repair.

2) **DNA damage tolerance (DDT)**: DDT is defined as a mechanism that aids the bypass of lesions encountered by the DNA replication machinery. In other words, unlike DNA repair mechanisms that help in the removal of DNA

damage, DDT pathways prevent stalling of DNA synthesis, even before the damage has been attended to; thus mitigating the associated lethality of arrested replication fork(s) or transcription stalling. Depending on the involvement of the damaged strand, DDT mechanisms fall under two categories.

a) Error free: These mechanisms do not require the damaged DNA strand as a template for DNA synthesis. Therefore, these mechanisms allow for the persistence of damage in the genome without generating mutations in the newly synthesized DNA.
    i) Recombination based repair - DNA synthesis is reinitiated downstream of the site of arrested replication. The gap generated in the affected newly formed DNA can be filled concomitantly by recombination between the damaged and the undamaged newly synthesized daughter DNA duplexes.
    ii) Replication fork regression - DNA synthesis at the site of the arrested replication fork is carried out using the newly replicated template strand instead of the parental damaged template strand. This involves folding the arrested replication fork back on itself.
b) Error prone: Primarily the translesion DNA synthesis, this mechanism can cause mutations at sites at and around the arrested replication forks, also called mutagenesis. They employ a switch from high fidelity replication polymerases to low fidelity error-prone polymerases that can resume replication albeit with incorporation of promiscuous nucleotides. Mutagenesis by Translesion Synthesis (TLS) is thought to be one of the primary sources of induced mutations in cells.

3) **DNA damage checkpoint (DDC)**: Conventionally discussed in eukaryotic model systems, DDCs are points in the cell division cycle that are sensitive to DNA damage. They result in an arrest of cell cycle progression. This buys repair and tolerance mechanisms time to attend to the damaged sites.

4) **Cell death**: Other than the above cellular responses, as a strategy to weed out cells with irreconcilable damage, the point of no return, from the cellular population(s), programmed cell death mechanisms are employed.

DNA damage is a stochastic process. Depending on the metabolic state of the cell, certain regions might be more prone to a particular kind of damage than others. The choice of cellular responses depends on a number of factors with respect to the genomic region under attack (14). Some of these factors that we have started to understand, include - Is the region replicating or non-replicating? (15) Is it transcriptionally active or silent? (13) Furthermore, multiple cellular responses could act at the same time, either within the same cell or at a population level in a subset of cells.

The diversity of cellular responses to DNA damage exists across all organisms, intra- and inter-specific, at two levels. First, in a given clade, the exact pool of repair mechanisms could differ among species. For example, photoreactivation repair pathways are present in prokaryotes like *E. coli* and eukaryotes like *S. cerevisiae*, whereas other eukaryotes like humans lack photoreactivation (17). Second, other differences in repair pathways among species exist in the subtypes of a given repair pathway. For example, it has been suggested previously that the substrates for MMR pathway are highly species specific (16). These differences in the diversity of cellular responses to DNA damage can have different downstream effects. For example, certain studies have suggested that the increased mutation rate in the human pathogen Mycoplasma may be partly due to lack of DNA repair pathways (34,35), while a hypermutator state is linked to the presence of NHEJ in *P. putida* (32). Other biological effects that have been associated with a difference in repair repertoires include codon usage and GC content (36,37), evolutionary rates (38), speciation events (39,40), pathogenesis (21,22), cancer rates (23,24), increased mutation rates (18) and withstanding extreme environments (19,20).

The association of these different biological effects with repair systems have been based majorly on one-off repair characterization studies in different species, proposing similar associations might exist in other, if not all, species. For example, a Nucleotide excision repair independent pathway called the MrfAB (consisting of a helicase and an exonuclease) has been reported in the

soil bacterium *Bacillus subtilis*. This study by *Burby et. al.* showed that deletion of MrfAB results in sensitivity to mitomycin C, a DNA cross-linking agent, but not to any other type of DNA damage tested. The study suggested that this repair pathway could be an adaptation to environments with mitomycin C producing bacteria like *Streptomyces lavendulae*. Collectively, these studies are important in understanding the evolution of different repair proteins. An evolutionary perspective provides the depth in understanding repair systems that goes beyond just identifying and characterizing the differences and similarities that exist among organisms. In that, they help us understand "how" and "why" those differences and similarities might have come to be. For example, a recent study (147) has suggested that photolyase—catalyzed direct damage reversal (DR) and apurinic/apyrimidinic endonuclease dependent nucleotide incision repair (NIR) were most likely present in the Last Universal Common Ancestor (LUCA). These two mechanisms require less energy and very few steps and are not dependent on a multitude of proteins to carry out repair. Conditions on Earth in the beginning presented life forms with a limited DNA damage, owing to factors like Infrared and Ultra-violet radiations, and that DR and NIR would have been sufficient to repair those damages. With the oxygen catastrophe that occurred much later during evolution, there was an increase in the complexity and spectrum of DNA damage. This most likely corresponded to the appearance of multi-protein complexes associated with Base Excision repair (BER) and Nucleotide Excision repair (NER), that presented more efficient and versatile catalytic mechanisms to combat DNA damage emerging from oxidative stress.

Given the recent advances in genomic sequencing technologies since the early 2000s, it is now possible to incorporate comparative frameworks to understand the evolution of repair pathways across organisms; something that one-on-one studies lack to offer. There is, however, a dearth of such comparative genomic studies. Studies in the existing literature are either based on only a small number of genomes or are focused on only a few repair pathways lacking in-depth exploration (25–30).

The work I present in this thesis, takes advantage of recent advances in sequencing technologies resulting in a surge in the number of completely

sequenced genomes and improvements in techniques like phylogenomics, phylogenetic comparative methods and biostatistics, to

1) design comparative and evolutionary frameworks and,
2) use them to understand the evolution of different repair proteins

In **Chapter 2**, I review various statistical methods and algorithms employed in the field of comparative and evolutionary genomics that have been used in the work presented in this thesis.

In **Chapter 3**, I take a case study of an error-prone double strand break repair mechanism, Non-homologous end joining, to understand the role of different processes like gene gain-loss and horizontal gene transfer events in shaping its evolutionary history. Furthermore, I study its association with three central genome characteristics - genome size, growth rate and GC content.

In **Chapter 4,** I take an alternate strategy to understand the evolution of *alkB*, an oxidative demethylase involved in direct reversal repair of lesions targeting nitrogenous bases under alkylation stress. I incorporate a machine learning based approach to understand the factors that could be dictating the distribution of *alkB* across bacteria.

In **Chapter 5**, I present a general discussion on the findings and the implications reported in this study. I also point out problems that remain unanswered and suggest directions for the future studies.

Chapter 2: Review of evolutionary and comparative genomics methods

2.0 Statistical primers

2.0.1 Maximum likelihood

Let us understand how maximum likelihood (abbreviated as ML for this section) can be applied to a statistical analysis by using a simple coin toss example. We would later see how this would be used in estimating molecular phylogenies in the sections to come.

For an ML analysis, one requires an observed data and a probabilistic model for how the data that we observe could have been produced. A probabilistic model allows one to compute the probability of any given outcome for a set of specific model parameter values. What that means will be clear with the example that follows below.

Let us generate some data by tossing a coin ten times. Let us say, the observed data that we get is 6 heads and 4 tails.

Let us define an appropriate model that could have generated this data. Let us say that the coin had a probability *p* for generating heads and *1-p* for generating tails. Using a binomial model, the probability of observing x heads among n tosses is given by:

$P(x \text{ heads}) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x}$

Given the observed data and the probabilistic model, what maximum likelihood allows us to do is to calculate the “best” point estimate parameter value *p* that could have generated the data. Please note, unlike the multitude of parameters that are estimated when constructing phylogenies, as we will discuss, here we are concerned with just one parameter, *p*. More formally, a likelihood is the probability of data given a probabilistic model. As follows, the best point estimate is the parameter value that gives the highest possible likelihood.

In our example, we find that the best parameter estimates of p that explains the observed data of 6 heads/successes and 4 tails is p=0.6 with,

Likelihood (p=0.6 | x=6) = Probability (x=6 | p=0.6) = 0.25 (see Table 2.1)

p	x	n	Likelihood
0.1	6	10	0.00013
0.2	6	10	0.0055
0.3	6	10	0.036
0.4	6	10	0.111
0.5	6	10	0.205
0.6	**6**	**10**	**0.2508**
0.7	6	10	0.2001
0.8	6	10	0.088
0.9	6	10	0.011

Table 2.1 Maximum likelihood table This table outputs the likelihood for different parameter values, p. Here number of successes, x, are 6 and number of trials, n, are 10.

In the coin tossing example above, the parameter estimate can be calculated using both analytical methods, like the binomial formula above, and numerical simulation methods (as would be explained in the next sections). When dealing with phylogeny construction, where estimating parameter values like branch lengths, topologies etc., analytical solutions are not available, we find the best phylogeny (and the corresponding parameters) estimate(s) using numerical solutions.

2.0.2 Bayesian inference

In the scope of this thesis, algorithms that implement Bayesian inference have been used to study character trait evolution in bacteria, as would be discussed in the next sections. We will take a different example to understand Bayesian inference - rolling a dice and the proportion of times, *p*, the number 6 comes up.

In general, the Bayesian inference involves the following steps:

1) **Defining a prior:** express an opinion about the proportion *p* before sampling.

2) **Estimating likelihood:** take a sample and record the proportion that 6 comes up

3) **Posterior:** use Bayes' rule to update the prior belief *p* given the information from the data sample.

2.0.2a Defining the prior:

The proportion of times 6 comes up could be anywhere between 0 and 1. Let us assume that 0 and 1 are not likely proportions here. The number of values then proportion can take, with an increment of 0.1 are,

$$p_i = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$$

Next, we assign probabilities to all the possible proportion values defined above. In other words, this would entail our prior probabilities for these proportion values. If you are carrying out the experiment for the first time, and using the dice for the first time, a safe bet is to consider a Uniform prior (π) i.e all proportions are equally likely.

$$\pi_{Uniform}(p_i) = (1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9)$$

If, however, you have some prior experience using the dice, and let's say you have a hunch that not all proportion values are equally likely, you can reflect your belief in the prior probability distribution. Let us say that you have a hunch that 0.6 and 0.7 proportions are twice as likely as other proportions. The corresponding Non-uniform prior can then be calculated,

$$\pi_{Non\text{-}Uniform}(p_i) = (0.09,0.09,0.09,0.09,0.09,0.18,0.18,0.09,0.09)$$

Let us work with the Non-Uniform prior, assuming you have some experience working with the dice. This can be written in a tabular format as follows:

proportion	prior probability
0.1	0.09
0.2	0.09
0.3	0.09

0.4	0.09
0.5	0.09
0.6	0.18
0.7	0.18
0.8	0.09
0.9	0.09

2.0.2b Estimating likelihood:

The next step is collecting data and estimating likelihood (**L(p_i)**). Let us say you roll the dice 20 times and you get a six 12 times i.e. your success. The likelihood of this experiment can be calculated using a *binomial model*, as done in the previous section using the binomial formula. Here n = 20 and x = 12. We can update the table above, with the likelihood,

proportion, p_i	prior probability, $\pi_{Non\text{-}Uniform}(p_i)$	Likelihood, **L(p_i)**
0.1	0.09	5.4e-08
0.2	0.09	8.6e-05
0.3	0.09	3.8e-03
0.4	0.09	3.5e-02
0.5	0.09	1.2e-01
0.6	0.18	1.8e-01
0.7	0.18	1.1e-01
0.8	0.09	2.2e-02
0.9	0.09	3.5e-04

2.0.2c Calculating posterior distribution:

The last part of Bayesian inference is updating your prior belief based on the data that you observe i.e posterior belief.

We will use Bayes' formula,

$$\boldsymbol{\pi(p_i|x)} = \frac{\pi(p_i)\ x\ L(p_i)}{\sum_j \quad \pi(p_j)\ x\ L(p_j)}$$

$$\boldsymbol{\pi(p_i|x)} \propto \pi(p_i)\ x\ L(p_i)$$

where, the term on the left is the posterior probability of a given proportion given the observed number of successes, here x=12. The numerator on the right is prior belief times the likelihood for the given proportion, as can be calculated from the table above. The denominator is the marginal distribution of x=12, that is, all the possible prior and likelihood products that would generate x=12. It is a normalizing constant, so that the values all add up to 1, across all p_i values.

proportion, $\boldsymbol{p_i}$	prior probability, $\boldsymbol{\pi_{Non\text{-}Uniform}(p_i)}$	Likelihood, $\boldsymbol{L(p_i)}$	Product $\boldsymbol{\pi(p_i)\ X\ L(p_i)}$	Posterior probability
0.1	0.09	5.4e-08	4.9e-09	7.0e-08
0.2	0.09	8.6e-05	7.8e-06	1.1e-04
0.3	0.09	3.8e-03	3.5e-04	5.0e-03
0.4	0.09	3.5e-02	3.2e-03	4.6e-02
0.5	0.09	1.2e-01	1.0e-02	1.5e-01
0.6	0.18	1.8e-01	3.2e-02	4.6e-01
0.7	0.18	1.1e-01	2.0e-02	2.9e-01
0.8	0.09	2.2e-02	2.0e-03	2.8e-02
0.9	0.09	3.5e-04	3.2e-05	4.6e-04

Table 2.2 Bayesian table This table outputs the posterior probability for different parameter values, p. Here number of successes, x, are 12 and number of trials, n, are 20. These calculations can be easily done using the Bayes' formula.

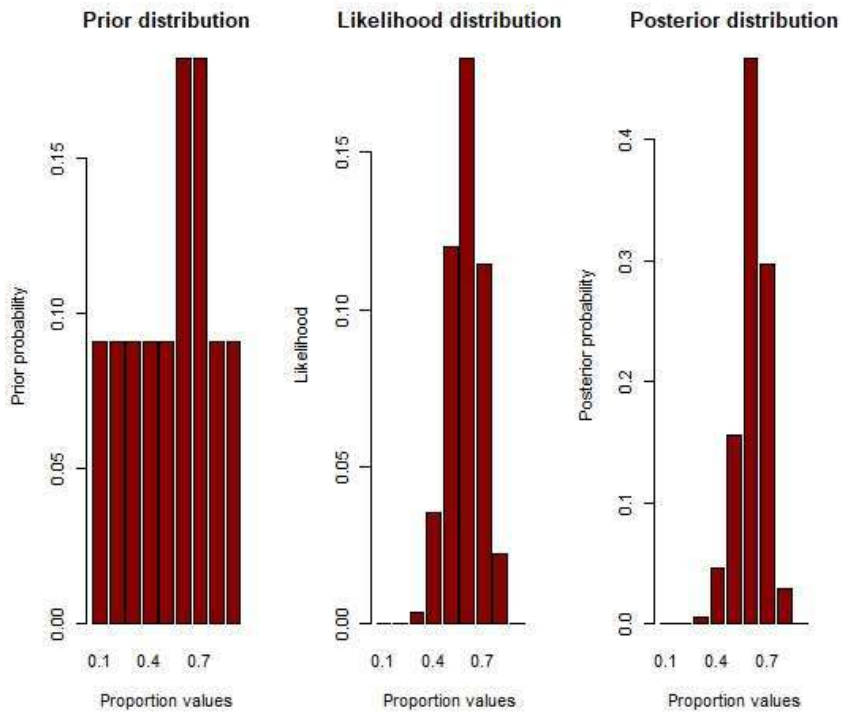


Figure 2.1 Distribution of Prior, Likelihood and Posterior values for an experiment of rolling a dice for n=20 and getting a six, a success for x=12.

In the example taken here, we are dealing with a discrete prior probability distribution. One could also have parameter values that can take continuous values, like in our example *p=0.55*, in which case the model requires continuous prior probability distributions. The Bayes formula then takes the following form, replacing summation with integral,

$$\boldsymbol{\pi(p_i|x)} = \frac{\pi(p_i)\ x\ L(p_i)}{\int_j \quad \pi(p_j)\ x\ L(p_j)}$$

One such family of continuous distributions that can model such prior probabilities, where the proportion can be any value between 0 and 1, is beta distribution. In that case, the prior (beta distribution) times likelihood (binomial model) would be proportional to a posterior distribution that would take the form of beta distribution as well. Because the prior and posterior belong to the same family of distribution, the prior in this case is called a *conjugate prior*. Therefore, here an analytical solution is possible with a solvable Bayes formula as mentioned above.

In many other problem statements, however, including the analysis related to phylogeny, the integral in the denominator does not have an analytical solution. This becomes even more intractable in molecular phylogenetics, where there are multivariate parameters, unlike the case presented here with just one parameter of interest. The only other solution that has been worked out in recent times is to use numerical methods; this bypasses the need to calculate the integral in the denominator. One of the widely used numerical methods, and the one that has been used in the algorithms that were applied for the analysis presented in this thesis, is called the ***Markov Chain Monte Carlo*** or ***MCMC***. MCMC is a general class of algorithms that can be used to simulate posterior distributions from general Bayesian models i.e even if there is no analytical solution available. Obviously since these are only approximations to the underlying real yet intractable posterior distributions, there are diagnostics that need to be run to check the output of any Bayesian analysis, as will be discussed later.

2.0.3 Markov chain

MCMC is based on a general probability model called the Markov chain. Let us take an example, similar to the ones above, where the possible outcomes are finite.

Let us say that there is a hypothetical organism that can undergo reverse metamorphosis much like batflies that can revert back and forth between a winged form and a wingless maggot form. Only that our hypothetical organism can revert back and forth via five states, A to E. If this organism is in the first (A) or the last state (E), it can either stay in this state or move to the right or left respectively. If it is in any of the other middle states, it can either move to the left or the right or stay in the same state. This movement among states can occur with an associated probability. The transition to a given state only depends on the current state and not on the previous state. This is exactly the kind of problem that can be modeled using a Markov chain. And these probabilities are called *transition probabilities* that represent all the likelihood values of moving among the states in a single step. This can be summarized in the form a transition matrix T,

T =

0.5	0.5	0	0	0
0.25	0.5	0.25	0	0
0	0.25	0.5	0.25	0
0	0	0.25	0.5	0.25
0	0	0	0.5	0.5

There are certain properties worth mentioning here with respect to this Markov chain that form the basis for phylogenetic analyses as well.

1) **Irreducibility:** Given the Markov chain above, the organism can transition from any given state to any other given state in one or more steps.

2) **Periodicity:** The organism present in a given state can only come back to the given state in regular intervals. However, over here because the organism can undergo reverse metamorphosis in any direction, the Markov chain in our example is aperiodic.

3) **Unique stationary distribution:** If a Markov chain is irreducible and aperiodic, as is the case here, it will have a unique stationary distribution *i.e as one takes infinite number of steps, the probability of transitioning to a given state does not depend on the initial state*. This is a very useful property that is exploited when carrying out phylogenetic analysis as well. Let us illustrate this with an example:

Let us say that the organism, to begin with, is in state B. Therefore, the probability is 1 and other probabilities are 0.

S_1 = (0, 1, 0, 0, 0)

Now let us say that the organism transitions to another state. After one transition step, the state to which it could have transitioned can be obtained by Multiplying S with the transition matrix T. After three transitions, the state that it could be present in can be obtained as follows,

$S_{1+3} = S_1 X T^3$

One can calculate using matrix multiplication the chances of being in a state given that the starting state was B. (Implemented using an R script)

S_{1+1} = (0.25, 0.5, 0.25, 0, 0)

S_{1+3} = (0.21, 0.38, 0.25, 0.11, 0.03)

As the matrix exponentiation tends to infinity, i.e. the number of states the organism transitions to, the chain reaches a stationary distribution. The matrix then looks like below,

T^{100} =

0.125	0.25	0.25	0.25	0.125
0.125	0.25	0.25	0.25	0.125
0.125	0.25	0.25	0.25	0.125
0.125	0.25	0.25	0.25	0.125
0.125	0.25	0.25	0.25	0.125

The chain is said to reach an equilibrium. It does not matter what the initial state was (we know it was state B here), the probability of transitioning to a given state is independent of it. Any further multiplication of T with the row vector (r) of the T^{100} matrix would yield the same row vector again.

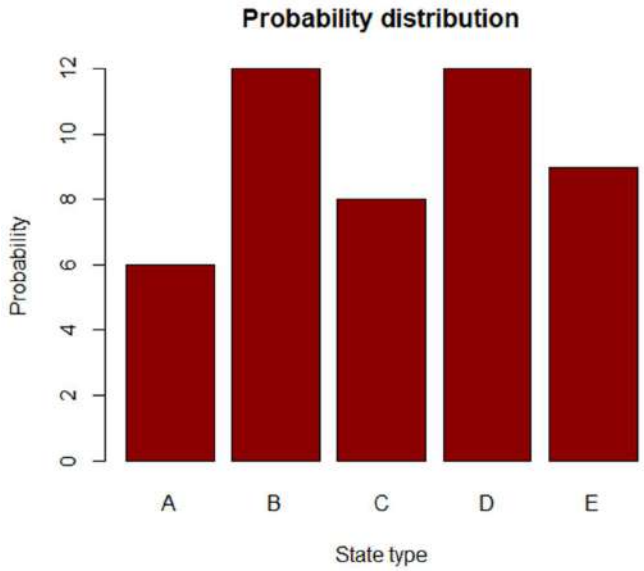
r X T = r

2.0.4 The Metropolis algorithm

In the previous section we discussed what Markov chains are and some of their important properties that are worth knowing. We can use Markov chains to sample from any arbitrary probability distribution. Remember, this is particularly important to be able to approximate a probability distribution of parameter(s) of interest that cannot be obtained analytically.

Let us continue with the example in the previous section. Let us say the organism can be in any of the five states with probabilities proportional to

6,12,8,12,9. Note that these numbers do not add to one. That is the beauty of this algorithm, we do not need to know the probabilities to approximate from this distribution. Recall that this is equivalent to just computing prior times likelihood and not dividing by the normalizing constant to know the posterior probability, as seen in the Bayes' formula.



Metropolis algorithm allows one to simulate from the above probability distribution using a simple random walk:

1) We start from any state the organism can be in, A to E.

2) To decide which state the organism would take next, a fair coin is flipped. Let us say, if the coin flip lands up heads, we move one state to the left. If it lands up tails, we move one state to the right. This is called the *candidate state*.

3) Next we calculate the ratio of the probabilities of the candidate state and the current state.

   R = probability (candidate state) / probability (current state)

4) Next we choose a number *N* between 0 and 1 at random. If *N* is smaller than *R*, we move to the candidate state, otherwise we remain at the current state.

Steps 1 to 4 define an irreducible, aperiodic Markov chain. Step 1 allows one to start from a random position in space. The rest of the steps define a transition matrix T. If this process is repeated a large number of times (millions of iterations in practice), the resulting distribution of our visits should approximate the actual distribution that we defined above, thanks to the stationary distribution property of an irreducible aperiodic Markov chain. Let us say we start at position 2 and simulate for 10000 iterations, we get the following approximation, which is very similar to the actual distribution, (implemented using an R script)

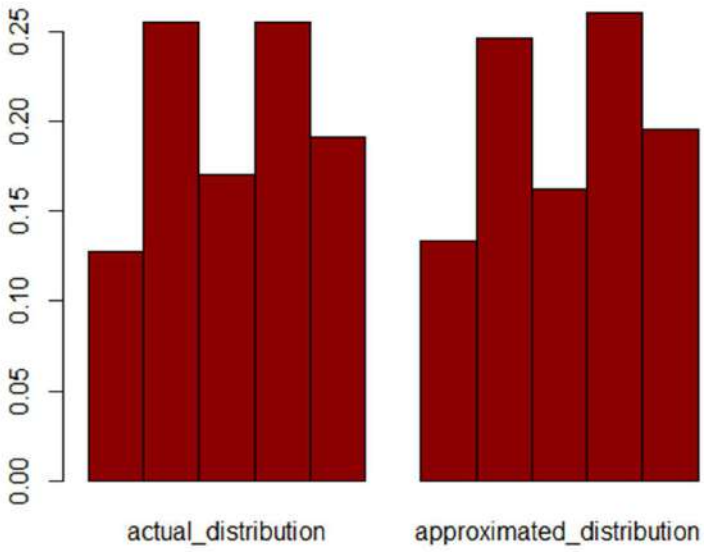


Figure 2.2 Actual and MCMC approximated probability distributions of finding an organism in any of the metamorphosis states A through E (left to right)

In this example, we considered a discrete probability distribution. The same algorithm can be extended for continuous distributions as well. Furthermore, there are multiple flavors to this algorithm, like 1) Metropolis-Hastings (MH), where the proposal distribution can even be non-symmetrical. The Metropolis algorithm presented above forms a special case of the MH algorithm, where the proposal distribution is constrained to be symmetrical. 2) Gibbs Sampling, which can deal with more than one unknown parameter, for example when dealing with parameter distributions in phylogenetic analysis. (Note that we have dealt with distributions until now that have only one unknown parameter.) The mathematical proofs for these are beyond the scope of this thesis. They are all, however, just variants of the Metropolis algorithm presented here.

2.0.5 Simulation of phylogenetic data

Evolution takes place across multiple scales spread both temporally and spatially. This makes it appropriate for simulations, helping one to make theoretical predictions without any complex mathematical tools that are conventionally used to solve problems analytically. Any prediction under a probabilistic model of evolution can be attributed to the nature of species frequency in a given trait state as being modeled as a random variable, allowing for simulations to paint what its distribution could look like. For discrete traits, the basic model that is used is a Markov chain.

The simulations can be either deterministic or stochastic. Deterministic simulations use a related set of equations to simulate evolution, like when one wants to model a population of infinite size under selection. Stochastic simulations involve random numbers, and form the majority of cases seen in phylogenetic comparative methods. We will discuss these with specific use cases in the sections to come.

Another way to categorize simulations is in either continuous or discrete time. A discrete time model, simpler than the continuous time model, can be expressed as follows:

$X_{t+1} = f(x_t)$, where x is a simulated trait and f is some function

A continuous time model is harder to derive and expressed as a differential equation that defines how a simulated trait could change over an instantaneous time interval dt,

$dx/dt = f(x_t)$

The above equation can be solved in two ways: a) To find an analytical solution with a form,

$x_{t+1} = f(x_t, t+1)$

and b) to find a solution to the differential equation numerically, mostly because either the analytical solution is too complex, time consuming or the analytical solution is not available.

For discrete traits, as mentioned before, Markov chains form an appropriate general probabilistic model. The simplest model would be the one with two discrete trait states, X and Y and a rate parameter where the transition from X to Y (forward rate) and Y to X (backward rate) occur at the same rate. In continuous time, the parameter is the rate subject to the constraint 0 <= r. It quantifies the speed of transition in an instantaneous time interval, where the time interval is small enough to disallow multiple changes. A rate matrix, R, can be built where the off-diagonal elements are filled with the rate parameters (in this example both will be the same, forward rate = backward rate). The diagonal elements are filled such that the row sums are set to zero i.e rate of being in the same state and the negative sign depicts the resistance in changing to another state. From this rate matrix, the probability of change between the two states during some time interval t (usually branch length over an edge of a phylogenetic tree representing evolutionary distance) is calculated using matrix exponentiation of *Rt* i.e exp(Rt). This matrix gives the probability that a character in *ith* state will be in *jth* state after time interval t (or evolutionary distance in terms of branch lengths). The rows of the resulting probability matrix sum to one. In the context of phylogeny, the probability matrices over all branches (with given branch lengths) are then used to calculate the overall likelihood of the resulting model using *Felsenstein's pruning algorithm* (described in the next section). With the advent of methods in phylogeny, continuous time markov models (CTMM) are employed universally and this thesis has also used algorithms that implement CTMM.

2.1 Evolutionary change of DNA sequences

This section reviews statistical methods that are employed to study evolutionary changes in DNA sequences. These changes vary depending on the type of region in the DNA - protein coding, RNA coding, flanking regions, repetitive sequences and insertion sequences. Therefore, it is important to know the kind

of DNA sequence that is under investigation. For example, if we consider the protein coding regions, the nucleotide substitution patterns are different at the first, second and third codon positions. Some regions are subject to natural selection less often than other regions, contributing to a variation of evolutionary pattern. This chapter focuses on protein-coding and RNA-coding regions.

2.1.1 Estimation of overall sequence divergence

When two sequences of DNA emerge from a shared ancestral DNA sequence, they diverge gradually by nucleotide substitutions. This extent of divergence can be quantified by simply calculating the proportion of sites at which the descendants are different. This is termed as the p-distance.

$p = n_d/n$, where n_d is the number of sites different between the two descendant sequences and n is the total number of sites.

p-distance gives a metric of the overall nucleotide differences. This can further be broken down into frequencies of nucleotide substitution pairs. For four nucleotides A, T, G, C, there are sixteen nucleotide substitution pair frequencies:

1) Identical (I): AA , TT, CC and GG
2) Transitions (Ti): CT, TC, AG, GA
3) Transversions (Tv): AT, TA, CG, GA, AC, CA, GT, TG

For any two descendent sequences, if the nucleotide substitutions occur at random, Tv is expected to be greater than Ti. However, as has been observed, Ti occur more frequently than Tv. When the sequence divergence is low, the transition/transversion ratio (R) can be estimated by,

$\hat{R} = \widehat{Ti}/\widehat{Tv}$,

where $\widehat{Ti}$ and $\widehat{Tv}$ are the observed values of Ti and Tv, respectively. As follows, if the number of nucleotides that are examined are small, $\hat{R}$ is subject to a large sampling error.

p-distance can be expressed in terms of Ti and Tv as follows,

$$p = \widehat{Ti} + \widehat{Tv}$$

p-distance is a reliable measure of estimating the number of nucleotide substitutions per site when the sequences are closely related. However, when the p-distance is large, owing to highly diverged sequences, the measure gives an underestimate of the distance as it does not account for backward and parallel substitutions. Probabilistic models, discussed in the next section, provide a way of incorporating the latter and provide a better estimate of distances.

2.1.2 Probabilistic models of molecular evolution

Given a sequence alignment, modeling evolution using a probabilistic model is most commonly carried out by treating the evolution of each character column as a continuous-time Markov process. Here the assumption is that when a (ancestral) sequence duplicates, the character(s) in the descendent sequence can be randomly selected from a distribution that only depends on the current state of that character. Mostly the character remains unchanged, but changes to other character states are possible and the probability of change depends on dynamics of biochemical processes. Inherently, these processes are thought to be stochastic in nature and that there is no known mechanism by which the character states in the ancestral state could affect the outcome in the next generations. Rather than assuming discrete generations, time here is modeled as a continuous variable owing to the slow nature of evolutionary changes, spread across a very large number of generations.

The model is captured using a transition rate matrix, R, which describes different substitution rates, and the initial state of character(s). For a DNA sequence, the transition matrix would have dimension of 4X4, owing to the four nucleotide characters. Many different classes of these transition matrices have been worked out in the past decades; each matrix makes slightly different assumptions of the probabilistic model of molecular evolution. We are going to discuss a few of these models with increasing complexity in the next section.

At this point, it is worth mentioning the assumptions these models make: a) in a given alignment, mutations are identically and independently distributed, b) lineages arise strictly through a tree-like evolution, c) reversibility: mutations can revert to a previous state, d) stationarity: mutational processes are consistent through time i.e non-directed nature of evolutionary processes like neutral evolution and e) Markov process: mutation events are not influenced by previous mutations at the site under study.

Last thing to note with respect to these probabilistic models is that they cannot be used to estimate simultaneously both the overall substitution rate and the amount of time the Markov process has been evolving. These models instead only allow us to estimate their product, that is the evolutionary distance i.e the mean number of substitution events expected to occur per site.

2.1.3 Estimation of nucleotide substitutions

To overcome the limitations of p-distance, especially for distantly related sequences, different nucleotide substitution models have been proposed in the field. They aid in *distance correction*. This section reviews some of the most common models employed in evolutionary studies.

2.1.3.a Jukes and Cantor's Model

This model was proposed by Jukes and Cantor in 1969 (41) and is one of the simplest models for nucleotide substitutions in the field. This model assumes:

1) Substitutions at any given site occur with equal frequency,
2) At each site, a nucleotide changes to one of the other three nucleotides with a probability of $\propto$ per unit time.

	A	T	G	C
A	-	$\propto$	$\propto$	$\propto$
T	$\propto$	-	$\propto$	$\propto$
G	$\propto$	$\propto$	-	$\propto$

C	$\propto$	$\propto$	$\propto$	-

For example, at a given site, nucleotide A can change to either T or G or C, with a probability of change,

$r = 3 \propto$, where r is the rate of nucleotide substitution per site per unit time.

Let there be two sequences, A and B that diverged from a common ancestor X, t years ago.

Let q_t be the proportion of identical nucleotides between A and B at time *t*.

Let p_t be the proportion of non-identical nucleotides between A and B at time *t*. ($p_t = 1 - q_t$)

Proportion of identical nucleotides, q_{t+1}, at time *t+1* can then be calculated in the following way:

1) For a given site with same nucleotide in A and B sequences at time *t*, the probability of it remaining the same at time *t+1* is (1-r) *(1-r). Since r is a small quantity, r^2 can be neglected. Therefore, the probability can be approximated to 1-2r.

2) Another possibility is that at time *t* the two sequences for that given site had different nucleotides, and at time *t+1* have the same nucleotide. This can happen in two ways. For different nucleotides *i* in A and *j* in B at time *t*,

    a) *i* in A changes to *j* and *j* in B remains the same, or

    b) *j* in B changes to *i* and *i* in A remains the same

    Taking event *a* above, the probability of its occurrence is $\propto$(1-r) or r*(1-r)/3. Similarly, the probability of occurrence of event b is also r*(1-r)/3. Therefore, the total probability is 2r(1-r)/3 or 2r/3, if we neglect the r^2 term.

3) Taking 1 and 2 together,

$$q_{t+1} = (1-2r)\, qt + 2/3r(1-q_t),$$

$q_{t+1} - q_t = 2r/3 - (8r/3)\, qt$

If we consider a continuous time model, the above equation can be represented by the following differential equation,

$$\frac{dq}{dt} = \frac{2r}{3} - \frac{8r}{3}q$$

Solution of the equation above with an initial condition q=1 at t=0, i.e when the ancestral sequence has not diverged, is given by

$$q = 1 - \frac{3}{4}(1-e^{-8rt/3})$$

Under the Jukes and Cantor model, the expected number of nucleotide substitutions per site, denoted as *d*, for two sequences is r*t +r*t = 2rt. Therefore, rearranging the above equation, gives us

$$d = -(\frac{3}{4})\ln[1-(\frac{4}{3})p]$$

where p = 1 - q i.e the proportion of nucleotide differences between A and B.

An estimate of d, $\hat{d}$, can be obtained by replacing p with $\hat{p}$ i.e the observed proportion of nucleotide differences between A and B.

$$\hat{d} = -(\frac{3}{4})\ln[1-(\frac{4}{3})\hat{p}]$$

with a sample variance of $\hat{d}$,

$V(\hat{d}) = \frac{9p(1-p)}{(3-4p)^2 n}$, where n is the number of nucleotides included for analysis.

2.1.3.b Kimura's two parameter model

Changes between nucleotides that are chemically similar (transitions) are more likely than those with different chemical structure (transversions), as it is energetically less disruptive. Moreover, the genetic code allows for more transitions over transversions without the replacement of an amino acid. Kimura's two parameter model allows one to account for this feature (42).

The rate of substitutions per site per unit time owing to transitions (∝) are assumed to be different from those owing to transversions (2β). The total substitution rate per site per unit time is given by, r=∝+2β.

	A	T	G	C
A	-	β	β	∝
T	β	-	∝	β
G	β	∝	-	β
C	∝	β	β	-

Kimura showed that the transition and transversion frequencies, Ti and Tv, respectively, can be obtained as follows,

$$Ti = (\frac{1}{4})(1 - 2e^{-4(\beta+\propto)t} + e^{-8\beta t})$$

$$Tv = (\frac{1}{2})(1 - e^{-8\beta t}),$$

Therefore, the expected number of nucleotide substitutions per site between two sequences A and B can be derived as follows,

$$d \sim 2rt = 2\propto t+4\beta t = -(\frac{1}{2})\ln(1- 2Ti - Tv) - (\frac{1}{4})\ln(1-2Tv)$$

2.1.2.c Other nucleotide substitution models and extensions

The stationary distributions of the two models discussed so far are assumed to be uniform distributions for all nucleotide frequencies. Other models, more complex, exist where the stationary distributions can be allowed to be non-uniform. It is important to model the fact that base frequencies vary significantly both within a genome and among species. One such model is Hasegawa, Kishino, and Yano (HKY) (43). Another modification of this model is Tamura-Nei 1993 (44), which in addition to allowing different base frequencies as HKY, models A<->G, C<->T and transversion substitutions separately. Finally, the

General Time reversible (GTR) (45) model is the most flexible model of nucleotide substitution that preserves the time reversibility property of Markov chains. It also allows for all types of character substitutions to occur at distinct rates and for arbitrary equilibrium frequencies. Not all models have analytical solutions like Jukes-Cantor and Kimura's two parameter model. The transition matrices for these models then can be calculated using numerical methods via simulations similar to the ones discussed previously.

There are few other characteristics, learnt based on empirically dealing with nucleotide sequences, that are not included in the models discussed so far. These characteristics can be modeled additionally. First, allowing certain sites to be constrained as *invariants*, where any substitution is not allowed. This is based on the observation that certain positions are more important to sequence function than others and these sites experience stronger purifying selection. Second, incorporation of *rate categories*. This allows for the observation that different sites in a sequence might evolve at different rates. The rate categories are modeled using a gamma distribution.

2.2 Molecular phylogenetics

While there are many methods that have been employed to construct phylogenies, right from distance based methods like UPGMA, Neighbor joining to methods like Maximum Parsimony, we will be discussing in length how phylogenies are made using the Maximum Likelihood (ML) method. ML uses a continuous time Markov process as discussed in previous sections. Maximum likelihood has several advantages over other methods mentioned in the previous line. It allows for a maximum use of information, including the information of branch length. It allows for the use of different mutational models of how the sequences and therefore species could have evolved. It not only tells us which tree could be preferred but also quantitatively tells us, by how much should that tree be preferred.

Let us see how a phylogeny is constructed using a maximum likelihood approach, an algorithm called *Felsenstein's pruning algorithm* (46,47). We start with a multiple sequence alignment of homologous sequences. Typically, 16S

rRNA sequences are chosen as they evolve slowly and are a part of a housekeeping machinery in bacteria, thereby allowing for hypothesizing the relationship among organisms, giving a decent resolution at least at the genus level and in certain clades at the species level.

The probabilistic model that is tested, in general, is how an ancestral sequence could have evolved into the sequences present in the multiple sequence alignment (our given data). The model consists of several parameters. For the simplest case, these parameters include tree topology, branch lengths, nucleotide frequencies, nucleotide-nucleotide substitution rates (or probabilities, that are calculated using matrix exponentiation as explained in the section on Statistical primers using a simple 2 X 2 case).

Let's take an example to understand how maximum likelihood is used:

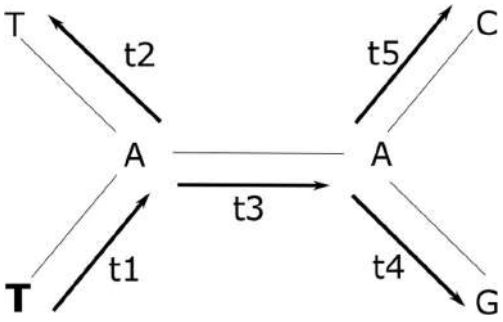
Given a multiple sequence alignment, let's focus on column two (in bold)

Bacteria 1: A**T**GCATTTT

Bacteria 2: A**T**G-ATTTT

Bacteria 3: A**C**GGATTTT

Bacteria 4: A**G**GCATTTT

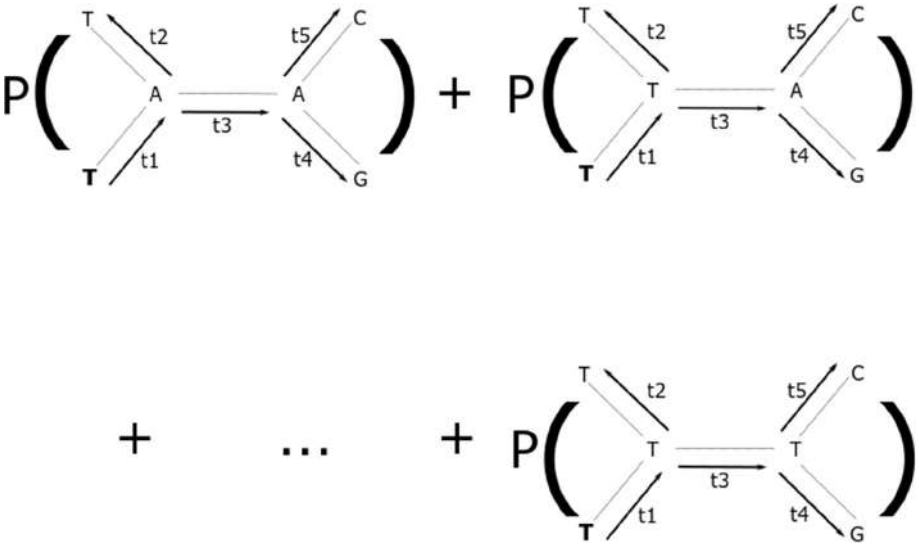


Assuming that we randomly select certain parameter values for branch lengths, nucleotide frequencies, substitution rates (and probabilities that can be calculated using substitution rate matrix exponentiation as discussed in previous section), tree topology and we start at some node. In this example, we start at the bottom left T (in bold), the likelihood for that column is given by the

following (here we assume that the ancestral nucleotides at that column are A and A),

Probability = f_T x $P_{TA}(t1)$ x $P_{AT}(t2)$ x $P_{AA}(t3)$ x $P_{AG}(t4)$ x $P_{AC}(t5)$

We repeat for the same parameter values the calculation of likelihood for this column, but now for all possible combinations of ancestral nodes and not just A and A. Since these are OR cases, these probabilities for a given column would be summed over each other. Here we have two internal nodes, therefore we will add up 16 possible combination terms as above. Likelihood for column 2 here will then be given by, **L_2 =**



To get the likelihood of the entire model i.e over the entire sequence alignment, we multiply the individual probabilities of each column. These likelihoods are usually reported as log-likelihoods to deal with very small values of probabilities, which often lead to computer underflow technical problems. For a multiple sequence alignment of length n, the overall likelihood is given by,

$L = L_1 \times L_2 \times L_3 \times \ldots \times L_n$

$\ln(L) = \ln(L_1) + \ln(L_2) + \ldots + \ln(L_n)$

Now that we understand how the likelihood of a model is calculated given an initial set of parameter values, the above is repeated by randomly choosing different values of all parameter values. For each parameter value combination,

the likelihood is calculated and a step in a direction that increases the likelihood is taken. By the end of this exercise, we chose the phylogeny that maximizes the likelihood for all parameter values - tree topology, branch lengths, substitution rates (and probabilities), base frequencies etc.

Apart from the above estimates, we also get a measure of how well the data fits the model by the end of this exercise i.e the overall log-likelihood of the model. This can be used to compare different models (Jukes-Cantor versus Kimura versus HKY versus GTR) and select the one that best explains the data at hand. Two main methods are employed for model selection: a) Likelihood ratio test and b) AIC or Akaike Information Criterion based on information theory. These will be discussed in the next section.

2.3 Model selection

2.3.1 Likelihood ratio test

A direct way of comparing two given models is by taking a ratio of their associated likelihood values, as calculated in the previous section.

Likelihood Ratio (LR) = Likelihood of Model 1 / Likelihood of Model 2

If model 1 has a higher likelihood, LR > 1.

Mathematically it can be shown for nested models i.e. models where one model (Model 1, let us say) is a special case of a more general model (Model 2), that

$\ln(LR^2) = 2 \ln(LR) = 2 \times (\ln L \text{ of Model 1} - \ln L \text{ of Model 2})$, follows a chi-squared distribution with degrees of freedom that are extra in the more complicated model.

This allows us to test if the complicated model is significantly better than the simpler model.

2.3.2 Akaike Information Criterion

Recall that a probabilistic model of a system defines the probability distribution over possible outcomes. For example, probability of getting a specific alignment

for a GTR model on a given phylogeny. In information theory, assuming that the underlying true model is known, one can compare how much the approximated model (in this case GTR model) has diverged from the true model, using Kullback-Leibler divergence. If the true probability distribution is P, and there are multiple Q approximated probability distributions (JC, Kimura, HYK, GTR etc.), we can use KL divergence to find the model, out of the given models, that best approximates the true distribution. Therefore, KL divergence is the distance measure between two distributions and is given by,

Distance(P||Q) = $\sum_{i=1}^{N}$ $\mathbf{p_i \log(\frac{p_i}{q_i})}$

However, in our case, we do not know the true probability distribution. AIC offers a solution. AIC or Akaike Information Criterion, is an estimate of the expected, relative Kullback-Leibler distance between the true and the approximating model. (The mathematical proof of how AIC is an estimate of KL divergence is out of the scope of the thesis).

AIC = -2ln(L) + 2K

Here, ln(L) is the log-likelihood of a model calculated as discussed previously. K is the number of free parameters that the model is based on; K would be least for JC and higher for other models like HKY, GTR etc. The advantage of this model selection measure is that unlike the Likelihood ratio test, there is no requirement of the models being compared to be nested within one another. The lower the AIC, the better the model. Notice how a model with more parameters will increase the AIC, penalizing the model with more parameters in an attempt to over fit the model. Similarly, a higher likelihood will decrease the AIC, preserving the information that the model conveys.

2.4 Phylogenetic Comparative Methods

In the previous sections, we talked about the ways in which relationships among different species of interest can be constructed. These relationships are summarized in the form of a phylogenetic tree. We can use phylogeny to compare traits across species and test various hypotheses. We map the trait(s) onto a phylogeny and trace evolutionary history to make inferences about when,

where, and how the traits might have evolved. Such methods can be used to study data that is distributed in a discrete manner (e.g. presence/absence of a given trait) or continuously (measurement of a given trait like size). I will only be discussing the phylogenetic comparative methods related to discrete characters, as the work presented in this thesis deals majorly with them. Particularly, the following Phylogenetic Comparative Methods will be discussed:

1) Phylogenetic signal and phyloANOVA
2) Ancestral state reconstruction
3) Correlated Evolution

2.4.1 Phylogenetic signal

Phylogenetic signal is a measure that indicates the effect of shared ancestry in explaining the distribution of a given trait over a phylogeny. The idea is that descendants from a common ancestor are more likely to resemble each other (and the ancestor) as compared to any two organisms picked at random from anywhere over the entire phylogeny. Testing for phylogenetic signals is important in comparative phylogenetic studies since all the conventional statistical tests are based on the assumption that the data (points) included in the analysis are identically and independently distributed. However, traits with a high phylogenetic signal would indicate non-independence within the data, owing to shared ancestry. Studies have shown, starting from the seminal work by Joseph Felsenstein in 1980s (48), that inadequacy in controlling for the effects of phylogenetic conservatism could lead to false positives or Type I error in drawing inferences related to the trait under study. Therefore, before carrying out any phylogenetic comparative studies, it has been debated in the past that phylogenetic signals should be calculated to then proceed ahead with appropriate statistical tests. One such method that has been developed to compare the distribution of a trait of interest among two or more groups of organisms is phyloANOVA (49) (also used in the work presented in the thesis). The general workflow in this case would be to check for the phylogenetic signal for our trait. In case the phylogenetic signal is low, proceed ahead with ANOVA (without correcting for phylogeny). In case it is high, use phyloANOVA

(mechanism discussed in the schematic below) to compare the trait distribution among the groups of interest.

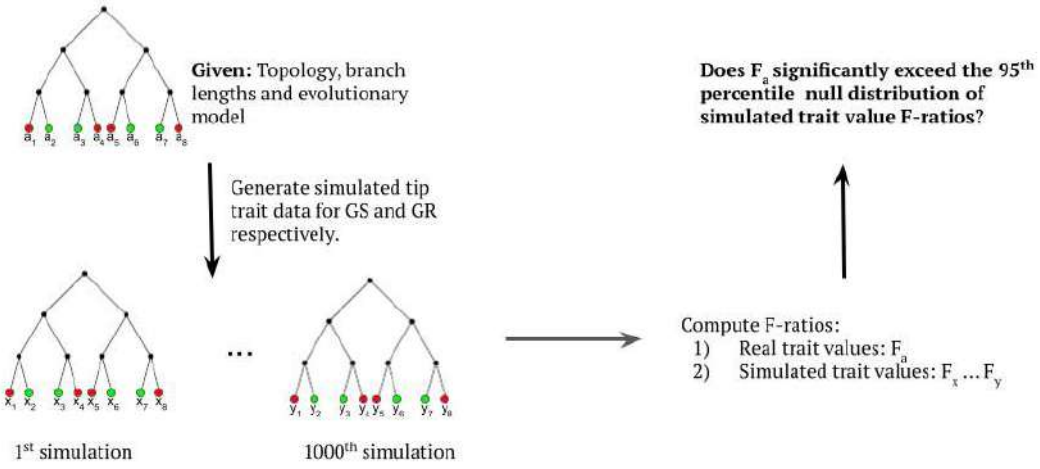


Figure 2.3 Cartoon explanation of phyloANOVA algorithm

Now let us discuss how a phylogenetic signal is calculated. For categorical traits, there are two methods that exist in literature to calculate the phylogenetic signal: a) D statistic by *Fritz and Purvis* (50), and b) delta statistic by *Borges et. al.* (51)

2.4.1a D statistic

The main principle behind the statistic is to calculate the sum of differences between sister-clades over the entire phylogeny. A trait that could either be completely clumped or over dispersed if the trait is in the same state in related species in the former but not the latter. As a result, the sum of these differences will be the lowest if the trait is strongly clumped and highest if the trait is over dispersed. The sum would also be a function of the trait prevalence and the phylogenetic tree shape and size. This is particularly important to consider (and normalize) if one wishes to compare the statistics across different datasets.

In terms of the calculation, the trait values at the internal nodes are calculated much like the independent contrasts method proposed by Felsenstein 1985 (48) - trait value at a given node is the mean of the descendent node values inversely weighted by their evolutionary distance. Then the difference between each pair

of sister clades is summed across the entire tree ($\sum d_{observed}$). This sum of difference is then scaled by the expected sum if the trait was distributed randomly over the phylogeny ($\sum d_{random}$ overdispersed) and the expected sum of difference under a given evolutionary model.

To scale under a given evolutionary model, continuous traits evolving under a Brownian motion i.e. a random walk with a temporally constant variance) are simulated. A threshold is chosen to convert these values into a binary trait. The threshold is set such that the resulting binary trait has the same prevalence as seen in the observed data. Species with continuous trait values below the threshold are given a score of 0 and those above are given a score of 1. The sum of difference thus generated is termed $\sum d_{brownian}$.

Using the above D statistic is calculated and scaled as follows:

$D = [d_{observed} - mean(d_{brownian})] / [mean(d_{random}) - mean(d_{brownian})]$

D = 1, if the observed trait of interest has a phylogenetically random distribution

D = 0, if the observed trait of interest is clumped and as if it has evolved under a Brownian model for some underlying continuous trait

Statistical significance for D using the permutation test can be calculated by observing where the observed sum of difference lies within the two expected distributions - a) sum of differences under the random model and, b) sum of differences under the threshold model where the underlying continuous trait is evolving under a brownian motion.

2.4.1b Delta statistic

This statistic exploits the entropy measurement contained in ancestral inferences. First an ancestral reconstruction is carried out using state-of-the-art methods for our categorical trait of interest. The output returns the probability of all k states that the ancestor could take at that node. The expectation is that if a phylogeny is better associated with our trait under study, one can retrace its evolution with minimal uncertainty. This uncertainty is captured using Shannon's entropy from Information theory.

The quantity of information (e_i^j) coded in the node probabilities, for a trait with k-states and a given state i at a node j, can be calculated using a linear version of Shannon entropy with the help of state probability p^j_i. If the state probability is 1 or 0, the entropy is 0 since there is no uncertainty for that state. As the probability departs away towards 1/k, entropy increases i.e. uncertainty increases. The node entropy (e^j) can be obtained by summing up all entropies for all k states for that node. e^j lies between 0 and 1 and corresponds to situations of absolute certainty and uncertainty.

Finally, a Bayesian inferential scheme is implemented to calculate the delta statistic, *δ*. Since e^i lies between 0 and 1, its likelihood follows a beta distribution with parameters α and β. To calculate the posterior distribution of these two parameters, the prior probability is taken as an exponential distribution with the lambda rate parameter. To sample from their posterior distributions an MCMC-Metropolis Hastings within a Gibbs sampler algorithm is used. Delta is defined as the expected ratio between posterior distributions of β and α, as follows:

$$\delta = E [p(\beta|\alpha, e) / p(\alpha |\beta, e)]$$

δ is greater than 1 when β > α. It is possible when the entropy distribution favors lower entropies over higher. Hence, the higher the δ-values, the greater the information the ancestral inferences provide i.e. higher phylogenetic signal.

2.4.2 Ancestral state reconstruction of discrete characters

The work presented in this thesis has used a specific method for calling ancestral states over the entire phylogeny called Stochastic character mapping (SCM) (52). What sets it apart from other ancestral state reconstruction methods is its ability to simulate state changes not just at nodes but also along the edges of the tree. In this method, given a tree and our data (the discrete trait that one wishes to map), the possible discrete character trait histories are randomly sampled in direct proportion to its posterior probability under a given evolutionary model. For discrete traits, a commonly used model is a continuous-time discrete-state Markov chain (as described in the previous section).

A joint reconstruction of the trait of interest is sampled across all tree nodes, sampled on an instantaneous rate transition matrix, R, between states and our discrete character data. The states are sampled from the joint posterior probability distributions, similar to methods discussed in previous sections. A rejection procedure is used to sample changes along the tree edges. Given an initial state i, an exponential distribution with rate - R_{ii} is used to randomly sample the waiting times for changes between the states. At a given branch with a length *l*, if the sampled waiting time is less than *l*, we simulate another change, check again, if it is still less than *l*, we simulate another; we do this till we reach the branch end. A new state is simulated at each change by randomly selecting a state with a probability $P(j) = R_{ij} / \sum_{j=1}^{n;\, j \neq i} R_{ij}$, for any derived state j. After this exercise, if the states at the start and end of the branch match the initially stochastic joint sampled node states, the stochastic history for that branch is retained. Otherwise we reject this simulation and repeat the procedure until the states at the beginning and the end match our stochastic joint node states.

2.4.3 Correlated evolution of discrete characters

2.4.3a Maximum likelihood

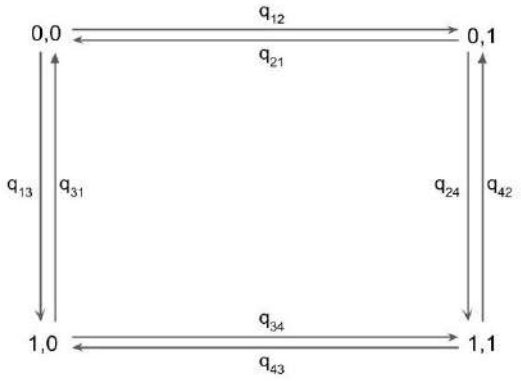
In the first step towards studying coevolution (defined here as a phenomenon where two entities have a causal relationship that dictates how one entity evolves in the absence or presence of another entity), one studies correlated evolution between the two traits under study. To study if the signal of correlation is statistically significant, there are a number of measures and tests in the literature. However, all these methods are based on the assumption that the data points included in any such analyses are identically or independently distributed (i.i.d). However, these assumptions are true for most real world problems, this might not be true for biological entities like species. In terms of character traits (here we only refer to discrete characters like presence/absence of a gene) they do not evolve independently. In other words, two descendent species can have similar traits to that of their ancestor by the virtue of a shared ancestor, unless acted upon by different selection pressures. Not accounting

for the non-independence in the data leads to false positive results. This was first pointed out by Joseph Felsenstein in his seminal paper (48). Therefore, special methods that can account for the non-independence data structure to identify if there is truly a correlated evolution between the two traits, are required.

Mark Pagel in 1994 (53) proposed a method based on maximum likelihood that is most widely used to study correlated evolution among discrete characters. It draws from all the methods that we have already discussed in the previous sections and is very similar to the method used to construct a phylogenetic tree where characters A, T, G and C are evolved under a continuous-time discrete state Markov process. In Pagel's method for a binary discrete character evolution, the transition matrix would replace A, T, G and C with the traits of interest under two models:

a) independent model (two 2X2 matrices, one for each trait) and,

b) dependent model (one 4X4 matrix).



	0,0	0,1	1,0	1,1
0,0	…	q_{12}	q_{13}	0
0,1	q_{21}	…	0	q_{24}
1,0	q_{31}	0	…	q_{34}
1,1	0	q_{42}	q_{43}	…

The figure above, part a, is a state transition diagram that defines the rates of transition among four state combinations. The values 1 to 4 in the rates correspond to the following pairs of states: {0,0}, {0,1}, {1,0} and {1,1}.

Under an independent evolution of both traits, the rate of transition between two states of trait 1 will be independent of the background state of trait 2. Therefore, under this model: $q_{12} = q_{34}$; $q_{13} = q_{24}$; $q_{21} = q_{43}$ and $q_{31} = q_{42}$. As follows, the independent model would use a maximum of four parameters and a minimum of one parameter (if all four pairs take the same value).

Under a dependent model, there are no restrictions on the rate parameters, thereby allowing certain transitions in one trait to depend on the background state of the other trait under analysis. What this would suggest is that pairs of states for the two traits would tend to be correlated more than what one would expect by chance. One can imagine the number of possible model combinations under the dependent model constrained on the eight rate parameters.

Both the models can be summarized in the form of a rate matrix $Q_{I,D}$. The elements in the diagonal of the matrix are defined as the negative sum of other rate elements belonging to the row matrix. This is done to ensure the sum of the row elements is zero. Please note that the rate of dual transitions is set to zero. This is justified because in an infinitesimal amount of time, the chances of both changes happening in both the traits being tested for correlated evolution would be negligible and therefore could be ignored.

From the above rate matrix, for a given time interval (in our case evolutionary distance measured as branch length between two nodes), probability of transition can be calculated as follows:

$P_{ij}\ (dt) = q_{ij}\ X\ dt$

$P(t) = e^{-Qt}$

These probabilities are calculated over all tree branches and all node states that are possible. This gives the likelihood of the data given the model of trait evolution (dependent or independent). The likelihood of the observed data (trait values of extant species included in the analysis) given the two models are

calculated separately using the same techniques as those used when constructing a phylogenetic tree (already discussed in previous sections) based on Felsenstein's pruning algorithm (46,47). Finally, using any model selection methods like Likelihood-ratio test or Akaike Information Criteria (discussed in the previous section), the better-fit model is chosen.

2.4.3b Bayesian

Using methods based on Maximum-likelihood, we find point estimates of parameters that give the best fit to our data. There could however be a range of parameters, with lower likelihoods that could provide a decent description of our data. Bayesian methods can be used to generate a posterior probability distribution of our parameter estimates. This method allows one to incorporate uncertainty associated with parameter estimation. One starts with a prior probability distribution, performs a likelihood analysis based on the given phylogenetic tree shape and the trait values of the extant species and updates those beliefs for our parameters in the form of a posterior distribution.

The posterior distribution is calculated using techniques based on MCMC as described in the previous sections. Since we do not know the underlying true posterior distributions, certain diagnostics are usually run to check the validity of our final findings. Two main issues are worth mentioning here:

a) **Chain convergence:** Recall how MCMC methods sample and approximate the distributions. Initially, MCMC starts at a point away from the posterior distribution, a low-likelihood region of the parameter space. This is called the 'hill climbing phase', until it reaches convergence to the posterior distribution space. Therefore, the initial few iterations of simulations are discarded as a pre-convergence phase or the burn-in period.

b) **Autocorrelation:** Recall that MCMC takes advantage of the parameter search space as being correlated in the direct neighborhood. Therefore, successive steps in the chain could give rise to biased results. And one might not be able to sample a true representative of the true posterior distribution. One solution is to run the chain for a long period of time, and

sometimes run multiple chains to avoid getting stuck at the local minima. However, this could also lead to a huge amount of output that takes up a lot of disk space (since 10s and 100s of iterations are run in a practical analysis). Therefore, thinning is one solution to circumvent this problem, where samples are collected at regular steps from each MCMC chain.

It is worth discussing the choice of priors, the subjective part of a Bayesian analysis. The simplest case is choosing a non-informative prior, for example a uniform distribution where each value is equally likely. This expresses no strong prior belief of the parameter value. Then there are informative priors. These can vary from being weak beliefs to strong beliefs, fine-tuned by the parameters of the distributions. One such case of an informative prior is an exponential distribution. This is particularly useful to estimate rates of evolution under a parsimony principle or among closely related species, where the lower values are more likely. Any parameter estimate created should justify the use of the prior that is employed. A common way of setting the parameters of the prior is to run a maximum likelihood analysis first, set the center of the prior distribution around the Maximum-likelihood point estimate. One way of choosing the limits of parameters of a prior distribution is by not hard-coding them but allowing them to change and be set using a set of hyper-priors. This is carried out using a procedure called reversible jump MCMC that helps choose the hyper-priors.

For carrying out a Bayesian correlated evolution analysis for two binary traits, *Pagel and Meade 2006* implemented all the above techniques in their seminal paper (54).

Let us see how Bayesian analysis can be used to model different models for independent and dependent evolution. From the rate matrix, the pairs as described in the maximum likelihood section of correlated evolution, can be written as follows: ($\{q_{12},q_{34}\},\{q_{13},q_{24}\},\{q_{21},q_{43}\},\{q_{31},q_{42}\}$)

Using integers (1 to 8) for each rate element, two elements that are assigned the same integer would fall in the same rate class. Therefore, for an independent model of evolution the above pairs would look like: (1,1,2,2,3,3,4,4) if there are four rate parameters, or (1,1,1,1,1,1,1,1) if there is one rate parameter. Similarly, for a dependent model with all different rates, the

parameters would take integers of the form: (1,2,3,4,5,6,7,8). Remember, for it to be a dependent model, the pairs should at least be different, it just so happens that in this conformation, all rates are different. One can imagine other conformations of dependent models where certain rates would belong to the same rate class.

For n objects and c classes, one can calculate the number of models that can be formed (including independent and dependent conformations), using Stirling numbers:

$$S_2(n,c) = 1/c! \times \sum_{i=0}^{c-1} (-1)^i \binom{c}{i} (c - i)^n$$

In our case, n = 8, corresponding to eight rate parameters, and c would vary from 1 to 8 i.e. rates either belonging to one rate class i.e. c= 1 or all rates belonging to different classes i.e. c=8.

$S_2(8,1) = 1$, $S_2(8,2) = 127$, $S_2(8,3) = 966$, $S_2(8,4) = 1701$, $S_2(8,5) = 1050$, $S_2(8,6) = 266$, $S_2(8,7) = 28$, $S_2(8,8) = 1$, that amounts to a total of 4,140 distinct models. The last model $S_2(8,8)$ is (1,2,3,4,5,6,7,8). One of the models with seven parameters could be (1,2,3,4,5,6,7,7). Using this scheme, fifteen out of 4140 models would take the conformation of an independent model. Therefore, $S_2(4,1) = 1$ i.e. (1,1,1,1,1,1,1,1) and $S_2(4,4) = 1$ i.e (1,1,2,2,3,3,4,4) and so on.

One can therefore imagine how using Markov Chain Monte Carlo (MCMC) one can estimate parameter values with uncertainty for a given independent model of evolution and a given dependent model of evolution. But because there could be multiple model combinations with different rate classes as discussed above to choose from, one needs to explore the space of these parameter values called the hyper-parameters i.e the parameters that control the parameters for a given model combination. Reversible jump MCMC is a widely used technique that helps choose these model combinations. In short, four operations are used in rj-MCMC:

a) Merging: This operation reduces the number of rate classes by combining two distinct rate classes. For a model combination of (1,2,3,4,5,6,7,8), if the last two classes are merged, it would yield (1,2,3,4,5,6,7,7).

b) Split: This operation does the reverse of the merge operation. For a model combination of (1,2,3,4,5,6,7,7) a split operation on the last two rate parameters would yield (1,2,3,4,5,6,7,8).

c) Reduce: Operations merge and split change the model dimensions, they do not remove any of these parameters. Reduce operations can explore a possibility that the rate parameter can be given a class where the rate is set to zero i.e it would remove the given evolutionary path from the model under test.

d) Augment: This operation returns the parameter from the rate class equal to zero.

With the augment and reduce operations, one can then calculate the total number of model combinations to increase from 4140 to 21,146 (since for n=7, the Stirling number would give 877 models and since any of the eight rates can be assigned a zero bin, that gives 8 x 877 additional models i.e. 7016). Including these models, fifty-one of the 21146 models would belong to the independent trait evolution.

Model testing in case of a Bayesian analysis is carried out using a Bayes factor. Likelihood ratio tests and tests based on information theory like Akaike information criterion as discussed in the previous section are useful only if testing single likelihood values for two models (in our case, independent versus dependent models). Bayes factor compares model i to model j by calculating the ratio of the marginal likelihood of model j to model i (i.e. all possible model combinations for independent model, i, tested against all possible model combinations under dependent model, j):

$BF = P(D|M_j) / P(D|M_i)$

Values of BF > 1 suggest that model j is preferred over model i.

2.5 Discussion

This chapter has discussed in detail the algorithms used in the evolutionary and comparative genomics studies, pertaining to the ones that were used in the work presented in this thesis. As discussed so far, a thorough understanding of the caveats and the assumptions associated with these algorithms is important to interpret the results obtained from these analyses. Especially, in cases related to phylogenetic analyses, it is common to misinterpret and wrongly extrapolate the results, which if left uncorrected could propagate through the studies, especially when experimental and computational biologists without a background in these sophisticated and niche algorithms use them.

Chapter 3: Evolutionary and Comparative Analysis of Bacterial Non homologous End Joining Repair

This work is published as

Mohak Sharda, Anjana Badrinarayanan, Aswin Sai Narain Seshasayee, Evolutionary and Comparative Analysis of Bacterial Nonhomologous End Joining Repair, *Genome Biology and Evolution*, Volume 12, Issue 12, December 2020, Pages 2450–2466, https://doi.org/10.1093/gbe/evaa223

3.1 Abstract

DNA double-strand breaks (DSBs) are a threat to genome stability. In all domains of life, DSBs are faithfully fixed via homologous recombination. Recombination requires the presence of an uncut copy of duplex DNA which is used as a template for repair. Alternatively, in the absence of a template, cells utilize error-prone Non-homologous end joining (NHEJ). Although ubiquitously found in eukaryotes, NHEJ is not universally present in bacteria. It is unclear as to why many prokaryotes lack this pathway. Toward understanding what could have led to the current distribution of bacterial NHEJ, we carried out comparative genomics and phylogenetic analysis across ~6,000 genomes. Our results show that this pathway is sporadically distributed across the phylogeny. Ancestral reconstruction further suggests that NHEJ was absent in the eubacterial ancestor and can be acquired via specific routes. Integrating NHEJ occurrence data for archaea, we also find evidence for extensive horizontal exchange of NHEJ genes between the two kingdoms as well as across bacterial clades. The pattern of occurrence in bacteria is consistent with correlated evolution of NHEJ with key genome characteristics of genome size and growth rate; NHEJ presence is associated with large genome sizes and/or slow growth rates, with the former being the dominant correlate. Given the central role these traits play in determining the ability to carry out recombination, it is possible that the evolutionary history of bacterial NHEJ may have been shaped by the requirement for efficient DSB repair.

3.2 Introduction

Accurate transmission of genetic material from parent to progeny is essential for the continuity of life. However, low rates of error during replication and DNA break-inducing mutagenic agents (such as ionizing radiation and reactive oxygen), while generating diversity for natural selection to act on (55), also adversely affect viability and could lead to diseases including cancer (56,57). Therefore, most cellular life forms invest in mechanisms that repair damaged DNA including double-strand breaks (DSBs).

Two major mechanisms of repair of DNA DSBs are homologous recombination and nonhomologous end joining (NHEJ). Recombination-based repair requires a homologous copy of the DNA around the damage site for repair to occur. In contrast, NHEJ (Fig. 3.1), the subject of the present work, directly ligates the DSB after detecting and binding the break ends (58,59). Where direct ligation is not possible—at breaks that generate complex ends—a processing step involving the removal of damaged bases and resynthesis of lost DNA is required. Such processing can be error-prone (60). Thus, NHEJ can be a double-edged sword: required for essential DNA repair when a homologous DNA copy is not available, but also prone to causing errors at complex DNA breaks.

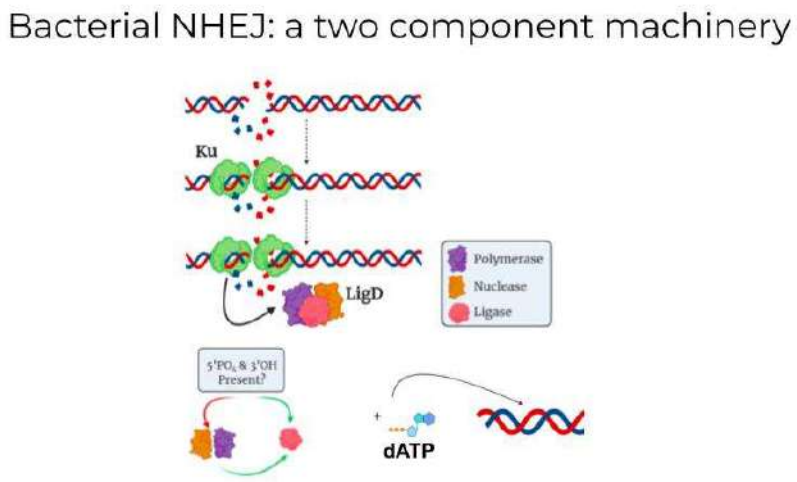


Figure 3.1 Bacterial Non-homologous end joining repair consists of a two component machinery - Ku and LigD. LigD is multidomain protein consisting of LigD-LIGASE, LigD-POLYMERASE and LigD-NUCLEASE domain

NHEJ is a major mechanism of DNA repair in eukaryotes. In bacteria however, homologous recombination-based repair is the most common mechanism of DNA repair; NHEJ, on the other hand, was described only recently (61), and its prevalence still remains to be systematically elucidated. Unlike eukaryotes, where NHEJ activity is regulated in a cell-cycle-dependent manner, it is unclear as to when NHEJ may be a preferred mode of bacterial DSB repair. Recent reports have shown that NHEJ can contribute to mutagenesis during a specific stage of bacterial growth, such as in stationary phase (32,59), raising the possibility that availability of a second copy of the genome for repair and/or growth phase may dictate whether recombination or NHEJ is employed for repair.

Experiments in *Mycobacterium* and *Pseudomonas* have shown that bacterial NHEJ repair machinery consists primarily of two proteins. The homodimeric Ku binds DNA break-ends and recruits the three-domain LigD harboring phosphoesterase (PE), polymerase (POL), and ligase (LIG) activity. These three domains are respectively required to process the ends, add bases if necessary, and subsequently ligate the break. Additionally, the POL domain mediates interaction between Ku and LigD (58,62–64). As an exception, studies in *Bacillus subtilis* found that these bacteria encode Ku along with a two-domain (LIG and POL) LigD (65). In the absence of LigD, it is also possible for LigC,

which contains only the LIG domain, to carry out repair (58,59). Though *Escherichia coli* does not encode NHEJ, expression of *Mycobacterium tuberculosis* Ku and LigD renders *E. coli* NHEJ proficient (66).

Studies in the early 2000s, with a small number of genomes, suggested that the distribution of NHEJ in bacteria could be patchy (61). Collectively, what does it mean for bacteria like *E. coli* to not code for NHEJ and others like *Mycobacterium tuberculosis* to harbor it? Given the large population sizes and relatively short generation times that make selection particularly strong, the question of the pressures that determine the deployment of the potentially risky NHEJ assumes importance. In this study, using bioinformatics sequence searches of Ku and LigD domains in the genomic sequences of ~6,000 bacteria, we have tried to 1) understand how pervasive their pattern of occurrence is, 2) trace their evolutionary history, and 3) understand what selection pressures could explain it.

3.3 Materials and Methods

3.3.1 Data

All “complete” and “latest” (assembly_summary.txt; as of January 2017) genome information files for ~6,000 bacteria were downloaded from the NCBI ftp website using in-house scripts—whole genome sequences (.fna), protein coding nucleotide sequences (.fna), RNA sequences (.fna), and protein sequences (.faa). All the organisms were assigned respective phylum and subphylum based on the KEGG classification (https://www.genome.jp/kegg/genome.html; as of May 2018).

3.3.2 Identification of NHEJ Repair Proteins

Bacteria were assigned to five broad categories, based on the status of NHEJ components—*NHEJ−*, *Ku only*, *LigD only*, *conventional NHEJ+*, and *nonconventional NHEJ+*. Conventional NHEJ was said to be present in bacteria harboring Ku and LigD having LIG, POL, and PE domains in the same protein. Nonconventional NHEJ included bacteria with Ku and at least one of the following: 1) all LigD domains present in different combinations in different proteins, 2) just the LIG and POL domains present in the same, or 3) different proteins, and 4) LIG with/without PE domain present in the same or different proteins. Organisms with PE and/or POL but not the LIG domain detected were assigned to the NHEJ− state if the Ku domain was not detected either; a *Ku only* state if the Ku domain was detected (fig. 1).

3.3.3 Ku and LigD Neighborhood Analysis

In-house python scripts were written to determine the proximity of *ku* and *ligD* on the genome using the annotation files. Taking one gene as the reference, the presence of the other gene was checked within a distance of ten genes upstream or downstream inclusive of both strands. The organization of NHEJ genes fell into three categories: 1) both genes on the same strand and within the distance range; this we call as operonic NHEJ, 2) both genes on different strands and within the distance range, and 3) genes outside the distance range.

This analysis was done only for *NHEJ+* organisms which code for LigD containing all the three domains as part of a single protein.

3.3.4 Identification of Ribosomal Ribonucleic Acid Sequences

A 16S ribosomal ribonucleic acid (rRNA) sequence database was downloaded from the Genomic-based 16S ribosomal RNA database (GRD) website (https://metasystems.riken.jp/grd/; last accessed November 2, 2020). An MSA and a profile of the database were made using muscle v3.8.31 (67) and hmmbuild (Finn et al. 2011), respectively. To detect 16S rRNA homologs in our database of 5,973 bacteria, nhmmer (68) was used with an *E*-value cutoff of 0.0001 where the GRD-based 16S rRNA sequence profile was used as the query. In-house python script was written to parse the output files and select the best hits for further analysis.

3.3.5 Calculation of Genome Sizes, Growth Rates, and G–C Content

In-house python scripts were written to calculate the genome sizes (GSs), growth rates (GRs), and G–C content of all bacteria in our data set (supplementary table 1). These calculations were made after excluding the plasmid sequence information from the whole genome sequence assembly files. Previous studies have shown a significant positive correlation between a bacteria's GR and the number of rRNA operons harbored in its genome (69,70). Therefore, the rRNA copy number was taken as a proxy for GR.

3.3.6 GS Randomization Analysis

For this analysis, 920 organisms harboring conventional NHEJ were considered. The total number of coding DNA sequences was used as the proxy for GS. Because the chance of finding a gene in a larger genome is more than that in a smaller genome, 920 genes were drawn at random from a pool of all the genes coming from 5,973 organisms in our data set and the genome coding DNA sequences size (GS) from which each gene was picked was noted. For each such iteration, the median GS was calculated. This was repeated for 100 iterations and a distribution of median GS was obtained. The nonparametric

Wilcoxon rank-sum test was used to compare this random distribution with the median GS of NHEJ-harboring bacteria.

3.3.7 HGT Analysis

Alien Hunter v1.7 (71) with default options was used to predict horizontally acquired regions (HARs)—based on their oligonucleotide composition—across bacterial genomes present in our data set. In-house python scripts were used to detect Ku and LigD in the predicted HARs. NHEJ was said to be acquired through horizontal gene transfer (HGT) if both Ku and LigD were present in the predicted HARs.

Phylogenetic incongruence between the archaeal–bacterial 16S rRNA-based species tree and the corresponding gene trees (for Ku, LigD-LIG domain, and RecA) was checked using the Approximately Unbiased (AU) statistical test (72). The AU test was carried out with 10,000 simulations with a constrained versus unconstrained approach (explained next). The MSA for each gene was run in an unconstrained mode for the given model parameters. Then, the alignments were run in a constrained mode with respect to their respective species tree topologies. p-AU gives the probability of identifying the gene family as having evolved according to the species, that is, the evolutionary history of the gene is the same as that of the species. This test was carried out using the -au options and -zb option implemented in IQTREE.

Further, to test the strength with which the aforementioned genes could be vertically transmitted or moved among closely or distantly related bacteria, nonparametric Mantel tests were carried out. Patristic distances were calculated between bacteria in the species tree and a given gene tree, respectively. A Pearson product–moment correlation coefficient r was calculated between these two distance matrices consisting of bacteria shared between the two phylogenies under test. The significance of correlation was assessed via randomization test by conducting 10,000 permutations of distance matrices. The correlation coefficient r was recalculated for each permutation to produce the null distribution and P value was obtained using the one-tailed test using the R package.

RANGER-DTL 2.0 (73) was used to predict the donors and recipients of NHEJ HGT events among and between bacterial and archaeal species. A non polytomous rooted 16S rRNA-based tree with archaea as an outgroup was used as a species tree. Optimal rootings for Ku and LigD-LIG domain-based bacterial–archaeal gene trees were determined using the *OptRoot* program with default options such that the duplication–transfer–loss (DTL) reconciliation cost was minimized. *Ranger-DTL* program was used to compute the optimal DTL reconciliation of a given rooted species tree—rooted gene tree pair. In case of multiple optimal reconciliations, the program inherently reports an optimal reconciliation sampled uniformly at random. Therefore, the analysis was run for 100 simulations each with a transfer cost $T = 1$, 2, or 3 (default), that is, a total of 300 simulations. The lower the transfer cost, the more the HGT events allowed during the reconciliations. *AggregateRanger* was used to compute support values for the most frequent mappings, that is, the donor species, by accounting for the variance due to multiple optimal reconciliations and alternative event cost assignments. An in-house python script was written to back trace the most frequent recipient for a given most frequent mapping for both Ku (supplementary table 3) and LIG domain (supplementary table 4). The results were overlaid on the respective species phylogeny using the features provided in the iTOL server (https://itol.embl.de/login.cgi; last accessed November 2, 2020).

Sequence alignments were viewed and pruned using Jalview (74). Principal component analysis of Ku domain sequences was carried out based on the method by (75).

3.3.8 Phylogenetic Tree Construction

For the construction of a species tree, one 16S rRNA sequence per genome was extracted into a multi-fasta file using an in-house python script. For bacteria with multiple 16S rRNA sequences, one 16S rRNA sequence was chosen such that it minimizes the number of *N*s in that sequence and has the maximum sequence length. In order to build a pruned phylogenetic tree, 970 bacteria were randomly selected such that a genus was represented exactly once for every NHEJ state. Please note that in the case of nonconventional NHEJ, all its four

subcategories, as described in the previous sections, were treated separately, when including bacteria at the genus level. An MSA was built using musclev3.8.31 (67) with default options. The conserved regions relevant for phylogenetic inference were extracted from the MSA using BMGE v1.12 (76). After the manual detection of the alignment, one spurious sequence was removed and the MSA was built again for 969 bacteria (supplementary table 5). Using IQTREE v1.6.5 (77), a maximum-likelihood (ML)-based phylogenetic tree was built with the best model chosen as SYM + R10 (LogL = −118,152.5876, BIC = 249,946.0604). ModelFinder (-m MF option) (78) was used to choose the best model for the tree construction compared against 285 other models. Branch supports were assessed using both 1,000 ultrafast bootstrap approximations (-bb 1000 -bnni option) (79) and SH-like approximate likelihood ratio (LR) test (-alrt 1000 option) (80).

A similar approach was used to build a phylogeny of 1,403 (supplementary table 6) organisms, nonredundant at the species level, comprising just two NHEJ states: 1) *NHEJ−* and 2) *conventional NHEJ+*.

For the construction of bacterial–archaeal gene trees (Ku, LigD-LIG, or RecA), only bacteria harboring one Ku and one conventional LigD in their genomes were included. Because the motivation behind the construction of gene trees was to study the origin of NHEJ in bacteria, all archaea harboring either Ku or LigD-LIG domain were included in the respective phylogeny. In species with multiple copies of RecA, RecA with the highest alignment length and identity was chosen for a given genome. The species in the bacterial–archaeal 16S rRNA-based phylogeny depended on the species included in the corresponding gene tree. The approach used to build these phylogenies was similar to the one described in the previous two paragraphs.

3.3.9 NHEJ Ancestral State Reconstruction Analysis

To trace the evolutionary history of NHEJ, four discrete character states were defined as follows: *Ku only*, *LigD only*, *NHEJ−*, and *NHEJ+*. States were estimated at each node using stochastic character mapping (52) with 1,000 simulations provided by the make.simmap() method in the R package phytools

v0.6-44 (81). The phylogenetic tree was rooted using the midpoint method and polytomies were removed by assigning very small branch lengths (10^{-6}) to all the branches with zero length. The prior distribution of the states was estimated at the root of the tree. Further, by default the method assumes that the transitions between different character states occur at equal rates. This might not always be true, especially with complex traits where it is supposedly easier to lose than gain such characters. Therefore, for the estimation of transition matrix *Q*, three discrete character evolution model fits were compared: Equal Rates (ER), Symmetric (SYM), and All Rates Different (ARD). This allowed for models that incorporate asymmetries in transition rates. Based on Akaike Information Criterion (AIC) weights, the ARD model ($w\text{-}AIC_{ARD}$ =1, $w\text{-}AIC_{SYM}$ = 0, and $w\text{-}AIC_{ER}$= 0) was chosen as the best fit with unequal forward and backward rates for each character state transition. Finally, *Q* was sampled 1,000 times from the posterior probability distribution of *Q* using Markov chain Monte Carlo and 1,000 stochastic maps were simulated conditioned on each sampled value of *Q*. This strategy was used to reconstruct ancestral states for both phylogenies comprising 969 and 1,403 organisms as described in the previous sections.

Ancestral states for GS, a continuous trait, were reconstructed using the fastAnc() method employed in phytools v0.6-44 based on the Brownian motion model. This model was found to be the better fit model as compared with multiple rate model (Bayesian Predictive Information Criterion $[BPIC]_{Brownian} < BPIC_{stable}$ and Proportional Scale Reduction Factor approaching <1.1 well within 1,000,000 iterations), assessed using StableTraits (82). The multiple rate model allows for the incorporation of neutrality and gradualism associated with Brownian motion and also includes occasional bursts of rapid evolutionary change. Ancestral states for GR were reconstructed by converting it into a binary trait. An organism was said to be slow growing if it encoded rRNA copy numbers less than or equal to the median rRNA copy number (median = 3) and fast growing otherwise. The reconstruction was carried out using the same approach that was used for estimating NHEJ ancestral states, as described in the previous paragraph.

3.3.10 NHEJ and Genome Characteristics Phylogenetic Comparative Analysis

The two genome characteristics—GS and GR—were compared across bacteria with different NHEJ states. The distributions across bacteria with different NHEJ states were first compared assuming statistical independence of bacteria, using Wilcoxon rank-sum test, wilcox.test() in R. Next, two measures of phylogenetic signal—Pagel's λ (83) and Blomberg's K (84)—were used for detecting the impact of shared ancestry, for GS and GR across bacteria, using the phylosig() routine in phytools R package (81). A phylogenetic analysis of variance (ANOVA), employed in phytools R package v0.6-44, was carried out with 1,000 simulations and Holm–Bonferonni correction to control for familywise error rate, based on a method by Garland et al. (1993)(49), to compare the genome characteristics in a phylogenetically controlled manner. Please note that GSs and rRNA copy numbers were $\log_{10}$ transformed for all the analysis.

3.3.11 Correlated Evolution Analysis

Two relationships— (NHEJ repair and GS) and (NHEJ repair and GR)—were quantified using a statistical framework. To test if changes in genome characteristics occur independently of NHEJ or whether these changes are more (or less) likely to occur in lineages with (or without) NHEJ, two models of evolution were considered—independent and dependent. In the independent model, both the traits were allowed to evolve separately on a phylogenetic tree, that is, non-correlated evolution. In the dependent model, the two traits were evolved in a non-agnostic manner, that is, correlated evolution. The NHEJ repair trait had two repair character states—NHEJ− (0) and conventional NHEJ+ (1). GSs, a continuous trait, were converted into a binary state as well. For this, the mean GS of organisms with 0 or 1 NHEJ repair state was computed. A "lower" state (0) was assigned if a value was less than the mean and a "higher" state (1) if the value was more. The same approach was used to convert GRs (rRNA copy number) into a binary state.

A continuous-time Markov model approach was used to investigate correlated evolution between NHEJ repair and genome characteristics. First, the ML

approach (53) was used to calculate log-likelihoods for the two models of evolution per trait pair: 1) NHEJ repair and GS; 2) NHEJ repair and GR. A LR statistic was calculated for both comparisons, followed by a chi-square test to assess if the dependent model was a better fit. The degrees of freedom are given by dfchi-square test= ($n_{\text{rate-dependent model}} - n_{\text{rate-independent model}}$). There are eight transition rates in the dependent model across four states (00,01,10,11) and four transition rates in the independent model across two states (0,1; 0,1). Therefore, the test was run with four degrees of freedom.

The ML approach implicitly assumes that the models used for hypothesis testing are free of errors. Therefore, to make the analysis robust, the Bayesian reverse jump Markov chain Monte Carlo (RJMCMC) approach was used to calculate the marginal log-likelihoods of the independent and dependent models of evolution (54). This approach takes into consideration the uncertainty and minimizes the error associated in calculating the parameters used in each of our model(s), ensuring reliable interpretations. Log Bayes factor was used to assess the better fit out of the two models.

BayesTraits v3 (54) was used to carry out both ML- and Bayesian RJMCMC-based correlated evolution analysis as described above for both the models. ML was run using the default parameters. Bayesian RJMCMC was run for 5,050,000 iterations, sampling every 1,000th iteration with a burn-in of 50,000. For the estimation of marginal likelihood, a stepping stone sampler algorithm was used where the number of stones was set to 100 and each stone was allowed to run for 10,000 iterations.

3.3.12 Phylogenetic Logistic Regression Analysis

A method developed by (85) was used to carry out the phylogenetic logistic regression analysis provided in the R package phylolm v2.6 as the subroutine phyloglm() with method = logistic_IG10. NHEJ repair was taken as the binary-dependent variable with two states: NHEJ− (0) and conventional NHEJ+ (1). The two independent continuous variables were GS and GR. Before proceeding with the analysis, these independent variables were checked for multicollinearity by calculating variance inflation factor (VIF),

$VIF = 1 / (1 - R^2)$

Three models were tested: 1) NHEJ ~ GS, 2) NHEJ ~ GR, and 3) NHEJ ~ GS + GR. The best model was chosen according to AIC scores.

All the scripts used for analysis were written in python, perl, or R. Statistical tests and data visualizations were carried out in R and iTOL server.

3.4 Results

3.4.1 NHEJ Is Sporadically Distributed across Bacteria

To identify NHEJ machinery across bacteria, we used the reference sequences of the Ku domain, and the LIG, POL, and PE domains of LigD from *P. aeruginosa* to search ~6,000 complete bacterial genomes for homologs (see Materials and Methods). We defined bacteria encoding Ku and the complete, three-domain version of LigD as those harboring a conventional NHEJ system. Organisms lacking the POL and/or the PE domains of LigD, and those encoding these domains in separate proteins, were defined as those carrying nonconventional NHEJ (fig. 3.1).

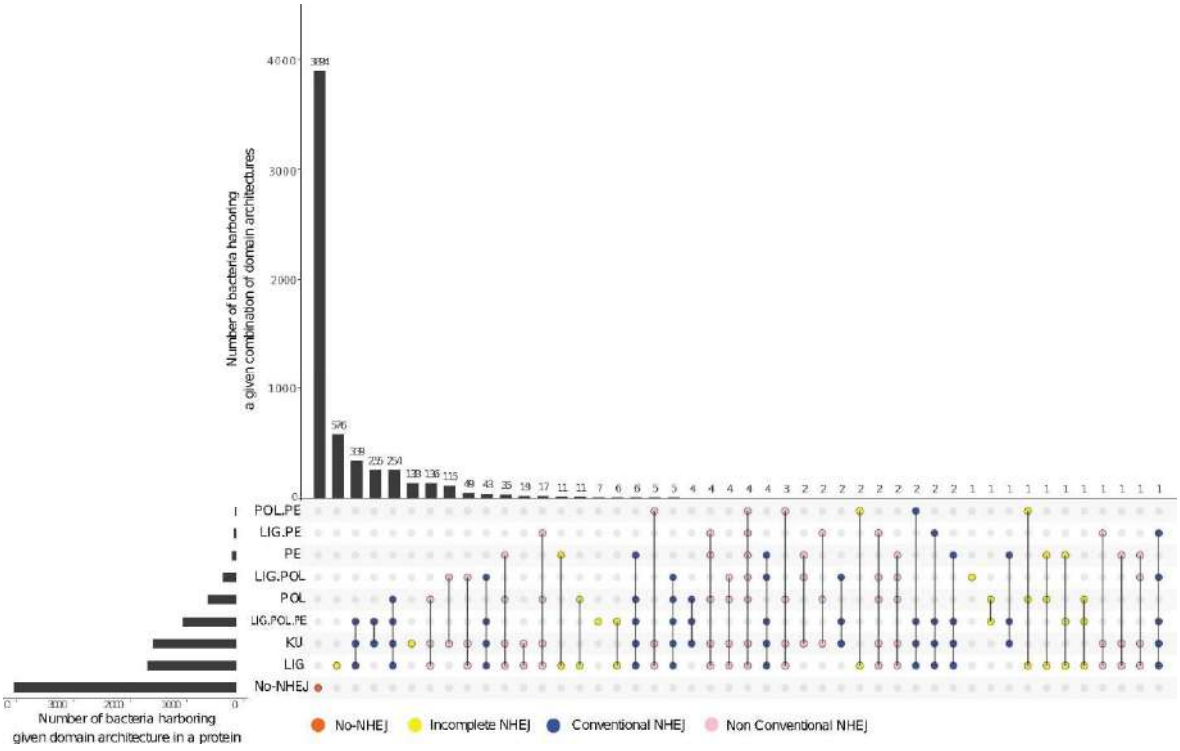


Figure 3.2 Distribution of NHEJ components in bacteria. An UpSet plot depicting the number of bacteria harboring a certain type of domain architecture per protein and number of bacteria harboring a combination of domain architecture in their respective genomes is shown for 5,973 analyzed genomes. NHEJ− *(orange),* Ku only *(yellow),* LigD only *(yellow),* Conventional NHEJ+ *(blue), and* nonconventional NHEJ+ *(pink).*

We found NHEJ in only ~1,300 (22%) genomes studied here. There were various combinations of Ku and LigD domains across these organisms, but a large majority (920) carried conventional NHEJ. Seventy-five percent bacteria harboring conventional NHEJ coded for Ku and LigD in a 10-kb vicinity of each

other, with 60% organisms carrying Ku and LigD on the same strand of the 10-kb vicinity. Most bacteria (84%) harboring NHEJ coded for a single copy of Ku, whereas the remaining coded for 2–8 Ku copies in their genomes. For example, as reported by (86,87), we identified four Ku-encoding genes in *Sinorhizobium meliloti*. About two-thirds of NHEJ positive bacteria carried multiple copies of the LIG domain, 37% carried multiple copies of the POL domain, and 8% bacteria had multiple copies of the PE domain (supplementary table 1). We also noticed that 138 (2.3%) organisms encoded Ku and not LigD, and 619 (10.3%) only LigD and not Ku (fig. 3.2 and supplementary table 1).

NHEJ was not restricted to specific bacterial classes (fig. 3.3) and was found in ten classes. We found a significant enrichment of conventional NHEJ in Proteobacteria (Fisher's Exact test, $P = 3.8 \times 10^{-5}$, odds ratio: 2.04) and Acidobacteria (Fisher's Exact test, $P = 5 \times 10^{-2}$, odds ratio: 5.286). All Bacteroidetes with NHEJ harbor a conventional NHEJ, although we could not assign statistical significance to it (Fisher's Exact test, $P = 0.14$, odds ratio: 1.38). In contrast, nonconventional NHEJ repair was significantly overrepresented in Firmicutes (Fisher's Exact test, $P = 1.23 \times 10^{-13}$, odds ratio: 6) and Actinobacteria (Fisher's Exact test, $P = 2.12 \times 10^{-15}$, odds ratio: 6.14). Twenty-four phyla did not include any NHEJ positive organisms (supplementary table 7).

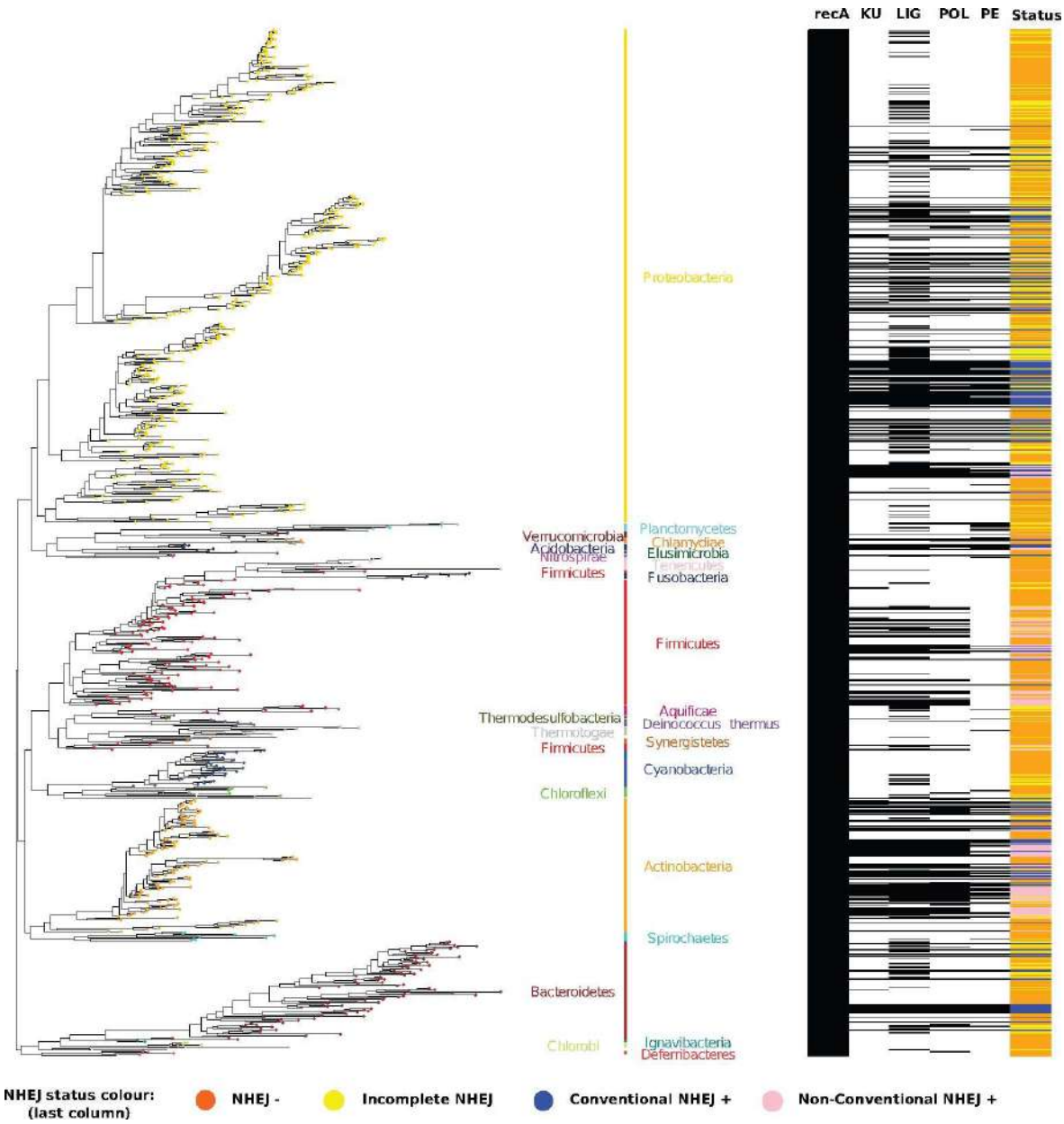


Figure 3.3 NHEJ is sporadically distributed across bacteria. 16S rRNA-based species phylogenetic tree of 969 bacterial species (left) with presence/absence matrix of RecA, KU, LIG, POL, and PE domains (right). These species were included such that each genus was chosen once for each NHEJ state (see Phylogenetic Tree Reconstruction section for further details). Tip labels, representing bacteria, are colored according to the phylum names and bars (right of the phylogenetic tree) that the given tip belongs to. The phylum names arranged on either sides of the vertical bars are for representational purposes only. The first five columns of the presence/absence matrix (extreme right of the figure) depict the status if the given protein (RecA) or domain (Ku, LIG, POL, and PE) is present (black horizontal bar) or absent (white horizontal bar) in the corresponding bacteria. Each horizontal bar maps to a bacterial tip on the phylogenetic tree (left). Horizontal bars in the last column of the matrix represent the overall NHEJ status for a given species (color legend same as in fig. 3.2). NHEJ status—orange: NHEJ−*; yellow:* incomplete NHEJ*; blue:* conventional NHEJ+*; pink:* nonconventional NHEJ+*.*

3.4.2 NHEJ Was Gained and Lost Multiple Times through Evolution

We traced the number of NHEJ gains and losses starting from the eubacterial ancestor to the species at the tips of the 16S-based bacterial phylogenetic tree. To trace the evolutionary history of NHEJ, we defined four discrete character states: *Ku only*, *LigD only* (conventional and nonconventional), *NHEJ–*, and *NHEJ+*. Note that an *NHEJ+* state is defined only when both Ku and LigD are present in a bacterium. We calculated the posterior probabilities (pp) of each character state per node on the phylogeny, the distribution of the number of times each of the 12-character state transitions occurred (fig. 3.4A) and the distribution of the total time spent in each state (supplementary fig. S1). We performed this analysis for a set of 969 genomes in which each genus was represented once for each state (see Materials and Methods).

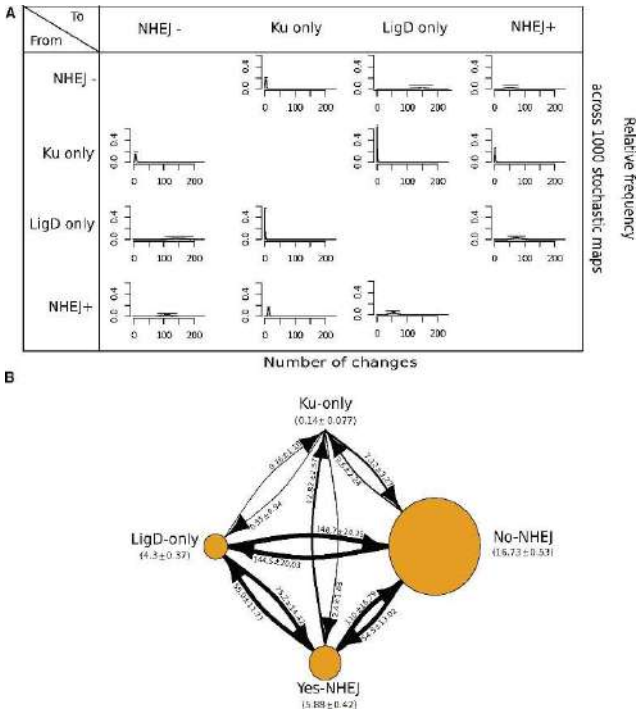


Figure 3.4 Transitions to a Ku only state are rare. (*A*) A matrix depicting relative frequency of number of changes of a state transition type across 1,000 stochastic maps. (*B*) A state transition diagram depicting the number of transitions between two given states and the time spent in each state during NHEJ evolution.

The node size is proportional to the amount of time spent in a particular state. The arrow size is proportional to the number of transitions from one state to another.

We first asked if NHEJ was present in the common eubacterial ancestor and, given the sporadicity of NHEJ, subsequently lost in several lineages (supplementary table 8). We assigned a major primary gain to an internal ancestral node if 1) all nodes leading to it from the root had *NHEJ−* state; 2) the pp of either ***NHEJ+***, ***LigD only***, or ***Ku only*** at that ancestral node was ≥0.7; 3) a gain of ***LigD only*** or ***Ku only*** was followed by a transition to ***NHEJ+***; and 4) if it had at least three descendent species. We observed multiple major independent primary gains at ancestral nodes within Bacteroidetes, Actinobacteria, Firmicutes, Acidobacteria, and multiple subclades of Proteobacteria (supplementary table 8). It follows that the common eubacterial ancestor likely did not have NHEJ (fig. 3.5).

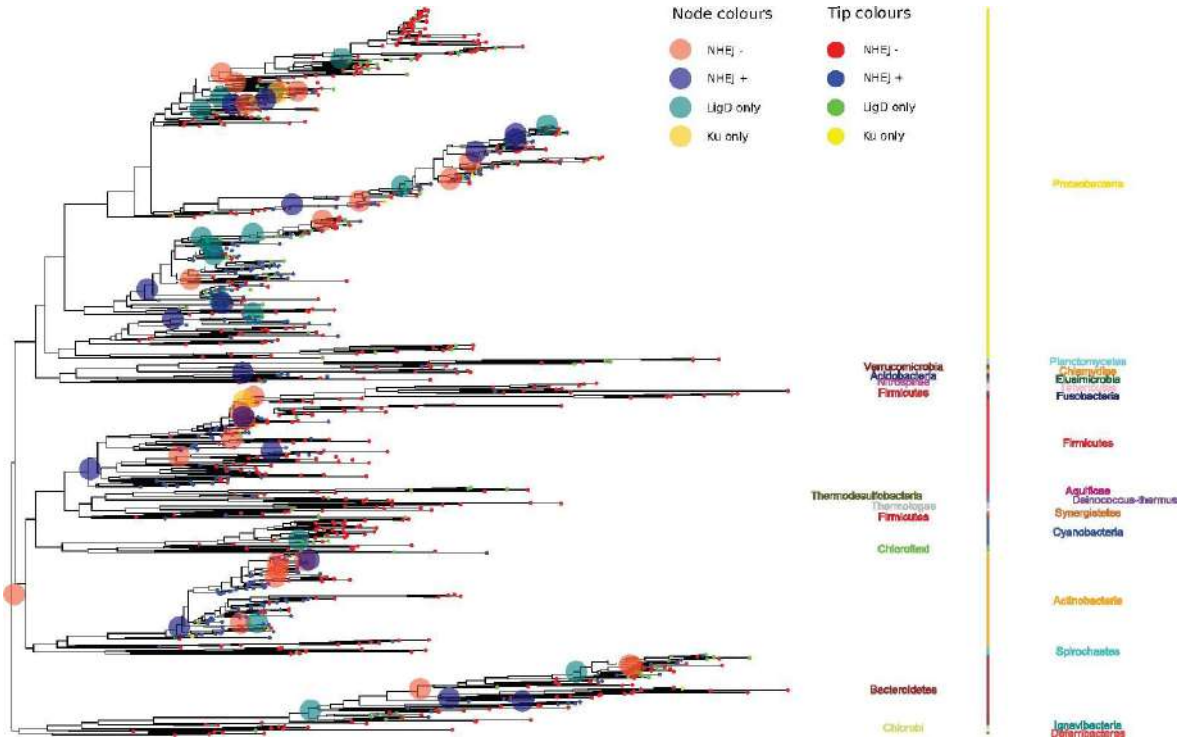


Figure 3.5 NHEJ was gained and lost multiple times through evolution. A trace of the evolutionary history of the two-component NHEJ system across 969 bacteria. These species were included such that each genus was chosen once for each NHEJ state (see Phylogenetic Tree Reconstruction section for further details). The phylum names arranged on either sides of the vertical bars are for representational purposes only. The tip and node labels are colored according to the NHEJ states—red: NHEJ−*; yellow:* Ku only*; green:* LigD only*; blue:* NHEJ+ *(conventional and nonconventional). NHEJ state for nodes is shown only when the posterior probability support is >70%; interpreted as change in NHEJ state at that node as compared with shallower phylogenetic depths.*

The gain of NHEJ can be sequential, gaining either *Ku only* or *LigD only* followed by the gain of the other component; or it can be a one-step acquisition of both components (fig. 3.4B). The most common transition from an *NHEJ−* state was to a *LigD-only* state. Also frequent was the direct acquisition of both components to transition from an *NHEJ−* to an ***NHEJ+*** state. Transition from *NHEJ−* to ***Ku only*** was negligible. In the reverse direction, a one-step loss of both Ku and LigD is most likely. Again, the *Ku only* state is rare.

A one-step transition from *NHEJ−* to ***NHEJ+*** is likely through HGT. Sixty bacterial genomes belonging to the phyla Alpha-proteobacteria (in particular the Rhizobiales)—and Beta-proteobacteria, and Streptomycetales carried their NHEJ components on plasmids (supplementary table 1). However, based on abnormal word usage statistics (see Materials and Methods), we could not find NHEJ to be a part of the horizontally acquired component of the chromosomes of any bacterial genome. At least two *NHEJ−* to ***NHEJ+*** transitions occurred close to the root, and it is possible that the predictions of horizontally acquired NHEJ systems made so far may be an underestimate (fig. 3.5). We investigate this in greater detail below.

In summary, 1) the common eubacterial ancestor was devoid of NHEJ; 2) NHEJ was gained and lost multiple times; and 3) transitions to a *Ku only* state are rare.

3.4.3 NHEJ and HGT

Experimental studies in the archaea *Methanocella paludicola* (88,89) have confirmed the presence of a functional NHEJ repair, with crystal structures revealing close relationship with the bacterial proteins (88).

To assess the possibility of horizontal transfer of NHEJ machinery across prokaryotes, we first performed a domain wise search in 243 archaea to complement the data we had assembled for bacteria. These searches revealed the presence of both Ku and LigD-LIG domains in ten archaeal species (see Materials and Methods; supplementary table 2). However, 230 archaeal

genomes encoded LigD but no Ku. This lends support to previous reports suggesting that NHEJ is rare in archaea (88).

In order to check for horizontal transfer events between bacteria and archaea, we used phylogenetic methods based on detecting conflicts between an organismal phylogeny and a phylogeny inferred for Ku and LigD-LIG domains, respectively. This method allowed us to test for any ancient transfers across bacteria as well. We found that NHEJ proteins undergo HGT events at significantly high rate (p-AU = 0; fig. 3.6B and C; see Materials and Methods) and these are not limited to closely related species or co-speciation events (fig. 3.6D; Mantel test, $P < 10^{-4}$, r_{Ku} = 0.25; $P < 10^{-4}$, r_{LIG} = 0.4). We also noted incongruence with respect to the RecA phylogeny (p-AU = 0; fig. 3.6A) as has been reported before (90,91). However, these were limited at best to transfers among closely related species (fig. 3.6D; Mantel test, $P < 10^{-4}$, r_{RecA}= 0.94).

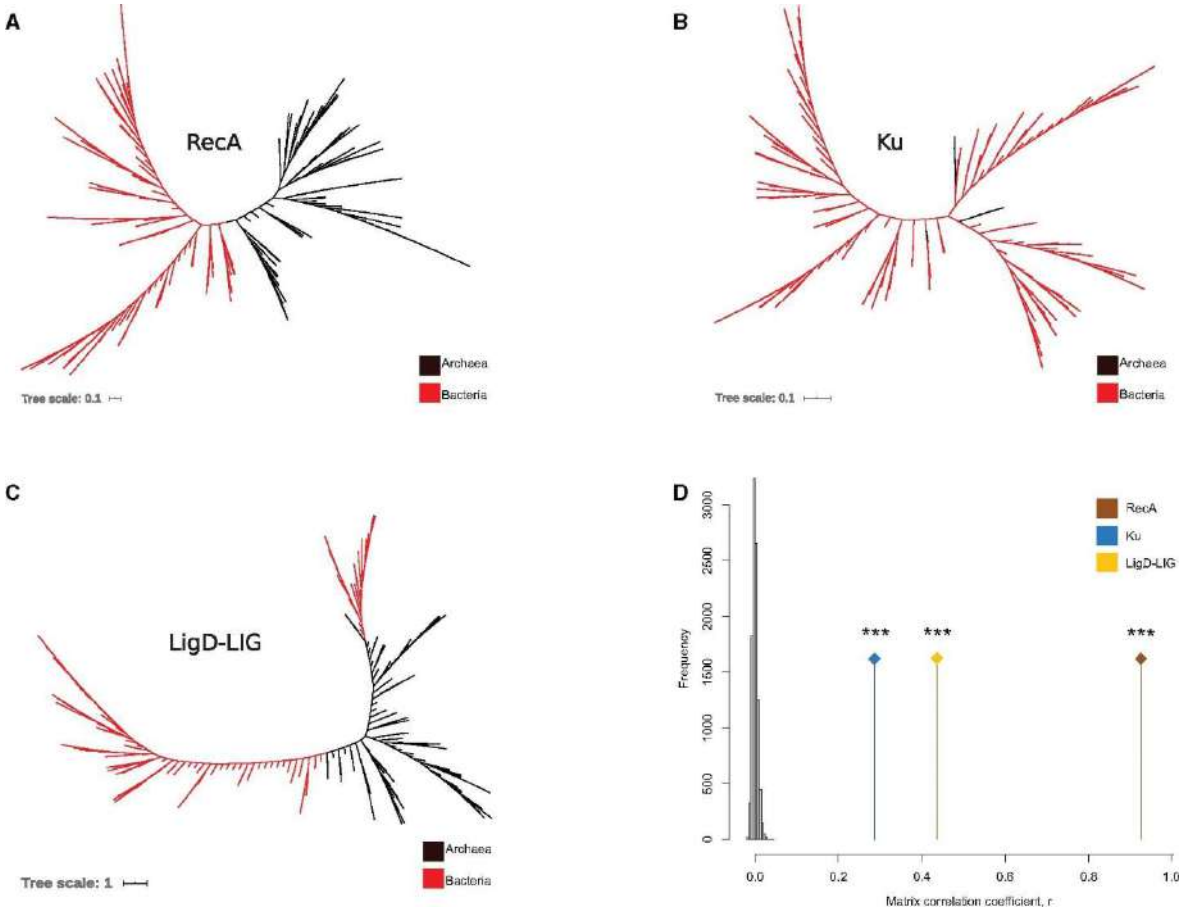


Figure 3.6 Phylogenetic methods suggest a strong role of HGT in NHEJ evolution. (A) *Unrooted RecA tree,* (B) *unrooted Ku tree,* (C) *unrooted LigD-LIG tree, and* (D) *Mantel test correlation coefficient* (r)

comparing RecA, Ku, and LigD-LIG distance matrices with 16S rRNA distance matrices, respectively, compared against a null distribution of r obtained by 10,000 matrix randomizations.

An incongruence between a species and gene tree could result due to processes other than HGT, like duplications and losses. Therefore, to predict the most frequent transfer events, we used a reconciliation approach based on the DTL model. DTL employs a parsimonious framework where each evolutionary event is assigned a cost and the goal is to find a reconciliation (possible evolutionary history of the gene tree inside a species tree) with minimum total cost. We observed a high rate of HGT between bacterial clades—Firmicutes, Actinobacteria, and Proteobacteria and Archaea, where each of these played the role of a donor and a recipient in Ku (fig. 3.7A) and LigD-LIG transfer events (fig. 3.7B). We observed that all proteobacterial species—with the exception of delta-proteobacteria, which was a donor of Ku to archaeal recipients—were recipients of both Ku and LigD from archaea or other distantly related bacteria. On one hand, we found evidence of Ku transfers from Archaea to Firmicutes and Actinobacteria and on the other hand, LigD-LIG transfers most likely occurred from Firmicutes and Actinobacteria to Archaea. Together, this raises the possibility of NHEJ transfers between bacteria and archaea.

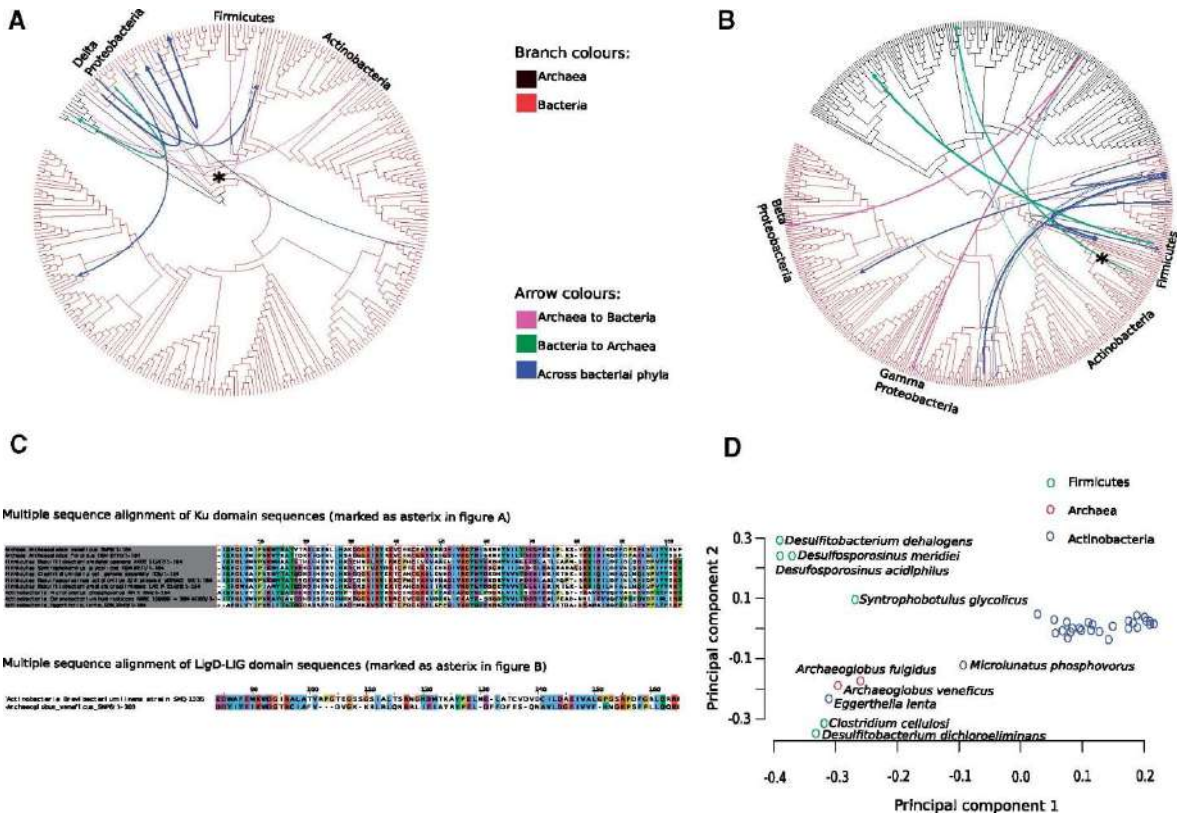


*Figure 3.7 Extensive HGT among bacterial phyla and between bacteria and archaea. 16S rRNA-based species tree depicting the most frequent donor–recipient pairs involved in (*A*) Ku HGT events and (*B*) LigD-LIG HGT events. The bacterial species included in both the species tree coded for one Ku and one LigD only. The archaeal species were included 1) in (*A*) if they had at least a Ku and 2) in (*B*) if they had at least a LigD-LIG domain (see Materials and Methods for more details). The width of the arrow corresponds to the number of reconciliations supporting a given transfer event. (C) Ku domain MSA (upper panel) of prokaryotes belonging to genus* Archaeoglobus *and phyla Firmicutes and Actinobacteria included in (A) (transfer event marked as asterisk). LigD-LIG pairwise alignment (lower panel) of an Archaea–Actinobacteria HGT transfer event in (*B*) (marked as asterisk). (*D*) Principal component analysis of Ku domain sequences evolved from ancestors included in transfer event in (*A*) (marked as asterisk).*

An example of the former is depicted as a MSA of Ku domain sequences belonging to Archaea, Firmicutes, and Actinobacteria in figure 3.7C (upper panel). This transfer event corresponds to the asterisk marked in figure 3.7A—corresponding to that between the ancestor of the genus *Archaeoglobus* and that of Firmicutes and Actinobacteria. We further carried out a principal component analysis of these domain sequences that had evolved from the aforementioned ancestors (fig. 3.7D). Along with the first principal component, all but two Actinobacteria—*Eggerthella lenta* and *Microlunatus phosphovorus*—form a distinct cluster from Archaea and Firmicutes. Along the second principal component, we see two distinct clusters. The cluster on the bottom left consists of anaerobic prokaryotes—Archaea (genus *Archaeoglobus*), Firmicutes (*Clostridium cellulosi* and *Desulfitobacterium dichloroeliminans*), and

Actinobacteria (*Eggerthella lenta*), highlighting the possibility of HGT among these prokaryotes. Another instance of a LigD-LIG transfer is depicted in figure 3.7C, lower panel. This transfer event corresponds to the asterisk in figure 3.7B, involving the Actinobacteria—*Brevibacterium linens*, and Archaea—*Archaeoglobus veneficus* (coding for both Ku and LigD-LIG domains).

In addition to the evidence supporting HGT of NHEJ components between Archaea and Bacteria, we observed transfers among different bacterial clades as well (fig. 3.7A and B; blue arrows). We found Ku transfers between donor–recipient pairs: 1) Alphaproteobacterium (*Asticcacaulis excentricus*) and common ancestor of Acidobacteria (genus *Acidobacterium*, *Granulicella*, and *Terriglobus*) and 2) common ancestor of Delta-proteobacteria genus *Geobacter* and *Chlamydiae* (*Parachlamydia acanthamoeba*). For LigD-LIG transfers, we observed the following donor–recipient pairs—1) Proteobacteria (*Phenylbacterium zucineum*) and Acidobacteria (*Terriglobus roseus*) and 2) common ancestor of Actinobacteria (genus *Eggerthella)* and Firmicutes (*Desulfitobacterium dicholoroeliminans*). A full list of donor–recipient events can be found in supplementary files 3 and 4, for Ku and LigD-LIG domains, respectively.

(92) carried out HGT detection of bacterial core genes among prokaryotes. They found that a majority of these transfers occurred from bacteria to archaea and that these genes were mostly metabolic genes. Overall, our study is consistent with their observation, with additional evidence showing a possibility of NHEJ transfers from archaea to bacteria as well. We also show evidence of NHEJ transfers between closely and distantly related bacteria. Using the approach used in our study, it remains to be tested how HGT events have shaped non-core genes like other repair pathways throughout evolution in prokaryotes.

3.4.4 NHEJ Occurrence Is Associated with GS, GR, and G–C Content

Recently, Ku-encoding organisms were shown to have higher genomic G + C content (93). Given its central role in DNA repair, we asked whether any other

genome characteristics could also be associated with the presence or absence of NHEJ. First, we verified that the findings of Weissman et al. on the correlation between the presence of Ku and G + C content held true for *NHEJ+* states as defined in our study (fig. 3.8A and supplementary figs. S4, S5, and S7*C*). Along with this, we tested two additional characteristics: GS and GR (as measured by the copy number of rRNA operons), both of which could determine the availability or the lack of a homologous template for high fidelity recombination-based repair. We restricted these analyses to conventional NHEJ-harboring bacteria as a proxy for repair proficiency and compared them with *NHEJ–* genomes. Data including nonconventional NHEJ are shown in supplementary figures S2 and S3.

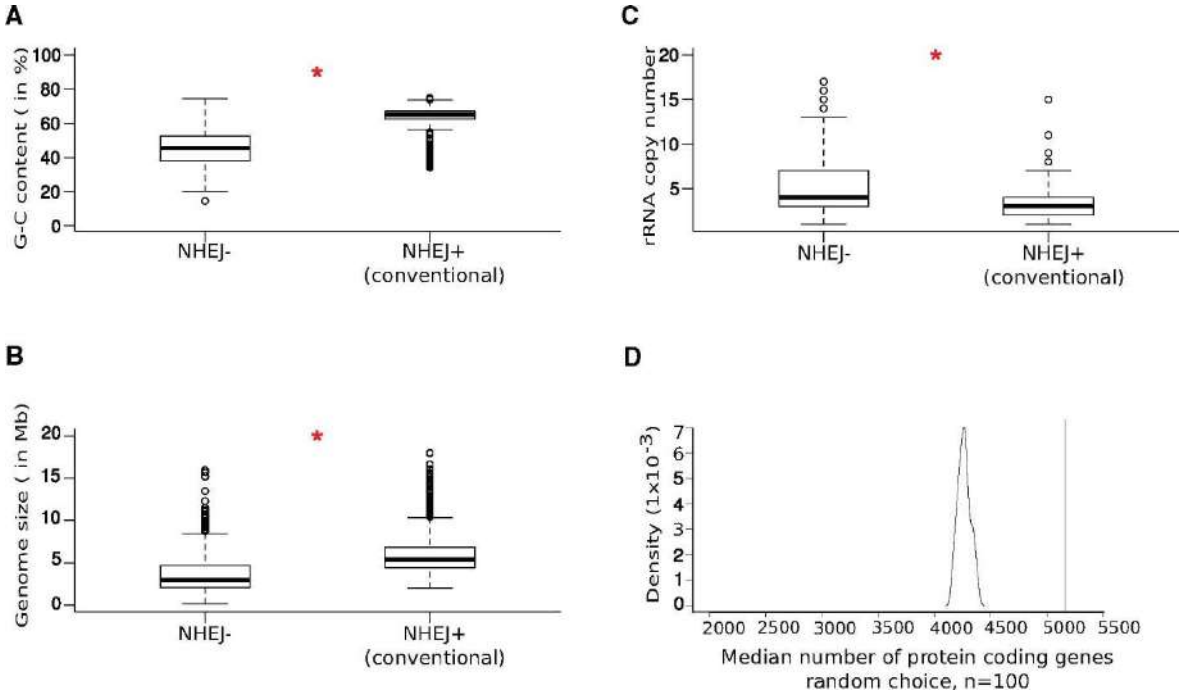


Figure 3.8 NHEJ presence and absence is associated with GS, GR, and G–C content. The red asterisk indicates statistical significance (Wilcoxon rank-sum test; P *value < 0.01). (*A*) Boxplot comparing the distribution of G–C content between* NHEJ– *and* conventional NHEJ+ *bacteria. (*B*) Boxplot comparing the distribution of GS between* NHEJ− *and* conventional NHEJ+ *bacteria. (*C*) Boxplot comparing the distribution of rRNA copy number between* NHEJ− *and* conventional NHEJ+ *bacteria. (*D*) A density distribution plot depicting the distribution of GS (median number of protein coding sequences) expected by a random distribution (black) where the probability of having NHEJ is linearly proportional to GS and the median GS of organisms harboring NHEJ (red).*

Bacteria with NHEJ were found to have larger genomes (median = 5.4 Mb) than those without NHEJ (median= 2.9 Mb; Wilcoxon rank-sum test, $P < 10^{-15}$; fig. 3.8B and supplementary fig. S7*A*) and significantly larger than that expected by

a random distribution in which the probability of having NHEJ is linearly proportional to GS (Wilcoxon rank-sum test, $P < 10^{-15}$, across 100 simulations; fig. 3.8D). This relationship was found to be true within the phylum Proteobacteria, Actinobacteria, Bacteroidetes, and Firmicutes as well (supplementary fig. S6).

In addition, bacteria harboring NHEJ were found to have significantly fewer rRNA copies (median = 3), and by inference slower GRs, than bacteria without NHEJ (median = 4; Wilcoxon rank-sum test; $P < 10^{-15}$; fig. 3.8C). Although the distribution of rRNA copy numbers for genomes without NHEJ was broad, those with conventional NHEJ fell within a narrow range, representing relatively slower growth (supplementary fig. S7*B*). At the phyla level, this relationship was found to hold true for Proteobacteria and Actinobacteria, whereas there was no significant difference for Bacteroidetes and Firmicutes, respectively (supplementary fig. S8).

In order to confirm the result in a phylogenetically controlled manner, Pagel's λ and Blomberg's K were used to first measure whether closely related bacteria tended to have similar GSs and GRs in the data set (see Materials and Methods). These measures suggest that phylogenetic coherence is significantly greater than random expectations for both the genome characteristics (supplementary table 9). Therefore, the distributions of GS between bacteria with different NHEJ status were compared while accounting for the statistical nonindependence of closely related taxa (see Materials and Methods). We found a significant difference in $\log_{10}$(GSs) between bacteria with conventional NHEJ and without the repair (phyloANOVA; $P = 6 \times 10^{-3}$); the characters being mapped on the phylogenetic tree of 969 bacteria with five discrete groups: *NHEJ −*, *Ku only*, *LigD only*, *conventional NHEJ+*, and *nonconventional NHEJ+*. However, we did not observe a significant difference in $\log_{10}$(rRNA copy number) between the two groups of bacteria (phyloANOVA; $P = 1$; see Discussion).

We used ML as well as Bayesian approaches to test whether the observed associations of conventional NHEJ individually with large genomes and slow GRs are indicative of dependent or independent evolution of these traits on the phylogenetic tree (see Materials and Methods). Both suggested that the phylogenetic data fit models of evolution in which conventional NHEJ presence or absence and GS or GR are evolving in a correlated manner (table 3.1). This strengthens the association between the tested variables in a phylogenetically controlled way. Phylogenetic logistic regression of the conventional NHEJ occurrence with both the continuous independent variables showed, however, that GS is the stronger correlate (see Materials and Methods; table 3.2).

Table 3.1
Maximum Likelihood and Bayesian RJMCMC Results for Two Character Pairs Tested for Correlated Evolution: 1) NHEJ State and Genome Size and 2) NHEJ State and Growth Rate

Method	Correlation Pair	Log-Likelihood (Independent Model)	Marginal Log-Likelihood (Independent Model)	Log-Likelihood (Dependent Model)	Marginal Log-Likelihood (Dependent Model)	Likelihood Ratio Test (LRT) Statistics	Bayes Factor (>2 = Better Fit)
Maximum likelihood	NHEJ state and genome size	−1,059.97	—	−1,010.002	—	LR = 99.94 $P < 0.001$	—
Bayesian RJMCMC	NHEJ state and genome size	—	−1,110.14	—	−1,042.25	—	135.78
Maximum likelihood	NHEJ state and growth rate	−975.301	—	−950.01	—	LR = 25.29 $P < 0.001$	—
Bayesian RJMCMC	NHEJ state and growth rate	—	−1,051.529	—	−985.579		131.9

NOTE. —A chi-squared-based LRT with four degrees of freedom was used to test the better model—1) independent model where the character pair evolved independent of each other and 2) dependent model where the characters were allowed to evolve assuming a correlated evolution—based on maximum likelihood. Bayes factor was used to test the better model, based on Bayesian RJMCMC analysis.

Table 3.2
Phylogenetic Logistic Regression for Three Models, Based on a Phylogenetic Tree of 1,403 Species of Bacteria Harboring Either *NHEJ−* or *Conventional NHEJ+* State

Model	AIC Value	Pen.LogLik	Alpha	Coefficient Estimate (CE)	*P* Value
NHEJ ~ GS	809.5	−383.3.	13.88	6.7×10^{-7}	10^{-15}
NHEJ ~ GR	916.7	−450.1	23.85	−0.133	7×10^{-4}
NHEJ = GS + GR	841.2	−395.1	23.51	$CE_{GS} = 8.1 \times 10^{-7}$ $CE_{GR} = -0.4$	10^{-15} 3.8×10^{-12}

NOTE.—Alpha represents the phylogenetic signal of the dependent variable, that is, NHEJ state in our case. The higher the alpha, the lesser the phylogenetic signal. AIC or the Akaike Information Criterion is used to select the best model for the NHEJ state, out of the three tested against two independent variables—GS and GR. The two independent variables were tested against multicollinearity, with variance inflation factor or VIF = 0.93715 (VIF < 10 is preferred), making the analysis reliable.

As a case study where the association of both GS and GR with NHEJ evolution was prominent, we found a gain of conventional NHEJ in the ancestor of two genera belonging to Corynebacteriales—*Mycobacterium* and *Corynebacterium*—where the former retained and the latter had a secondary loss of the machinery. Phylogenetic ancestral reconstruction analysis revealed an increase in GS in the ancestor of Corynebacteriales, followed by an NHEJ gain. Although *Mycobacterium* retained NHEJ, *Corynebacterium* lost the machinery along with a decrease in GS. Using a similar analysis, GR mapped to this subclade revealed an increase in rRNA copy number in *Corynebacterium* (supplementary fig. S9; see Materials and Methods).

3.5 Discussion

Taking advantage of the availability of a large number of genomes, we confirm the non-ubiquitous nature of NHEJ across the bacterial domain (94) with 22% bacteria coding for it. At the taxa level, although some phyla retain this sporadicity, others are devoid of NHEJ repair machinery, consistent with previous studies on distribution of bacterial NHEJ proteins (94). In line with previous reports on the multiplicity of NHEJ in certain bacteria (60,86,95–98), we further discovered a sporadic presence of multiple copies of NHEJ proteins extending to different subclades of Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes, etc. (supplementary table 1). A trace of evolutionary history suggested that NHEJ was most likely absent in the eubacterial ancestor. Furthermore, 96% of the archaeal genomes coded for LIG but no Ku. Therefore, by inference, NHEJ was absent in the Most Recent Common Ancestor of all prokaryotes as well. It was instead gained independently multiple times in different bacterial lineages. However, these primary gains were not sufficient to stabilize the repair states in subclades, because a large number of secondary losses and gains were observed across the phylogeny (elaborated further below).

Our analysis suggests that there are two common methods to arrive at an *NHEJ*+state: 1) the most common way to gain NHEJ was by the acquisition of LigD followed by Ku and 2) a direct transition from an *NHEJ*− to *NHEJ*+ state. We found that the LigD only state is more prevalent and is a prominent intermediate in the evolution of the *NHEJ*+ state in Proteobacteria and Bacteroidetes ($pp_{\text{Proteobacterial-subclade}}$ = 0.875 and $pp_{\text{Bacteroidetes}}$ = 0.7; supplementary fig. S10*C* and *D*). However, 90% of *LigD-only* genomes did not encode POL and PE domains, raising the possibility that *LigD-only* to *NHEJ*+ transitions could actually be *NHEJ*− to *NHEJ*+transitions.

Direct NHEJ gain can be explained by their acquisition via HGT. In this direction, we observed 60 *NHEJ*+ organisms coding this repair on their plasmids. However, our analysis based on the detection of base compositional differences revealed no horizontally acquired chromosomal NHEJ. These numbers may be

an underestimate because one is unlikely to pick HGT events if 1) the regions getting horizontally transferred have the same compositional biases in the donor and the recipient cells and/or 2) such HGT events occur early in bacterial evolution. Two instances of the latter case that we observe in our analysis are direct primary gains of NHEJ at the ancestral nodes of Firmicutes and Actinobacteria, with high posterior probability support ($pp_{Firmicutes}$ = 0.987 and $pp_{Actinobacteria}$ = 0.967; supplementary fig. S10*A* and *B*). (25) proposed a possibility of HGT between bacteria and archaea. In this direction, using a gene tree-species tree reconciliation approach, we observed evidence for HGT between bacteria and archaea. Furthermore, this method allowed us to find evidence of HGT between distantly related bacteria as well (supplementary tables 3 and 4). It should be noted that for the lack of methods to accurately date bacterial species trees and because transfers from unsampled species or extinct lineages could violate time constraints, we used an undated phylogeny for this analysis. Therefore, for each most frequent donor predicted, we noted the most frequent recipient as the potential donor–recipient pair. A similar search for NHEJ components in ~3,000 actinobacteriophages, yielded no hits, although, Pitcher et al. (2006)(99) have shown the employment of *Mycobacterium smegmatis* LigD by the Ku expressing mycobacteriophages Omega and Corndog for a successful infection. Thus, although our results suggest no transfer events between bacteria and bacteriophages, whether this is an artifact of sequence amelioration processes (100) remains to be tested.

A third route to an *NHEJ+* state could have been via a *Ku-only* state. However, we observed that the time spent in a *Ku-only* state is the least and NHEJ gains via this route are rare. Approximately 90% *Ku only* states are present in Firmicutes alone, specifically belonging to two genera, *Bacillus* and *Fictibacillus*. This suggests that *Ku only* state is largely avoided across most bacteria. This could be because Ku alone is nonfunctional in repair or in some cases could even block the access of break-ends for recombination-based repair (101–103). In 138 organisms, where Ku is retained in the absence of LigD, its function could be mediated by cross-talk with other ligases such as LigA. Consistent with this, it has been previously shown that if damage produces 3′ overhangs specifically, LigA could repair the lesion even in the

absence of LigB/C/D (58). In vitro studies have also shown that *Mycobacterium smegmatis* Ku can stimulate T4 DNA ligase (104). We observed the same trend in Archaea; Ku was always present with a LIG domain in ten archaeal species, whereas 230 archaea coded for LIG domain and no Ku. These observations are consistent with experimental studies reporting a fully functional complement of NHEJ being present only in a single archaeon, and with the likelihood of microhomology-mediated end joining (via ligase alone) being more prevalent in these organisms (88).

Furthermore, we observed that a primary gain does not stabilize the NHEJ trait in the subsequent lineages; with the most common type of loss being *NHEJ+* to *NHEJ–*. We discuss this, with examples, at two levels—1) across genera belonging to closely related phyla and 2) within the same genera. As discussed earlier, there was a direct primary gain of NHEJ at the common ancestor of the phyla Firmicutes and Tenericutes (fig. 4). First, we found that the monophyletic class Mollicutes (belonging to Tenericutes), which includes genera like the obligate plant parasites *Phytoplasma* and human parasites *Mycoplasma*, have had a one-step secondary loss of NHEJ during their evolution. This is consistent with the well supported hypothesis that these Mollicutes have evolved from Gram-positive bacteria including Firmicutes by reductive genome evolution (105–108). This suggests that NHEJ is either dispensable or costly to maintain a parasitic lifestyle in this class of bacteria. Second, although many species of the para- and poly-phyletic genus *Clostridium* (belonging to Firmicutes) including *Clostridium cellulosi* and *Clostridium stercorarium* retained NHEJ, others like *Clostridium clariflavum* and *Clostridium propionicum* had direct secondary losses. Furthermore, species like *Clostridium phytofermantans* and *Clostridium saccharolyticum* had direct secondary gains. A full list of NHEJ primary/secondary gains and losses at different taxa levels—direct and sequential—can be found in the supplementary table 8.

It is likely that several factors have contributed to the current distribution of NHEJ in bacterial systems. For example: 1) studies in both prokaryotes and eukaryotes have suggested a cross-talk between NHEJ and other repair mechanisms like recombination-based repair (29,95,96,109), base excision

repair (29), and mismatch repair (110). Shen et al. (2018) have shown that a bacteriophage infection is prevented by generating DSBs produced by MutL, a conventional mismatch repair endonuclease, which is subsequently repaired by NHEJ via a large deletion. However, this might not be a conserved pathway of repair as our analysis suggests an absence of NHEJ in a majority of actinomycetes belonging to the genus *Corynebacterium* including *Corynebacterium glutamicum* and 96% of archaea. 2) NHEJ could co-occur with other pathways that may or may not be directly involved in DSB repair. For example, NHEJ has been found to be active during sporulation in *Bacillus subtilis* (29,111), however, non-spore-forming bacteria also encode for NHEJ (86). Additional cross-talk at the level of regulation has also been reported, such as in *Mycobacterium*, where the deacetylase Sir2 has been implicated in regulating NHEJ (112). It is equally plausible that signatures of displacement of certain domains from genomes encoding NHEJ could inform us on how acquisition of NHEJ could have shaped genome architecture. In line with this, we did note 20 additional domains coded in Ku and LigD proteins (supplementary table 10) across different phyla. These would be useful to experimentally test in the future to understand whether they show functional interaction with NHEJ repair.

To understand the selection pressures that might play a role in shaping the evolutionary pattern of NHEJ, we focused our efforts on studying the genome characteristics associated with this DNA repair. We reasoned that under the event of a DSB, the unavailability of a template would prevent the highly accurate homologous recombination repair to occur (Fig. 3.9a, 3.9b). Thus, factors that affect rates of genome duplication or probability of multiple DSBs occurring could be central to determining the need for NHEJ-based repair. Therefore, one might expect a higher selection pressure of maintaining NHEJ in bacteria with slower GRs and larger GSs. In line with our hypothesis, we found that the organisms possessing conventional NHEJ tend to have significantly larger GS and slower GR as compared with those that are devoid of it. A phylogenetically controlled test suggested that this association held true for GS and not GR. We note that phyloANOVA used here for hypothesis testing assumes normality. However, our data for GS and rRNA copy numbers are not

normally distributed, even after converting them into a logarithmic scale (Shapiro–Wilk normality test; $P_{\text{genome size}} = 1.7 \times 10^{-10}$, $P_{\text{rRNA copy number}} < 10^{-15}$). Therefore, the result should be interpreted keeping this caveat in mind. Furthermore, ML and Bayesian inferences showed that there is a correlated evolution of NHEJ with GS and GR along the phylogenetic tree. Considering the two variables together, we found that GS is the better correlate of the *NHEJ+* state. Therefore, we conclude that the *NHEJ+* state is strongly associated with GS and to a much smaller extent with GR throughout its evolution.

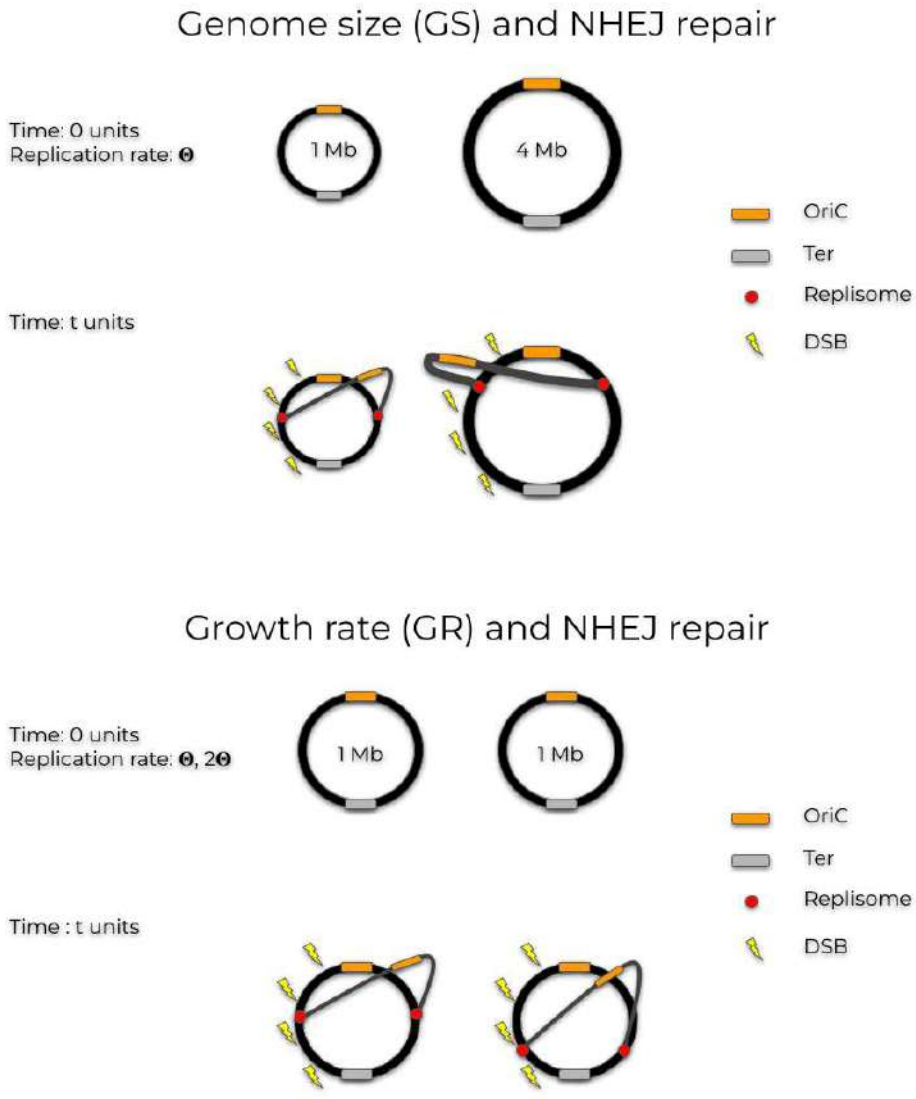


Figure 3.9 Higher selection pressure of maintaining NHEJ in organisms with larger genome size (upper panel) and slower growth rate (lower panel)

In sum, our study highlights the evolutionary trajectory of NHEJ and central characteristics that may have determined its sporadic distribution. DSB repair, including NHEJ, has been implicated in shaping bacterial genomes through mutagenesis (32), HGT (113), and their effect on genomic G–C content (93). Given this relationship between repair and genome evolution, it is important to ask how one factor may have influenced the other during bacterial evolution. It is also possible that similar forces have played a role in the evolution of other repair pathways and the genomes encoding them.

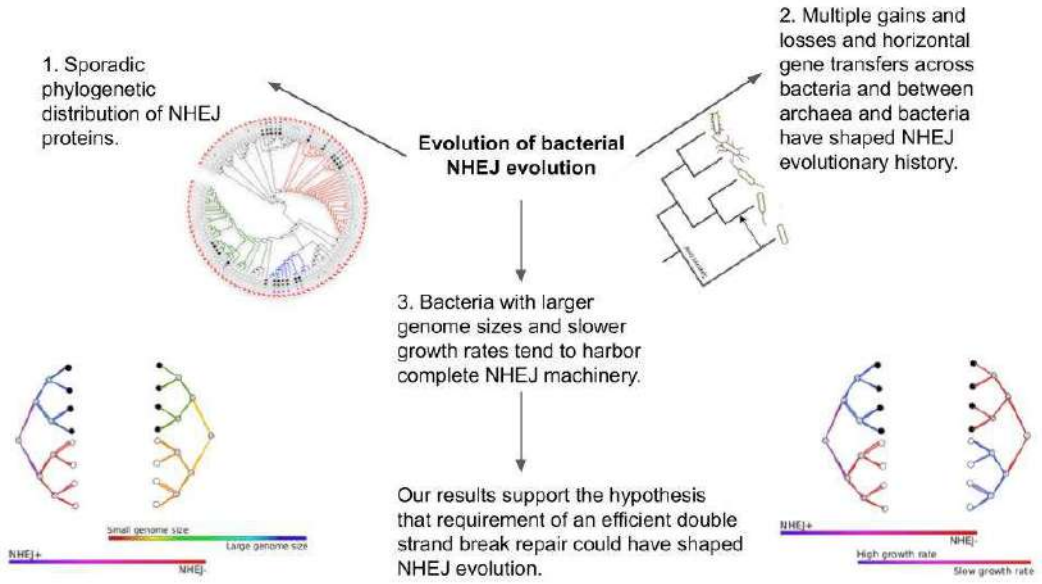


Figure 3.10 Summary of the evolution of NHEJ repair pathway in prokaryotes

Chapter 4: Evolutionary and predictive analysis of *alkB* prevalence in bacterial trait space

The manuscript for this work is under preparation and when published will have the following authors:

Mohak Sharda[1,2], Aditya Kamat[1], Inder Singh[1], Anjana Badrinarayanan[1], Aswin Sai Narain Seshasayee[1]

1 National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore - 560065, India.
2 School of Life Science, The University of Trans-Disciplinary Health Sciences and Technology (TDU), Bangalore, 560064, India

4.1 Abstract

DNA experiences atypical methylation patterns under alkylation stress. These lesions could be innocuous, mutagenic or cytotoxic in nature. Despite growing efforts towards mechanistic characterization of various bacterial cellular responses to alkylation damage, the underlying drivers of these responses remain largely unknown and predicting which bacteria will harbor which cellular response remains a challenge. In this direction we tried to understand the evolution of *alkb*, an oxidative demethylase involved in direct reversal repair, which presents an interesting case study. Although *alkB* exclusively targets specific mutagenic and cytotoxic lesions like 1meA, 3meC and other adducts bulkier than the methyl group, recent studies however, based on a small number of genomes, have shown that certain bacteria lack it. We asked two questions. 1) How pervasive is *alkB* incidence among bacteria? Using comparative genomics on ~6000 bacterial genomes, we confirm that *alkB* is sporadically distributed across the phylogeny, with only one-third species coding for it. Firmicutes completely lack *alkB* and it is the least conserved alkylation damage repair gene in all other major clades, with a non-significant phylogenetic signal. 2) What factors favor the maintenance of *alkB*? Using comparative phylogenetic methods, metagenomics and machine learning approaches, we show that oxygen requirement is a major, but not the only, habitat predictor of *alkb* prevalence. Taken together, we find evidence supporting the hypothesis that sources of genome instability and therefore the requirement for an efficient repair could have dictated *alkb* evolution.

4.2 Introduction

DNA methylation is a major mechanism of epigenetic regulation in bacteria, where other known epigenetic players like histones and nucleosomes are absent (114). Under alkylation stress, caused by various endogenous (115,116) and exogenous agents (117), DNA can experience atypical methylation patterns. The estimated damage, varies from innocuous lesions like 7meG, to less frequent but highly mutagenic and cytotoxic lesions like 3meA, 1meA and 3meC (117,118). Studies in *E. coli* have shown that two broad strategies are employed in specifically removing alkyl radicals from DNA- i) constitutive expression of repair proteins, like *ogt* (O^6-meG and O^4meT DNA methyltransferase) and *tag* (N^3-meA-DNA glycosylase), that are limited in the substrates they can repair, and ii) an ensemble of proteins forming an inducible adaptive response repair with broader target range of alkyl substrates (33,119–122). As described in *E. coli*, the adaptive response comprises four genes organized in the form of a discontiguous regulon: *ada, alkA, alkB and aidB*, where the respective promoters are induced by the activated (methylated) form of Ada (122–130). Additionally, other DNA repair and tolerance mechanisms like the SOS response also contribute in bypassing alkylation-dependent insults (131,132). In spite of an increasing in-depth understanding of how these mechanisms work, we lack a proper understanding of the factors that cause selection to favor one repair strategy over the other.

Here, we focus on *alkB*, which presents an interesting case study. It is a direct reversal demethylase (Fig. 4.1) that employs specific cofactors - oxygen, alpha-ketoglutarate and non-heme Fe^{2+} - to function (133,134). It repairs mutagenic and cytotoxic lesions primarily at two positions - 1meA and 3meC, which are involved in hydrogen bonding in a double stranded DNA; and it has been shown to do so preferentially when DNA is in a single stranded form (134). Other specific substrates that this protein has been reported to repair are N^1meG, N^3meT, other adducts bulkier than methyl like ethyl, propyl and hydroxyalkyl groups, exocyclic and ethano adducts (134). Recent studies have also suggested regulatory roles of AlkB, one particular case being its ability to demethylate the regulatory 5meC mark (28,135). Given this importance, *alkB* is known to be conserved from bacteria to humans. Surprisingly, *alkB* is absent in

certain bacteria like *Deinococcus radiodurans*, *Acinetobacter baumanii* and *Campylobacter jejunii*, suggesting a low phylogenetic signal (136). Additionally, unlike Ada and AlkA that have respective constitutive proteins Ogt and Tag in *E. coli*, there is no such known protein corresponding to AlkB. This begs the question, given its importance, how do bacteria lacking *alkB* deal with respective lesions?

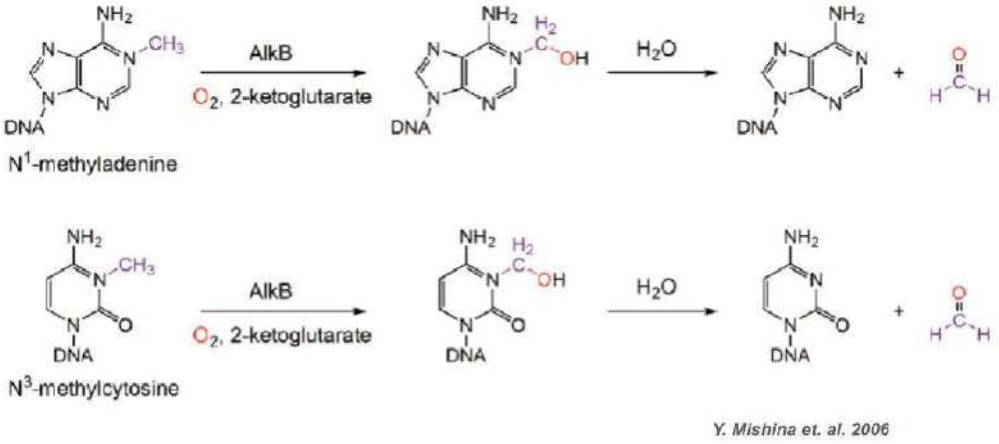


Figure 4.1 Mechanism of action of alkB in bacteria.

To the best of our knowledge, no comprehensive exploration of the factors that favour the maintenance of *alkB* has been made. What mechanisms might shape the distribution of *alkB* across bacteria? In this direction, we made the first attempt to identify different aspects of microbial habitats and lifestyles associated with *alkB* prevalence. In this study, we analyzed *alkB* evolution in three ways: (1) studying the distribution of *alkB* in an expanded set of bacterial genomes, (2) expanding the number of environmental and lifestyle traits considered as predictors using machine learning and metagenomics approaches and combination of large microbial trait databases like IMG/JGI and KEGG, and (3) incorporating appropriate statistical corrections for nonindependence among taxa due to shared evolutionary history as a confounding factor.

4.3 Materials and Methods

4.3.1 Genome data

All “complete” and “latest” (assembly_summary.txt; as of January 2017) genome information files for ~6,000 bacteria were downloaded from the NCBI ftp website using in-house scripts—whole genome sequences (.fna), protein coding nucleotide sequences (.fna), RNA sequences (.fna), and protein sequences (.faa). All the organisms were assigned respective phylum and subphylum based on the KEGG classification (https://www.genome.jp/kegg/genome.html; as of May 2018).

4.3.2 Identification of Adaptive response pathway homologues

Initial BlastP was run using Ada, AlkB and COG00745 protein sequences from *Escherichia coli MG1655* as the query sequence against the UniProtKB database (UniProt Consortium 2019) with an *E*-value cutoff of 0.0001. The top 250 full-length sequence hits were downloaded from UniProt. A multiple sequence alignment (MSA) was made using phmmer -A option with the top 250 hits as the sequence database and *E. coli* domain sequence as the query. A hmm profile was built using the hmmbuild command for the MSA obtained in the previous step. To find domain homologs, the command hmmsearch with an *E*-value cutoff of 0.0001 was used with the hmm profile as the query against a database of 5,973 bacterial genome sequences (supplementary table 1). These homolog searches were done using HMMER package v3.3 (137). An in-house python script was written to extract the results and assign organisms with protein sequences included in the study. Bacteria were assigned to two categories, based on the status of AlkB — *AlkB*− and ***AlkB*+**. Similar assignments were made for Ada and COG00745.

4.3.3 Identification of Ribosomal Ribonucleic Acid Sequences

A 16S ribosomal ribonucleic acid (rRNA) sequence database was downloaded from the Genomic-based 16S ribosomal RNA database (GRD) website (https://metasystems.riken.jp/grd/; last accessed November 2, 2020). An MSA and a profile of the database were made using muscle v3.8.31 (67) and hmmbuild

(137), respectively. To detect 16S rRNA homologs in our database of 5,973 bacteria, nhmmer (68) was used with an *E*-value cutoff of 0.0001 where the GRD-based 16S rRNA sequence profile was used as the query. In-house python script was written to parse the output files and select the best hits for further analysis.

4.3.4 Phylogenetic Tree Construction

For the construction of a species tree, one 16S rRNA sequence per genome was extracted into a multi-fasta file using an in-house python script. For bacteria with multiple 16S rRNA sequences, one 16S rRNA sequence was chosen such that it minimizes the number of *N*s in that sequence and has the maximum sequence length. In order to build a pruned phylogenetic tree, bacteria were randomly selected such that a genus was represented exactly once for every AlkB state. An MSA was built using musclev3.8.31 (67) with default options. The conserved regions relevant for phylogenetic inference were extracted from the MSA using BMGE v1.12 (76). Using IQTREE v1.6.5 (77), a maximum-likelihood (ML)-based phylogenetic tree was built. ModelFinder (-m MF option) (78) was used to choose the best model for the tree construction compared against 285 other models using the Bayesian Information Criterion (BIC). Branch supports were assessed using both 1,000 ultrafast bootstrap approximations (-bb 1000 -bnni option) (79) and SH-like approximate likelihood ratio (LR) test (-alrt 1000 option) (80).

4.3.5 Detection of methyltransferases

The protein sequences of 4-methylcytosine, 5-methylcytosine and 6-methyladenine methyltransferases, the respective methyltransferase target motifs and organisms habouring them, were obtained from REBASE. These were segregated into Type I, II, III and IV methyltransferases. To identify the protein sequences in our dataset of ~6000 organisms, phmmer was carried out with E-value cut off of 10^{-100}. Two matrices were obtained - 1) presence/absence and 2) number of methyltransferases of a given type per organism.

4.3.6 Trait data mining

An in-house python script was written to scrape the data from IMG/JGI website for 47 phenotypes for 71,000 bacterial genomes (.json). The phenotypes included:

1) auxotroph/prototroph information for all two amino acids, selenocysteine, biotin, coenzymeA, 2) Ability to utilize different carbon sources (maltose, sucrose, trehalose, glucose, fructose, galactose, xylose and L-arabinose), 3) chlorate reducer, 4) acetyl-CoA assimilator, 5) carbon fixation, 6) denitrifier, 7) ability to use nitrate as electron acceptor, 8) habitat, 9) oxygen requirement, 10) phenotype, 11) temperature range, 12) motility, 13) cell shape, 14) biotic relationship, 15) nitrogen fixer and 16) sulfur reducer.

An in-house python script was written to download the data of 183 KEGG metabolic pathways for 3387 organisms that were common with our dataset of ~6000 organisms using KEGG REST-API. Biopython module was used to download the data. An in-house python script was written to make the final presence/absence matrix of metabolic pathways for all organisms.

4.3.7 XGBoost classification pipeline

A random forest based Missing value imputation method employed in the python package MissingValuesHandler-1.1.6 was used to make informed guesses for different phenotypes with missing values. One hot encoding was used for feature engineering categorical variables. Stratified train/test split was used to ensure the same proportion of classes in both train and test sets. Cross validation was carried out using Grid search. Hyperparameter tuning was performed using the stratified KFold method with a scoring based on ROC-AUC. Hyperparameters tuned included ‘colsample_bylevel’, ‘colsample_bynode’, ‘gamma’, ‘learn_rate’, ‘learning_rate’, ‘max_delta_step’, ‘max_depth’, ‘min_child_weight’, ‘n_estimators’, ‘reg_alpha’, ‘reg_lambda’, ‘scale_pos_weight’ and ‘subsample’. The XGBClassifier with the objective function ‘binary:logistic’ and ‘gbtree’ booster was used to carry out classification with a validation metric of AUC-PR. Other metrics used were Matthew’s Correlation Coefficient (MCC), Precision and Recall, shown to provide better evaluation for imbalanced datasets like ours. All these steps were carried out using the *sklearn-1.0.1* python package. For a post-hoc explanation, TreeExplainer from the python package *shap-0.40.0*, based on Shapley values from game theory, was used.

4.3.8 Unsupervised learning analysis

Unsupervised learning was carried out on all microbial traits using Principal Component Analysis (PCA), t-distributed Stochastic Neighborhood Embedding (t-SNE) and Pairwise Controlled Manifold Approximation (PaCMAP). All these analyses were run using two python packages *sklearn-1.0.1* and *pacmap*.

4.3.9 Metagenome analysis

Normalized repair gene abundances as $\log_{10}$(FPKM) were calculated for each metagenomic community present in the TARA ocean dataset. This was compared with the metadata available for each community including the dissolved oxygen state, mean growth temperature, nitrite and nitrate content, using the Pearson's correlation coefficient along with the statistical significance calculation.

4.3.10 Detection of Phylogenetic signal

We used two statistics - a) the D statistics proposed by Fritz and Purvis (50). In order to standardize the effects of phylogeny size and the relative proportion of the two trait values, we simulated two null models (number of permutations = 1000): a) phylogenetic randomness and b) brownian threshold model. b) a phylogenetic analog of the Shannon entropy, delta (δ) (51). Since the D statistic assumes the discrete trait evolves under a Brownian motion threshold model, which might not be true for *alkB* evolution, we used another statistic, delta (δ). To test the statistical significance, we shuffled the *alkB* state values for 100 iterations and created a vector of random deltas that acted as our null hypothesis. To generate the p-value, we computed the probability *p(random_delta>deltaA)* in the null distribution.

4.3.11 Ancestral state reconstruction analyses

To trace the evolutionary history of alkB, two discrete character states were defined as follows: *alkB*+ and *alkB*-. Similarly, to trace the evolutionary history of oxygen requirement, two states were defined as follows: *anaerobe* and *non-anaerobe*. Non-anaerobe states included aerobes and facultative organisms. States were estimated at each node using stochastic character mapping (52) with 1,000 simulations provided by the make.simmap() method in the R package

phytools v0.6-44 (81). The phylogenetic tree was rooted using the midpoint method and polytomies were removed by assigning very small branch lengths (10^{-6}) to all the branches with zero length. The prior distributions of the states were set to *anaerobes* and *alkB-.* Further, by default the method assumes that the transitions between different character states occur at equal rates. This might not always be true, especially with complex traits where it is supposedly easier to lose than gain such characters. Therefore, for the estimation of transition matrix *Q*, three discrete character evolution model fits were compared: Equal Rates (ER), Symmetric (SYM), and All Rates Different (ARD). This allowed for models that incorporate asymmetries in transition rates. Based on Akaike Information Criterion (AIC) weights, the ARD model ($w\text{-}AIC_{ARD}$ =1, $w\text{-}AIC_{SYM}$ = 0, and $w\text{-}AIC_{ER}$ = 0) was chosen as the best fit with unequal forward and backward rates for each character state transition. Finally, *Q* was sampled 1,000 times from the posterior probability distribution of *Q* using Markov chain Monte Carlo and 1,000 stochastic maps were simulated conditioned on each sampled value of *Q*.

4.3.12 Correlated evolution analyses

The relationship between *alkB* and oxygen requirement was quantified using a statistical framework. To test if changes in *alkB* occur independently of oxygen requirements or whether these changes are more (or less) likely to occur in lineages with (or without) a given oxygen requirement, two models of evolution were considered—independent and dependent. In the independent model, both the traits were allowed to evolve separately on a phylogenetic tree, that is, non-correlated evolution. In the dependent model, the two traits were allowed to evolve in a non-agnostic manner, that is, correlated evolution. The *alkB* repair trait had two repair character states—*alkB-* (0) and *alkB+* (1). Similarly, the oxygen requirement trait had two states - anaerobe and non-anaerobe.

A continuous-time Markov model approach was used to investigate correlated evolution between *alkB* repair and oxygen requirement. An ML approach (53) was used to calculate log-likelihoods for the model of evolution. A LR statistic was calculated for both comparisons, followed by a chi-square test to assess if the dependent model was a better fit. The degrees of freedom are given by $df_{\text{chi-square test}} = (n_{\text{rate-dependent model}} - n_{\text{rate-independent model}})$. There are eight transition rates in the

dependent model across four states (00,01,10,11) and four transition rates in the independent model across two states (0,1; 0,1). Therefore, the test was run with four degrees of freedom.

A Bayesian approach was also used to test for correlated evolution based on RJ-MCMC (54). Different hypotheses were tested by comparing different transition rates under the dependent model of evolution.

4.4 Results

4.4.1 Differential conservation of *alkB* across bacteria

To identify the genes involved in the adaptive response pathway, we searched the *Escherichia coli* protein sequences as queries against the UniprotKb database. We searched the resulting HMM profiles against a database of ~6000 bacterial genomes (see Methods).

We found that *ada* was the most abundant adaptive response gene present in ~95% of the genomes, non-redundant at the species (n = 1992 of 2146 bacteria; **Figure 4.2a**) and genera (n = 972 of 1041 bacteria; **Figure 4.2b**) level respectively. It was present either alone (~30% of the genomes, n=591 species) or in combination with other adaptive response genes, *alkB* and *alkA*. Next, *alkB* was present in 35% (n=749 species) genomes. Both *alkB* and *ada* were present in 35% genomes, while only 4 genomes were coded for *alkB* alone. Next, we checked for the phylogenetic signal of *alkB* state to quantitate the effect of shared ancestry on its evolution. We used two statistics (see Methods) - a) the D statistics proposed by Fritz and Purvis and b) a phylogenetic analog of the Shannon entropy, delta (δ). We found a D = 0.06 (**Figure 4.2c**, black vertical line; p-value$_{\text{Brownian phylogenetic structure}}$ = 0.243, blue distribution; p-value$_{\text{random phylogenetic structure}}$ = 0, red distribution) and δ = 9.267 (**Figure 4.2d**, red line over the boxplot, *p-value=0.88*), suggesting that there is no significant evidence of a phylogenetic signal for *alkB* state. This is consistent with a previous study extensively carried out in the genus Pseudomonas that showed a variable distribution of *alkB* at the strain level (138).

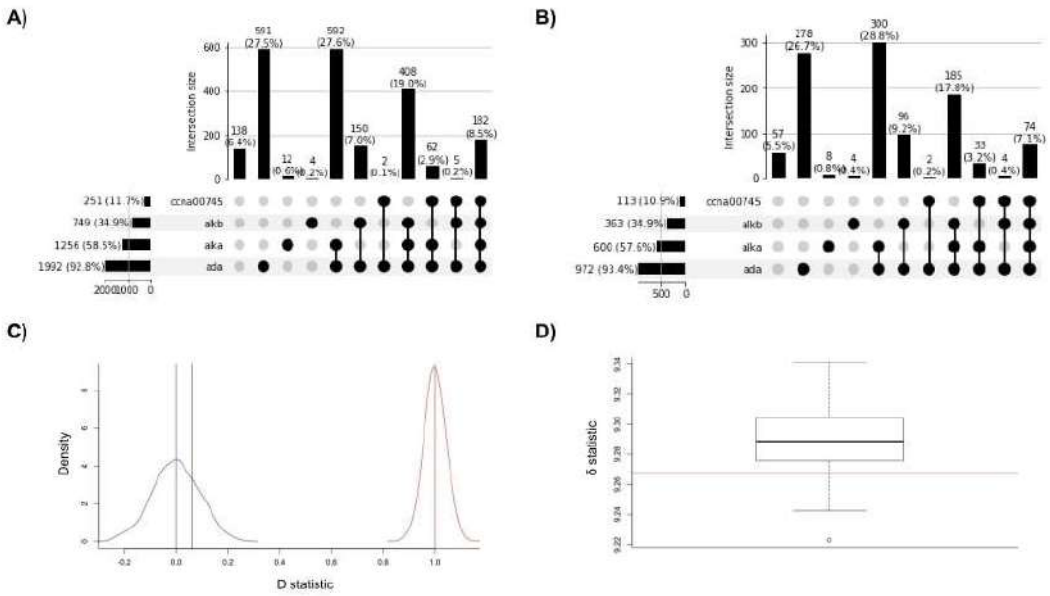


Figure 4.2 Conservation of alkylation damage repair genes across bacteria A) Upset plot showing the presence/absence status of genes involved in alkylation damage repair, non-redundant at the species level. B) Upset plot showing the presence/absence status of genes involved in alkylation damage repair, non-redundant at the genus level. C) D statistics showing a non-significant phylogenetic signal of *alkB*. D) Delta statistics showing a non-significant phylogenetic signal of *alkB*.

COG3826 is a newly annotated family of *alkB*-like genes implicated in DNA repair based upon an *in silico* study (Mello and Rigden, 2012). So far, no further studies have explored the role of this family to repair DNA alkylation damage. In this direction, we looked at COG3826 (*ccna_00745*) distribution among bacteria. COG3826 was present in a little over 10% (n=251) genomes. ~10% (n=187 species) genomes coded for both alkB and COG3826 while only 3% of the genomes (n=64 species) coded for COG3826 and not *alkB*. ~25% of the genomes (n=562 species) encoded alkB and not COG3826. At the phylum level, checked for major clades - Proteobacteria (**Figure 4.3a**), Firmicutes (**Figure 4.3b**), Actinobacteria (**Figure 4.3c**) and Bacteroidetes (**Figure 4.3d**) - we found that *ada* was present alone in the majority of the genomes when compared to other adaptive response genes. No organisms belonging to Firmicutes coded for *alkB*, while 12 species (3%) coded for COG3826 with *ada*. In the rest of the clades, organisms coding for *alkB* are more prevalent than COG3826.

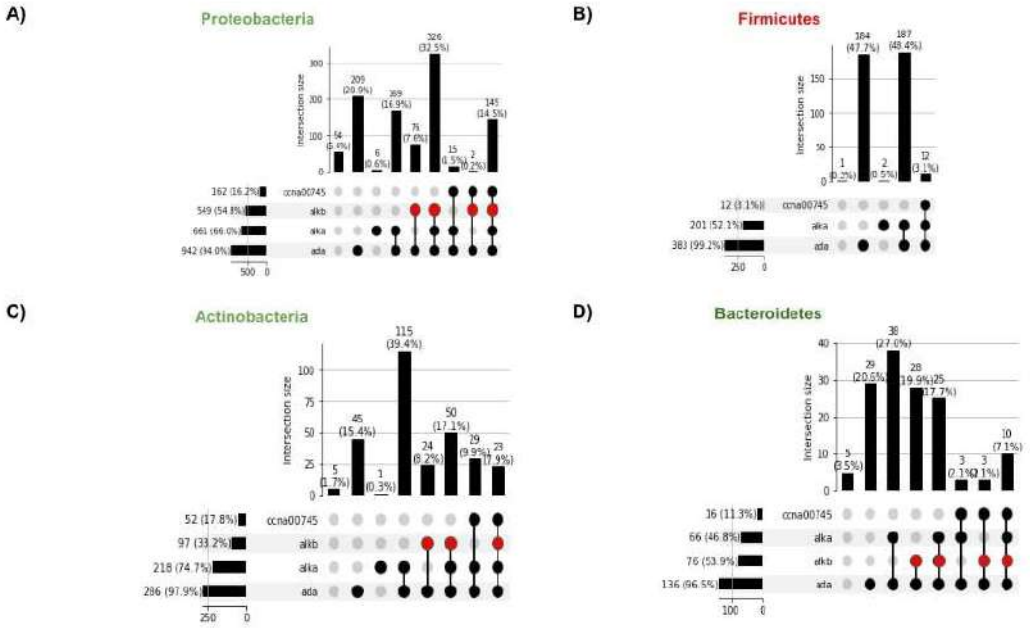


Figure 4.3 Comparison of conservation status of alkylation damage repair genes across clades A) Proteobacteria B) Firmicutes C) Actinobacteria D) Bacteroidetes. The red circles within the upset plot highlight *alkB* presence.

The presence of an *alkB* homologue (*ccna_00009*) and a COG3826 homologue in the model organism *Caulobacter crescentus*, provided us with an opportunity to study this novel family. We observed a growth defect when serial dilutions of cells lacking either genes were exposed to the methylating agent, methyl methanesulfonate (MMS) (**Figure 4.4**). This confirmed the essentiality of these genes under methylation damage, hinting at a possible role in repair. This growth defect was not observed when the cells were exposed to mitomycin C, a DNA alkylating agent that does not evoke the adaptive response. This illustrated the methylation-specific nature of the essentiality of these genes. Interestingly only the COG3826 deletion strain exhibited sensitivity towards another methylating agent, streptozotocin (STZ). STZ has a propensity to make O-adducts on DNA while MMS has a propensity to make N-adducts on DNA. Taken together, this suggests a possible difference in the repair substrates of *alkB* and COG3826.

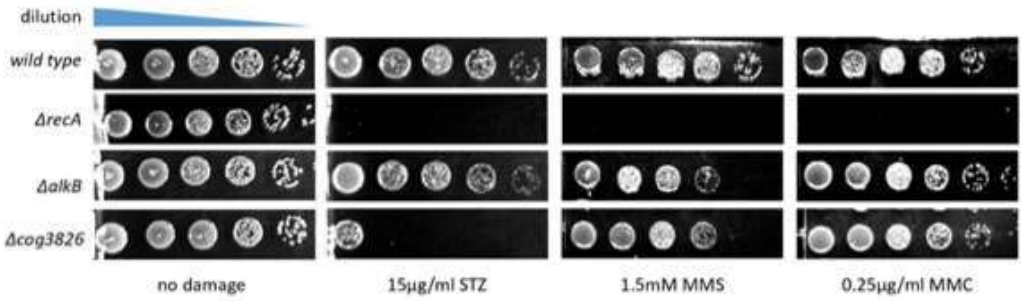


Figure 4.4 Caulobacter alkB and COG3826 deletion strains exhibit differential sensitivities upon exposure to different types of methylation damage. STZ: Streptozotocin; MMS: Methyl Methane Sulphonate; MMC: Mitomycin C Please note that this experiment was carried out by Aditya Kamat, Anjana lab. The figure was also prepared by him.

Despite the distribution patterns observed above, there were over 60% genomes that did not code for either *alkB* or COG3826. In this direction, studies have suggested the possibility of *alkA* repairing certain *alkB* substrates like 1meA and 3meC in a few experimentally studied organisms. We found that over 28% of the genomes (n=604 species) coded for *alkA* and neither *alkB* nor COG3826, raising the possibility of *alkA* repairing *alkB* substrates in these genomes (see Discussion). Finally, there were 6.4% (n=138 species) genomes that were coded for neither of the four alkylation damage repair genes tested here.

Taken together, we found that *alkB* is least conserved among all the known adaptive response genes with no evidence for a significant phylogenetic signal. We observed this trend to hold true even at the level of major clades tested here. The only other predicted *alkB*-like oxidative demethylase, COG3826, seems to be the least conserved gene overall with only 3% organisms coding for it but not *alkB*. Taken with our experimental results, this raises a possibility of a difference in the kinds of lesions both these proteins repair.

4.4.2 Fitness landscape of oxygen requirement and *alkB* suggests conditional synergism

Recent *in-vitro* studies have suggested that oxygen is a critical cofactor required for *alkB* function. We hypothesized that oxygen requirement could have dictated *alkB* maintenance. Towards understanding this, we grouped organisms included in our analysis, based on their ability to thrive under oxygen, into a binary trait - anaerobes and non-anaerobes, where the latter included both aerobes and facultative organisms.

We first checked for the phylogenetic signal of *oxygen requirement*, similar to the *alkB* state as described in the previous section. We found a D = 0.138 (**Figure 4.5a**, black vertical line; p-value$_{\text{Brownian phylogenetic structure}}$ = 0.06, blue distribution; p-value$_{\text{random phylogenetic structure}}$ = 0, red distribution) and δ = 17.02 (**Figure 4.5b**, red line over the boxplot; *p-value=0.06*), suggesting that there is no significant evidence of a phylogenetic signal for *oxygen requirement*.

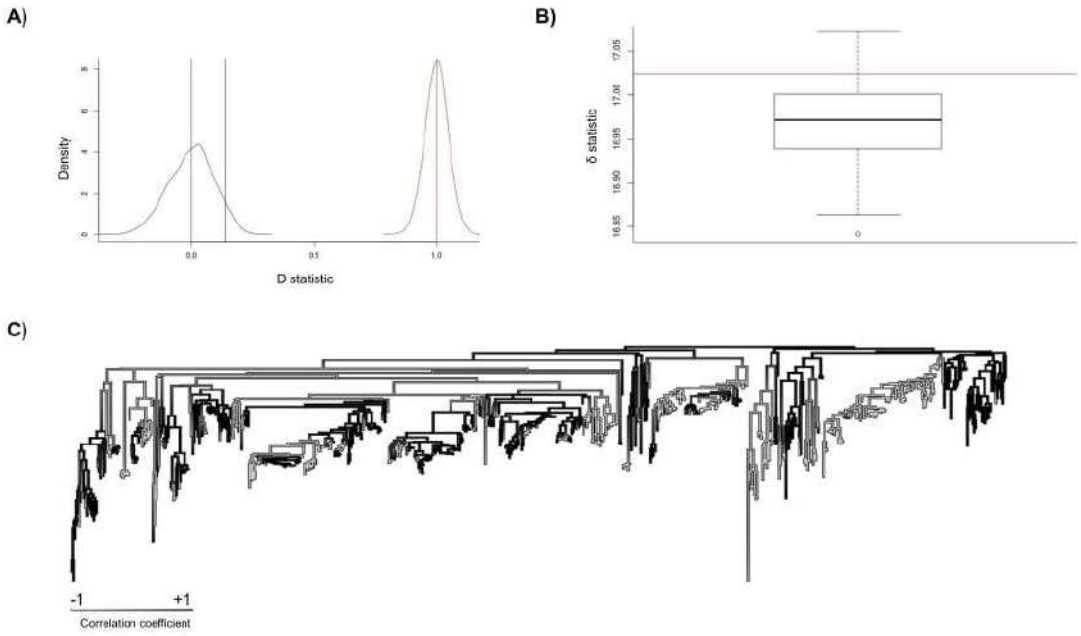


Figure 4.5 Correlated evolution of alkB with oxygen requirement A) D statistic showing a non-significant evidence for phylogenetic signal for oxygen requirement. B) Delta statistic showing a non-significant evidence for phylogenetic signal for oxygen requirement. C) Signal for correlated evolution of *alkB* and oxygen requirement. Black and white painted clades depict strong positive and negative correlation respectively. Gray clades indicate lack of any correlated evolution.

We next compared the evolutionary paths of both the trait states in order to test for signs of covariation. If there is an interaction between the evolution of the two traits, usually the joint changes in their paths will have a different impact on fitness than what one would expect by changes in either of their paths. All possible combinations of trait value pairs (00,01,10,11; 0 and 1 signify absence and presence of a trait respectively) can be mapped to their fitness values. This map is called a fitness landscape (Wright 1932, Yi and Dean 2019). Patterns of correlated evolution can help one understand the kind of fitness landscape two traits could be evolving in. For example, one of the strongest patterns of correlated evolution arises when changes in the two trait values together have a larger improvement in fitness than the sum of these changes for each trait

considered separately. This fitness landscape is said to be synergistic in nature. In this direction, we found evidence for correlated evolution between the two traits - *alkB* and *oxygen requirement*, in a phylogenetically controlled manner (**Figure 4.5c;** Marginal-likelihood$_{\text{Dependent model}}$ = -605, Marginal-likelihood$_{\text{Independent model}}$ = -805, log (Bayes Factor) = 400).

In support of correlated evolution, we further probed the eight different transition rates to test various evolutionary hypotheses. The eight transition rates were the following: a) q_{12} - gain of *alkB* in the absence of oxygen requirement, b) q_{13} - gain of oxygen requirement in the absence of *alkB*, c) q_{24} - gain of oxygen requirement in the presence of *alkB*, d) q_{34} - gain of *alkB* in the presence of oxygen requirement, e) q_{21} - loss of *alkB* in the absence of oxygen requirement, f) q_{31} - loss of oxygen requirement in the absence of *alkB*, g) q_{42} - loss of oxygen requirement in the presence of *alkB* and h) q_{43} - loss of *alkB* in the presence of oxygen requirement (**figure 4.6a**). We used *Z*-score, as described by *Pagel et. al. 2006*, to represent the posterior support for a zero transition rate. A Z-score ranges from 0% to 100%, indicating a decreasing uncertainty that the value of a given transition rate from one state to another is zero.

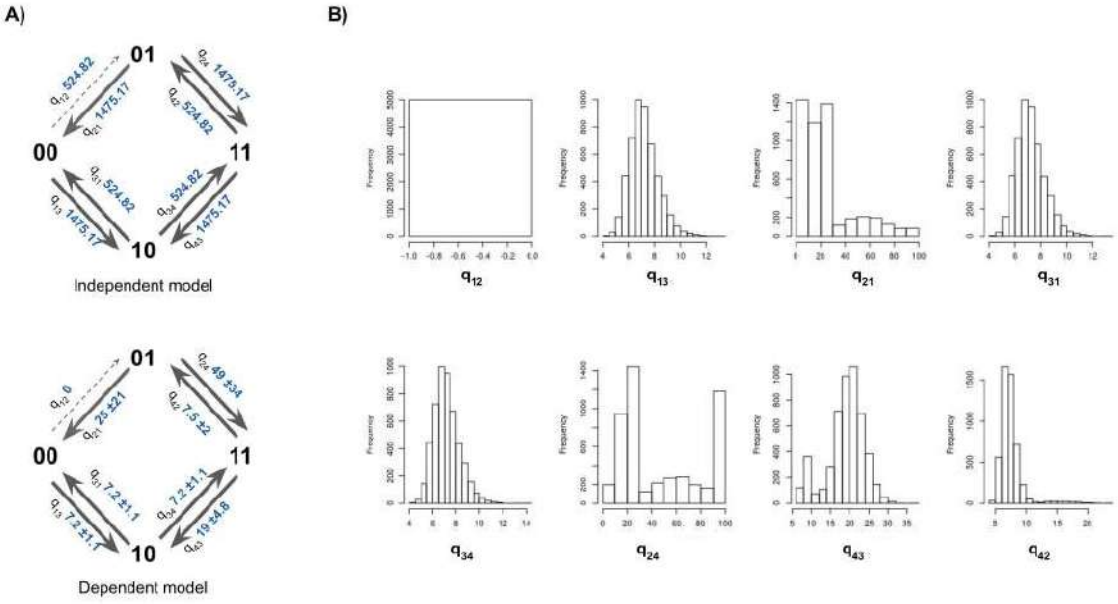


Figure 4.6 Transition rates suggest a change in oxygen requirement a pre-requisite for alkB gain but not maintenance A) State transition diagram depicting estimated four transition rates under an independent model of evolution (Left-above) and eight transition rates under a dependent model of evolution (Left-below). B) Bar plots depicting the rate classes traversed for each transition rate during the RJMCMC search. The transition rates are arranged in pairs q12-q34, q13-q24, q21-q43, q31-q42, such that a

difference in distribution shapes within each pair would contribute in favoring a dependent or correlated evolution model.

We found a significant pattern to changes in the two traits during the course of evolution: change in *anaerobe* to *non-anaerobe* state precedes *alkb*- to *alkb*+ transition, with a Z-score of 100% for the transition rate of *anaerobe:alkb*- to *anaerobe:alkb*+ (**Figure 4.6a**). Furthermore, we found that the *anaerobe:alkb*+ state is avoided, with q_{21} greater than q_{42} 67.38% of the time and is in the same rate class 32.2% of the time, and q_{24} greater than q_{42} 96% of the time and is in the same rate class 3.88% of the time (**Figure 4.6b**). In light of an adaptive interpretation, this suggests that the oxygen requirement changes first in evolution, and this selects a gain in *alkB*.

To test the degree of synergism between the traits, we compared two rate pairs: a) q_{34} is greater than q_{13} only 0.14% of the time and is in the same rate class 99.66% of the time and b) q_{34} is greater than q_{43} 0% of the time and is in the same rate class 10.88% of the time (**Figure 4.6b**). That Z-scores for $q_{13,}$ q_{43} and q_{34} are 0% (**Figure 4.6a**) indicate that both *non-anaerobe:alkb*- and *non-anaerobe:alkb*+ states are relatively stable with a tendency to lose *alkB* in a *non-anaerobe* background. Using stochastic character mapping, we observed that the time spent in the *non-anaerobe:alkb*+ state was only 20% (**Figure 4.7a, orange**). Infact, we found that the maximum time was spent in *anaerobe:alkb*- (red) and *non-anaerobe:alkb*- (blue), ~40% each, and the least time was spent in *anaerobe:alkb*+ state (cyan), ~0-1%, over the entire evolutionary history. A *non-anaerobe:alkb*- state could be achieved in two ways: a) from an *anaerobe:alkb*- state (q_{13}) as a virtue of shared evolutionary history with a lack of *alkb*- in the ancestor, or b) a secondary loss of *alkb* from a *non-anaerobe:alkb*+ state (q_{43}) (**Figure 4.7b**). We observed that q_{43} was greater than q_{13} with a posterior probability support of 89.1% and in the same rate class with a posterior support of 10.86% (**Figure 4.6b**), suggesting that *alkB* could be either costly or dispensable in these genomes.

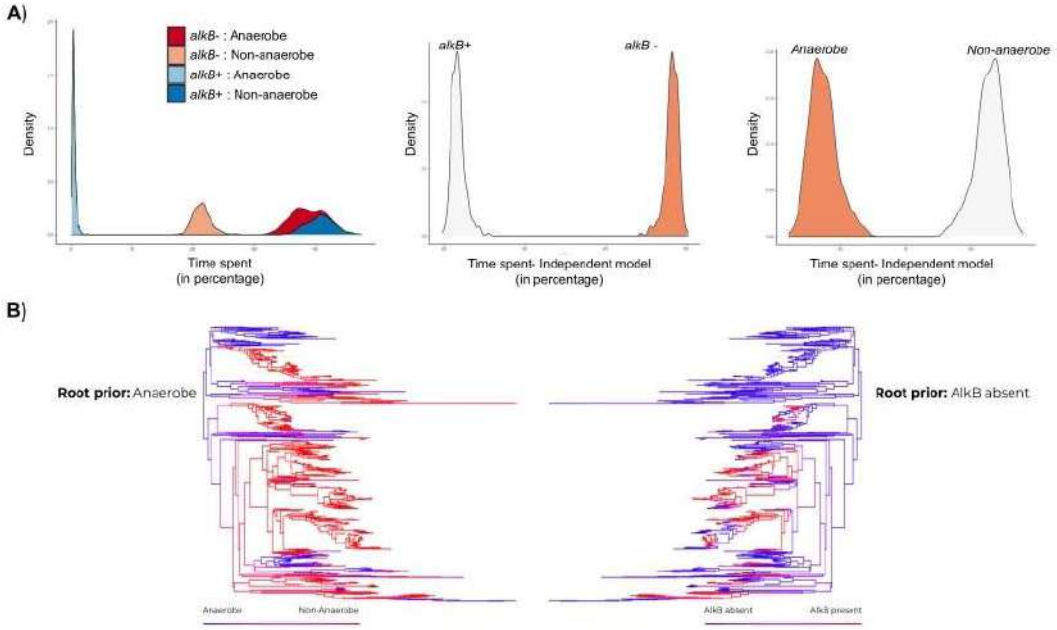


Figure 4.7 Ancestral state reconstruction of alkB and oxygen requirement using stochastic character mapping **A)** Time spent in each state in percentage across all the 1000 stochastic character maps analysed. ***Leftmost:*** The time spent calculated when both the traits alkB and oxygen requirement were evolved together over a phylogeny during the reconstruction of ancestral states. **Middle:** Time spent calculated when *alkb* state was evolved alone over a phylogeny during the reconstruction of its ancestral states. **Righmost:** Time spent calculated when *oxygen-requiremnt* state (Anaerobe versus non-anaerobe) was evolved alone over a phylogeny during the reconstruction of its ancestral states**. B)** Density plot showing the ancestral reconstruction of oxygen requirement (left tree) and alkB (right tree) states, juxtaposed opposite to each other. The prior for root node states were set to Anaerobe and *alkB* absent respectively during ancestral reconstruction.

Taken together, we found that oxygen requirement is a major driving force in *alkB* evolution with a gain in *non-anaerobe* state preceding a gain in *alkB*. That the maximum amount of time is spent in *non-anaerobe:alkb-* state with the rate of transition of *non-anaerobe:alkb+* to *non-anaerobe:alkb-* greater than the rate in the opposite direction indicates that additional selection pressures are required to maintain *alkB* in bacteria.

4.4.3 Linear and non-linear bacterial trait associations explain *alkB* prevalence

In order to understand the relative importance of different mechanisms that can explain *alkB* prevalence, we used different machine learning techniques to associate various microbial traits with *alkB* presence/absence. These microbial traits range from physiology (e.g heterotrophs/prototrophs), morphology (eg. motile/non-motile) to different methyltransferases (eg. 5-methyl cytosine methyltransferase) and metabolic pathways.

First, we tested the traits for their predictive potential to explain *alkB* incidence among bacteria. We used three unsupervised learning techniques that allow one to visualize high dimensional trait space in lower dimensions. We found that overlaying the information of *alkB* state on bacterial trait dimensional space places them in distinct clusters. Using Principal component analysis (PCA), we observed a separation of the two groups of bacteria, *alkB* harbouring and *alkB* lacking, along the first principal component axis (20% variation explained) (**Figure 4.8a**). There was no separation on the second (8% variation explained) and third (5.6% variation explained) principal components (**Figure 4.8b**). Top variable loadings along the first principal component axis can be attributed to different xenobiotic degradation pathways (including geraniol, caprolactam, chlorocyclohexane and chlorobenzene, toluene, aminobenzoate, novobiocin); biosynthesis of unsaturated fatty acids; synthesis of novobiocin, carbapenem, acarbose and validamycin; prototrophs for certain amino acids.

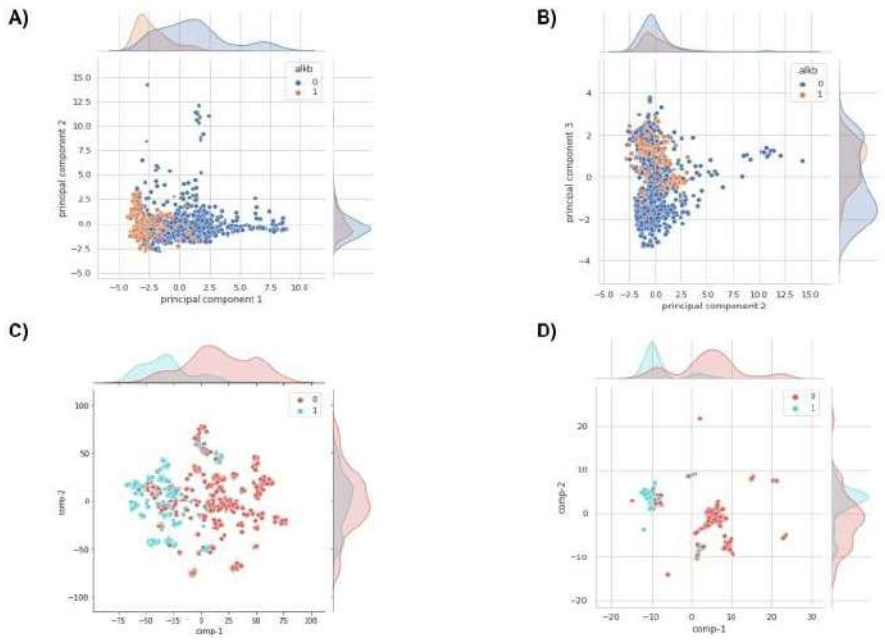


Figure 4.8 Visualising bacteria harboring and lacking alkB in microbial lifestyle and habitat trait space using different unsupervised learning algorithms A) PCA plot PC1 versus PC2 (*alkB* absent: blue and *alkB* present: orange), B) PCA plot PC2 versus PC3 (*alkB* absent: blue and *alkB* present: orange), C) t-SNE plot component 1 versus component 2 (*alkB* absent: red and *alkB* present: green), and D) PacMAP plot component 1 versus component 2 (*alkB* absent: red and *alkB* present: green).

PCA only accounts for global maximum variance and linear relationships that exist among bacterial traits. Therefore, we used two manifold learning visualization algorithms, t-SNE (**Figure 4.8c**) and PaCMAP (**Figure 4.8d**), to visualize *alkB*-

harbouring and *alkB*-lacking bacteria taking local and non-linear effects among traits into account. Both the methods revealed distinct clusters of *alkB*-harboring bacteria along the first component axis, further confirming that different ecological conditions play an important role in dictating *alkB* prevalence.

We superimposed COG3826 harboring and lacking organisms' information on the microbial trait space obtained above, for all the three techniques. We could not observe any difference in the two distributions either at the global or local level, suggesting no predictive potential to distinguish bacteria for COG3826 based on the traits included in our model. This further lends support to the results established so far and suggests a difference in the selection pressures of COG3826 and *alkB* (see Discussion).

The above techniques do not explicitly take into account the information of *alkB* states and therefore might miss out on some associations to better segregate *alkB*-harbouring organisms from those that lack it. However, they importantly established that *alkB* state is impacted by microbial traits, and that they hold a predictive potential. Furthermore, they highlighted that the relationships that exist among these microbial traits, segregating *alkB*-harbouring from *alkB*-lacking bacteria, are both linear and non-linear in nature.

4.4.4 Sources of genome instability dictate *alkB* prevalence

To better explore the trait space for their ability to predict *alkB* state, we used supervised learning that takes into account the additional information of *alkB* prevalence. As highlighted by t-SNE and PaCMAP in the previous section, there exist subtle local and non-linear associations among microbial traits. This guided our choice of the supervised algorithm, XGBoost, to capture these associations, along with linear effects, that explain *alkB* incidence. We built the model on the *training set* and made predictions on a separate set, called the *test set*; each containing an imbalanced proportion of 34% *alkB*-harbouring organisms (see Methods and Discussion). The model performance was assessed using appropriate classification metrics employed for imbalanced datasets such as ours. We confirmed the predictive power of our model with an Area-Under-Precision-Recall-Curve (AUCPR) of 91% and Matthew's Correlation Coefficient (MCC) of 0.7. The recall of our model was 0.88 and a precision of 0.73, indicating its

robustness against misclassification of *alkB*-harbouring organisms. The ability of the model to accurately predict *alkB* states of bacteria present in the test set indicated that microbial traits influence how *alkB* is distributed among bacteria.

We quantified the relative importance of each of the microbial traits in explaining *alkB* distribution. We took an approach where we added different independent features (microbial traits) in steps to our model. First, we incorporated the microbial physiological and morphological traits and habitats obtained from the IMG/JGI database. Using the post-hoc explainability method SHAP (SHapley Additive exPlanations; see methods), we found that *oxygen requirement* was among the top habitat predictors of *alkB* state (**Figure 4.9a**). This further confirmed the predictive reliability of our model. Oxygen requirement was split into three dummy states - 1) aerobicity was the top habitat predictor, with aerobes having a higher SHAP value of harbouring *alkB* as compared to non-aerobes, 2) anaerobicity, where anaerobes had a lower SHAP value of harbouring *alkB* and 3) facultative, with a mixed signal. Next, we observed different physiological trait predictors, prototrophs/auxotrophs for amino acids, that were associated with *alkB* incidence. We found that the log-odd of harbouring *alkB* was lower in auxotrophs for certain amino acids like serine, cysteine, threonine, arginine, tryptophan and phenylalanine.

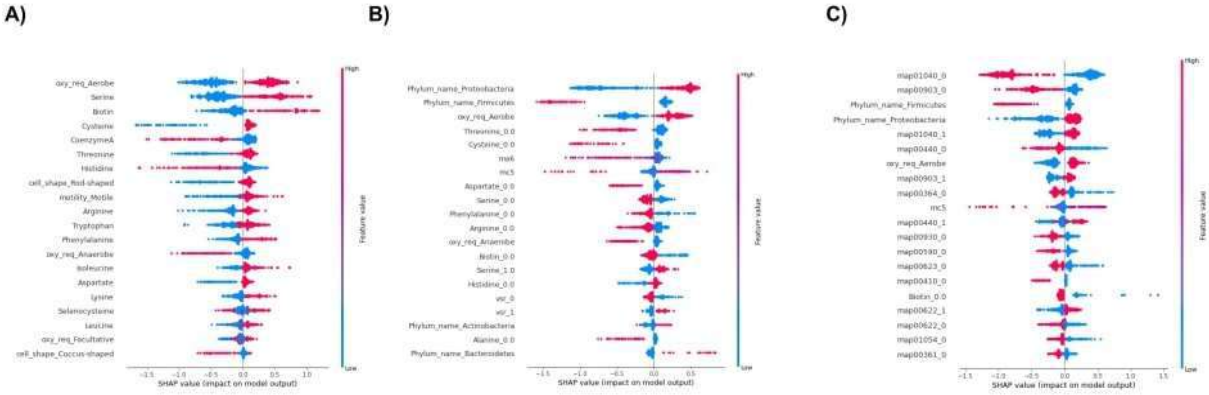


Figure 4.9 XGBoost based feature importance using SHAP values.

S. Rosic et. al. showed that DNA alkylation lesions in the genus Caenorhabditis, in particular the *alkB* substrate 3-methyl-cytosine, is a byproduct of 5meC DNA methyltransferase activity. They suggested a role of methyltransferases as endogenous agents for alkylation damage with a correlated evolution of *alkB* and methyltransferases in this genus. In bacteria, three major methyltransferase

activities are employed in gene regulation and innate immunity, namely 4meC, 5meC and 6meA. This raises the question whether the presence of methyltransferases plays a role in maintaining *alkB* in respective bacterial genomes. Therefore, we included the number of type II methyltransferases in our model. We found that 6meA and 5meC were among the top predictors of *alkB* state (**Figure 4.9b, Supplementary fig S11**). Bacteria coding for higher numbers of 6meA methyltransferases tend to lack *alkB* whereas there is a bimodality in terms of *alkB* presence in bacteria coding for higher numbers of 5meC methyltransferases (see Discussion). 4meC methyltransferases did not seem to have any predictive value for *alkB* state according to our model.

Towards understanding the metabolic pathways that *alkB* could be coevolving with, we added 183 pathways from the KEGG database into the model. We found that oxygen requirement was replaced by biosynthesis of unsaturated fatty acids and xenobiotic degradation pathways as the top predictors of *alkB* state (**Figure 4.9c,** see Discussion).

As an additional confirmation, we took advantage of the TARA ocean metagenomics dataset and observed an increase in *alkB* abundance with an increase in mean dissolved oxygen across different metagenomics communities ($r=0.4$, p-value < 0.01) (**Figure 4.10a**). Additionally, we took advantage of other environmental variables describing each metagenomics community. We found a strong negative correlation between temperature and *alkB* abundance ($r=-0.56$, p-value<0.01) (**Figure 4.10b**) and a weak correlation of *alkB* abundance with nitrite-nitrate concentration ($r=0.22$, p-value $= 0.008$), however the latter's distribution was broad (**Figure 4.10c**).

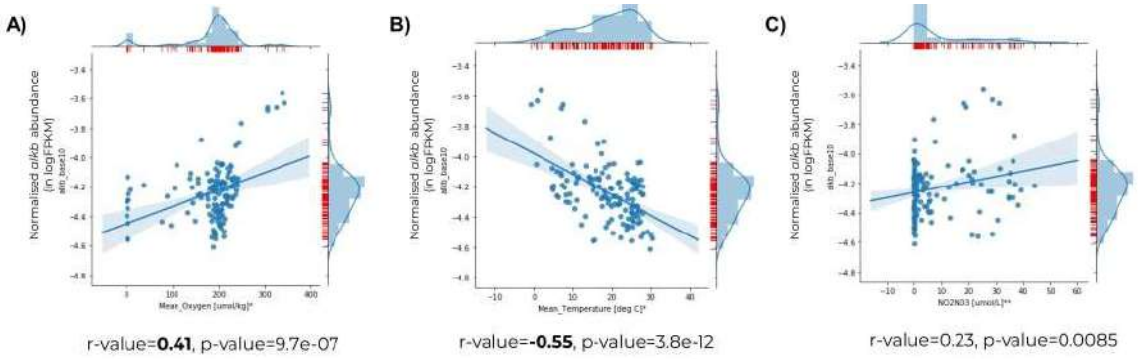


Figure 4.10 Scatterplots comparing normalized alkB abundance (Y-axis) in a given metagenomic community present in TARA ocean dataset against different community parameters (X-axes). A) Mean dissolved oxygen, B) Mean temperature and C) NO_2NO_3

Taken together, we found evidence for 1) oxygen requirement as one of the main selection pressures in *alkB* evolution, corroborated by phylogenetic, machine learning and metagenomic approaches, and 2) the role of different sources of genome instability in dictating *alkB* maintenance in bacterial genomes.

4.5 Discussion

In this study, we looked at ~6000 complete bacterial genomes to understand *alkB* prevalence. We found that ~34% bacteria code for *alkB*, with Firmicutes completely lacking it. That there is no significant phylogenetic signal associated with *alkB* suggests that environmental pressures play a more significant role in its maintenance than shared ancestry alone. It is possible that in organisms that lack *alkB*, there could be other mechanisms playing a substitutive role and repairing its substrates. In this direction, we looked at *alkA* prevalence, a DNA repair glycosylase, which has been shown to repair 3meC and 1meA lesions, but not other adducts heavier than the methyl group, in a few organisms like *Archaeoglobus fulgidus* (139) and *Deinococcus radiodurans* (140) but not in others like *Escherichia coli* (141). Around ~60% bacteria code for *alkA,* with only ~25% bacteria coding for *alkA* and not *alkB*. However, *Mielecki et. al. 2013* (138) in *P. putida* showed that AlkA can repair 3meC and 1meA lesions despite coding for a constitutively expressed AlkB. This suggests that AlkB could be important in repairing other alkyl lesions not repaired by AlkA. Furthermore, studies have shown that any base modification under alkylation stress also weakens the N-glycosyl bond and that exposure to huge amounts of alkylating agents lead to depurination/depyrimidination, giving rise to abasic sites (142). It could be possible then, in organisms that lack *alkB* and undergo high levels of alkylation stress frequently, that base excision repair is sufficient for repair.

Based on a multi-pronged bioinformatics approach, *Mello et. al. 2012* (143) predicted a new family of proteins - COG3826, and proposed that it could complement or replace AlkB function in the few genomes tested for their mutually exclusive distribution. However, the biochemical investigations related to COG3826 have not been carried out till date. In this direction, we started by looking at the distribution of *alkB* and COG3826 among our expanded set of bacteria. We found that only 3% organisms coded for COG3826 but not *alkB*, whereas ~26% organisms coded for *alkB* but not COG3826. There were 62% organisms that were coded for neither. Using spotting assays, we showed that COG3826 is important for survival under stress caused by the antibiotic streptozotocin but not *alkB*. However, cells under alkylation stress caused by

Methyl methanesulphonate (MMS) were sensitive only when they were double knockouts for *alkB* and COG3826 (data not shown here). This suggests a difference in the lesions that both the homologues repair, with a possible overlap of certain substrates (especially those caused by MMS) that cannot be resolved to a finer resolution in the scope of this study and therefore remains to be tested.

alkB requires three cofactors to function - alpha-ketoglutarate, non-heme Fe^{2+} ions and oxygen. *In vitro* studies have suggested that the former two cofactors are substitutable (134,144). We then asked if oxygen requirements shaped *alkB* evolution. In this direction, we found a correlated evolution of *alkB* and oxygen requirement. A number of studies have established that most likely the eubacterial ancestor thrived in the absence of oxygen (145–147), dictating our prior probabilities at the root to favour *anaerobic* and *alkB*- states. Given this, we observed that gain of *alkB* was always preceded by a change from *anaerobe* to *non-anaerobe* state throughout the entire evolutionary history. Furthermore, we observed that *alkB*+*:anaerobe* state was avoided with the least amount of time spent in it during bacterial evolution. Finally, we observed that relatively more amount of time was spent in a *non-anaerobe:alkB*- state (40%) as compared to *non-anaerobe:alkB*+ state (20%) and that there is a tendency of losing *alkB* during evolution in a *non-anaerobe* background. That there was no posterior support for zero rates for either combination indicated the stable nature of both the states. This suggests that even though oxygen requirement is a major driving force for *alkB* presence, its maintenance however depends on additional selection pressures.

Research on alkylation stress and the associated cellular responses have resulted in an extensive and detailed understanding of their biochemical and mechanistic nature (125,148–152). However, in a paradoxical sense, the nature of alkylation stress under natural conditions remains largely unknown. We still do not understand how prevalent stress is and how it varies across different environments, and if it occurs in bouts or has other factors dictating its nature. In this direction, although there have been advancements based on antimicrobial studies, that have helped us understand the variation in the severity of stress and the corresponding cellular responses (153), we argue that these stresses are only recent in nature, accelerated due to human interventions. That stress is an

evolutionary pressure that has existed since the beginning of life on earth, understanding that would help us in turn improve our understanding of how different DNA repair responses have evolved. This becomes important over long evolutionary timescales, where an absence of stress would result in a stochastic decay of cellular responses specific to it, subject to less effective selection.

In this direction, we took advantage of machine learning approaches combined with large databases on microbial habitats and lifestyles in order to understand the relative importance of mechanisms that contribute to stress and thereby could play a role in shaping *alkB* incidence among bacteria. Using unsupervised learning, we found evidence towards an association between microbial traits and *alkB* prevalence. Two kinds of approaches that are widely employed in machine learning studies were used - a) Principal Component analysis that preserve the global structure (bacteria distinctly dissimilar to each other) in the data, b) t-distributed Stochastic Neighbourhood Embedding that preserves the local structure (bacteria relatively similar to each other). Even though the latter can incorporate the non-linear component of the data, which the former approach cannot, it is subject to changes based on hyperparameter choice (**Supplementary figure S12**). We therefore, employed PaCMAP, that preserves both the local and the global structure and is robust to hyperparameter tuning. Among the top predictors were aerobicity, motility, prototrophicity for certain amino acids, methyltransferases involved in regulation, biosynthesis of unsaturated fatty acids and xenobiotic degradation pathways. We also tested for an association between COG3826 and microbial traits, and we found that with the present features available through the databases, there was no detectable predictive value. This suggests, however, that the selection pressures for *alkB* and its distant homologue COG3826 might be different.

To test associations in a robust manner, we took a supervised learning approach. The ability of the model to correctly predict *alkB* states was taken as a read-out for significant relationships between microbial traits and *alkB* prevalence. We predicted states on a test set. The set was chosen in such a way that the imbalanced proportion of *alkB* of 34% was maintained. We are aware that a test set should be independent of the set on which the model is trained and that this could be even more important in biological datasets where there is non-

independence among organisms that are evolutionarily close. However, since we were dealing with microbial trait data and we found no phylogenetic signal for *alkB*, we randomly segregated bacteria into the two sets for this analysis. Moreover, in genera *Pseudomonas*, studies have shown high variability in terms of *alkB* distribution even at the strain level (138), further justifying the approach incorporated by us.

We found that among different habitats tested, oxygen requirement, specifically aerobicity, was the top predictor. On incorporating other predictors in the model, we found that biosynthesis of unsaturated fatty acids replaced oxygen requirement as the top predictor. This lends support to the proposed idea in the field that lipid peroxidation might be a major contributor of alkylation stress produced intracellularly (154). Furthermore, our results suggest that it could, in fact, be a strong selection pressure acting on *alkB* maintenance. A number of xenobiotic degradation pathways that showed up in our model show a signal for multicollinearity with other well correlated pathways with *alkB*, suggesting a similar stress environment in organisms that harbour *alkB* (**Supplementary fig S13, Figure 4.8c**). Among other known significant alkylating agents produced intracellularly are the ones that are formed via nitrosations of amides, amines, amino acids, and peptides (116,155,156), especially in bacteria that occur in decaying matter, acidic soils, or putrid water. In this direction, we obtained certain physiological predictors like prototrophs for serine, cysteine, biotin, tryptophan, phenylalanine. Studies have shown that these molecules, in particular, have a higher propensity to undergo N-nitrosation reactions (157,158), raising the possibility of maintaining *alkB* in such genomes to reverse any resulting damage.

Another factor worth discussing includes methyltransferases that have been recently shown to act as endogenous alkylating agents in the genus Caenorhabditis, specifically the 5-methyl cytosine methyltransferases (28). In bacteria, methyltransferases are present in a huge diversity both in terms of the regulatory marks that they target and the kind of motifs they recognize, increasing the likelihood of acting as an endogenous alkylating agent. Therefore, we asked if there is a correlated evolution between methyltransferases and *alkB* in bacteria. We could not observe any differences in the distributions of the three kinds of type II methyltransferases - 4meC, 5meC and 6meA - in organisms that harbour *alkB*

and those that lack it (**Supplementary Fig S14**). Next, we reasoned that the presence of even one methyltransferase could be sufficient to cause lesions in a cell. In this direction, we found that organisms harbouring *alkB* are enriched for the presence of all three types of type II methyltransferases (Fisher test, P-value < 0.01), with a phi correlation coefficient of 0.25. We also repeated the analysis by taking into account the phylogeny, where we allowed the trait – presence/absence of methyltransferases, to evolve under a threshold model (this model assumes that the discrete trait is controlled by an underlying continuous trait evolving under a brownian motion). We found a correlation coefficient of 0.15 (**Supplementary fig S15**). A similar analysis between methyltransferases and COG3826 showed no significant enrichment and correlation. However, using the multivariate modeling approach that controls for other features in the model, we found that the number of 5meC and 6meA methyltransferases were among the top predictors albeit with a mixed signal. The electrostatic potential of each nitrogenous base is different (159) and studies have shown that it could be prone to lesions in a nonrandom fashion (160,161), especially where nitrogenous bases could affect the electrostatic potential of the neighbouring bases. It is therefore possible that methyltransferases that target a certain motif specificity could be more prone to causing lesions that are repaired by *alkB.* In organisms that have a very high number of methyltransferases and that lack *alkB,* either the lesions could be sufficiently repaired by other adaptive repair proteins or directly by base excision repair.

Finally, as a secondary confirmation for our results, we used the TARA ocean metagenomics dataset. This approach has an advantage that it allows one to analyze strategy shifts actually occurring in a changing environment. We found a moderate positive correlation between *alkB* abundance and dissolved oxygen levels among the microbial communities. Surprisingly, we also found a strong negative correlation between mean temperature of a given metagenomics community and *alkB* abundance. One possibility that could explain this association is the fact that at lower temperatures, psychrophiles are enriched with mechanisms involved in synthesis of unsaturated fatty acids, to help them survive in such extreme conditions (162–165). Since peroxidation of lipids, specifically unsaturated fatty acids, is one of the major contributors of endogenous alkylation

stress (163,166); maintaining *alkB* in such genomes could be important for survival.

We have tried to incorporate bacteria, as phylogenetically diverse as possible, in our analysis to associate *alkB* prevalence with ecological drivers. Though it is possible that with the current databases on microbial traits, we might have missed some important associations that could explain *alkB* distribution. For example, temperature was not among the top predictors in our supervised learning model. However, we picked up the association using the metagenomics approach. That could be explained by the fact that our dataset of complete bacterial genomes with habitat information was highly skewed for mesophiles and a severe underrepresentation of bacteria with other temperature profiles (**Supplementary figure S16**), thereby disallowing the model to learn this feature effectively.

In this study (Figure 4.11), based on a machine learning augmented approach, we have highlighted the possibility of maintenance of *alkB* driven majorly by sources of genome instability. Experimental studies would further help validate the mechanisms that we propose here, to better understand the role of microbial ecology in the evolution of *alkB*. Finally, this approach can be extended to understand the evolution of other repair pathways and the role different selection pressures play in shaping these repertoires in bacteria.

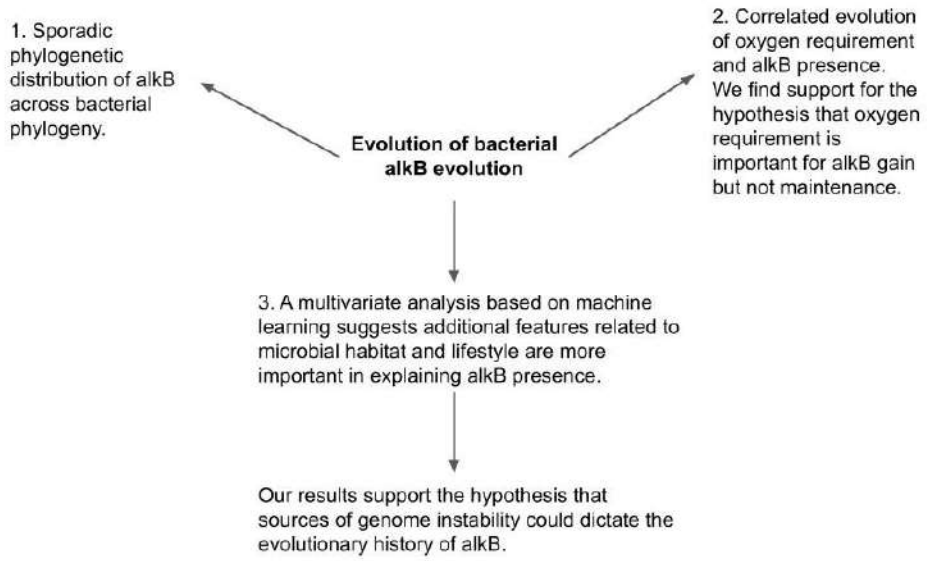


Figure 4.11 Summary of the findings on evolution of alkB in bacteria.

Chapter 5: General Discussion

All comparative and evolutionary genomics studies are dependent on the information encoded in a genome. With advances in sequencing technologies, including next-generation and third generation sequencing, the availability of genomic data is ever increasing, especially for relatively smaller genomes of prokaryotes (167). In addition to this, there have been improvements in the cost of sequencing a genome, accuracy and storage of the resulting data (168). This has opened up avenues of testing various long-standing hypotheses in the field that were not possible before the current upsurge.

The majority of the algorithms that were developed to carry out comparative and evolutionary studies exist from a time when the genomics era had not ushered in (47,53). In this direction, in the past few decades, there have been improvements in algorithm designs that can much better adopt current datasets (169). In fact, in parallel, there is an active ongoing research in furthering these improvements to better accommodate the ever increasing datasets, both in number and in kind. This is specifically important when one talks about biological data, where organisms are evolutionarily related to each other (48). This calls for incorporation of special statistical methods in the already existing analysis frameworks, that can take into account the non-independent nature of species. This has been discussed in length in Chapter 2 of this thesis. The general consensus in the field is to carry out these analyses in a phylogeny-unaware and phylogeny-aware manner to better deal with any false-positives that might skew the results.

It is worth discussing that the current methods are far from perfect. For example, the current DNA substitution matrix models used to construct phylogenies make a number of assumptions (discussed in Chapter 2) that might not be true for the 'real datasets'. There is a need to incorporate matrices that could model different aspects of evolution, some of them being: a) non-reversible nature or asymmetricity, b) non-stationarity i.e. the composition of nucleotides could change over time, c) non-independence of neighboring sites in an alignment, d) different rates of evolution in different branches of a tree, especially when one needs to make deep phylogenies. Due to the inherent complexity, although there have been efforts in improving these algorithms, they are still in their nascent stages and are not yet incorporated in the tools that are available at the

time of writing this thesis (170). Similarly, there is also a need for improvement in algorithms employed in phylogenetic comparative methods as well, especially when one wishes to employ these methods in inferring adaptation. Moreover, simulation studies would be helpful in understanding the amount and type of deviations of inferred results from the true underlying evolutionary structures, for any given methodology (171). Nevertheless, as some studies have reported in the past, current methods that incorporate phylogenies are better than carrying out an analysis without taking phylogeny into account (172,173). Needless to say, in the future, as the methodologies become more sophisticated, it remains to be seen how the evolutionary inferences made so far, including the work presented in this thesis, deviate.

The work presented in this thesis, in Chapter 3 and Chapter 4, has taken advantage of the already existing methods and incorporated them in building the pipelines to understand the evolution of DNA repair pathways. Much of the research in the field of DNA repair has focused on the mechanistic characterization of different repair pathways in model organisms. Despite these efforts, we still do not understand how, when and why these mechanisms evolved a certain way. Answers to these questions are important to understand the impact of DNA repair systems on genome evolution and vice versa.

In Chapter 3, we tried to understand the evolution of a double strand break repair pathway in prokaryotes, called the Non-Homologous End Joining repair. Previous studies that have attempted to understand its evolution have all been based on either a small number of genomes or have not gone beyond just cataloging the presence and absence of the proteins involved in this pathway (29). Our approach incorporated a domain wise search, to study the phylogenetic distribution of NHEJ. We confirmed, on an expanded set of organisms, the sporadic nature of this pathway in bacteria and archaea. Going beyond this, we also looked at how this pathway could have evolved. In this direction, we looked at the role of vertical gains and losses and horizontal transfers in shaping its evolution. A key finding from our analysis is that this pathway could have evolved independently with multiple gains and losses assuming a tree-like evolutionary model and a most likely absence of a functional NHEJ pathway at the eubacterial ancestor. We further found that

horizontal gene transfers had an important role in its evolution, not just among bacteria, but also between archaea and bacteria. Finally, in order to understand the selection pressures that could have played a role in NHEJ evolution, we looked at two genome characteristics- genome size and growth rate. We used correlated evolution analyses, Maximum Likelihood and Bayesian based, to detect a signal favoring a dependent model of evolution dictating genome size and NHEJ evolution, and growth rate and NHEJ evolution. A regression analysis, assuming a Markov process based evolutionary model, however, suggested that genome size alone is a stronger correlate for NHEJ evolution. Overall, it is worth highlighting the importance of our findings as they lend support to an already proposed hypothesis in the field - the requirement for an efficient double strand break repair could have played an important role in the evolution of NHEJ.

Few caveats are worth mentioning here:

1) We have taken the maximum likelihood based phylogeny that acted as our hypothesis of what the 'true' relationship among bacteria used in our analysis could have been. It remains to be seen in the future work(s) how robust our findings are when one takes phylogenetic uncertainty into account, using methods like Bayesian based phylogenies.

2) Another caveat is the conversion of continuous traits like genome size and growth rate into binary traits (small versus large genome, slow versus fast growth rate), prior to carrying out the correlated evolution analysis. For the lack of similar methods to study correlated evolution between a continuous trait and a discrete trait (presence and absence of NHEJ), we had to develop the approach mentioned in the previous line. We call out the readers highlighting the need to develop methods that can deal with mixed trait datasets, so that the issue of reduction in signal (continuous data has more signal than discrete data) can be circumvented.

3) Recent studies have reported the bias in the sequencing of prokaryotic genomes, sequencing those with a higher commercial value more than the others (174). It is possible that the inferences made in our study with

respect to the evolution of NHEJ repair could change, once the underrepresented and undersampled phyla (and species) are included in the analysis in the future; especially towards the internal nodes.

4) Because ours is an observational, rather than a case-control study, our findings related to NHEJ evolution and central genome characteristics, at best remain correlational in nature. Whether they really have any direct adaptive value, as we hypothesize, remains to be tested using experimental approaches. One way of testing this would be to use long term lab evolution experiments. This would help understand if these correlations are a result of *direct* or *indirect* selection (175).

In Chapter 4, we tried to understand the evolution of a reversible alkylation damage repair protein, an oxidative demethylase called AlkB. Despite its importance in exclusively repairing specific DNA lesions, we observed a sporadic distribution of this protein across bacteria. Because AlkB requires oxygen as one of the co-factors, we tested for a correlated evolution between the two traits. A key finding that came out of our analysis is that oxygen requirement is an important driver of AlkB evolution. We confirmed this finding using machine learning and metagenomic approaches as well. Another key finding that we report in our study is that there are additional factors that are required to maintain AlkB in bacteria, with genomes having a tendency to lose AlkB in the background of a non-anaerobic state during evolution. To better understand the selection pressures that could shape AlkB evolution, we used an exploratory approach incorporating machine learning algorithms to test various predictors belonging to different microbial habitats and lifestyles.

To the best of our knowledge, we have incorporated a machine learning (ML) approach to understand DNA repair evolution for the first time. We propose to extend this approach in carrying out similar evolutionary studies. A key advantage of this approach is that unlike regression methods like logistic regression, one can explore even non-linear and interaction effects of various independent variables, especially when one wants to use them as an exploratory analysis to generate testable hypotheses, in addition to testing existing ones. Furthermore, unlike phylogenetic logistic regression (PLR), this

approach does not assume an underlying evolutionary model. However, we urge algorithm designs like PLR that can incorporate more sophisticated machine learning models in a phylogeny aware context with different evolutionary model assumptions, something that does not exist as of today. In the meantime, when there is a strong phylogenetic signal in the data and if one wishes to carry out a phylogeny controlled analysis using an ML approach, there are certain propositions worth mentioning here, to circumvent the current lack of these methods: a) Using evolutionary distance based clustering methods to design training, test and validation set, in order to achieve independent sets, b) during KFold based hyperparameter tuning, to use the same clustering based methods to split the folds such that they are as independent of each other as possible. For example, one fold could consist of just Firmicutes, another of only Actinobacteria and so on. These methods, however, are not free from limitations, one of them being an underrepresentation of certain clades in the current databases storing sequencing data.

In conclusion, we have reported a framework that can be used to understand the evolution of other repair pathways. A key finding from both the case studies discussed in Chapter 3 and 4 lends support to the hypothesis that sources of genome instability might play a significant role in dictating DNA repair evolution. In the bigger scheme of discussion, this might explain why we see differences and similarities in DNA repair repertoires across organisms, not explained by phylogeny alone.

Chapter 6: References

1. Friedberg EC, Walker GC, Siede W, Wood RD. DNA Repair and Mutagenesis. American Society for Microbiology Press; 2005. 2845 p.

2. Friedberg EC. A brief history of the DNA repair field. Cell Res. 2008 Jan;18(1):3–7.

3. An early suggestion of DNA Repair. Effect os sublethal doses of monochromatic ultraviolet radiation on bacteria in liquid suspensions. By Alexander Hollaender and John T. Curtis. Proc Soc Exp Biol Med, 33,61-62(1935). Basic Life Sci. 1935;5A:xi–xii.

4. Witkin EM. Inherited Differences in Sensitivity to Radiation in Escherichia Coli. Proc Natl Acad Sci U S A. 1946 Mar;32(3):59–68.

5. Dulbecco R. Reactivation of ultra-violet-inactivated bacteriophage by visible light. Nature. 1949 Jun 18;163(4155):949.

6. Lindahl T. An N-glycosidase from Escherichia coli that releases free uracil from DNA containing deaminated cytosine residues. Proc Natl Acad Sci U S A. 1974 Sep;71(9):3649–53.

7. Radman M. SOS repair hypothesis: phenomenology of an inducible DNA repair which is accompanied by mutagenesis. Basic Life Sci. 1975;5A:355–67.

8. Witkin EM. Ultraviolet mutagenesis and inducible DNA repair in Escherichia coli. Bacteriol Rev. 1976 Dec;40(4):869–907.

9. Lu AL, Clark S, Modrich P. Methyl-directed repair of DNA base-pair mismatches in vitro. Proc Natl Acad Sci U S A. 1983 Aug;80(15):4639–43.

10. Lindahl T. My Journey to DNA Repair. Genomics, Proteomics & Bioinformatics. 2013 Feb 1;11(1):2–7.

11. Bauer NC, Corbett AH, Doetsch PW. The current state of eukaryotic DNA base damage and repair. Nucleic Acids Research. 2015 Dec 2;43(21):10083–101.

12. Marshall CJ, Santangelo TJ. Archaeal DNA Repair Mechanisms. Biomolecules. 2020 Oct 23;10(11):1472.

13. Kamarthapu V, Nudler E. Rethinking Transcription Coupled DNA Repair. Curr Opin Microbiol. 2015 Apr;24:15–20.

14. Mohseni-Salehi FS, Zare-Mirakabad F, Ghafouri-Fard S, Sadeghi M. The effect of stochasticity on repair of DNA double strand breaks throughout non-homologous end joining pathway. Math Med Biol. 2018 Dec 5;35(4):517–39.

15. Cortez D. Replication-Coupled DNA Repair. Mol Cell. 2019 Jun 6;74(5):866–76.

16. Modrich P. Mechanisms and Biological Effects of Mismatch Repair. Annual Review of Genetics. 1991;25(1):229–53.

17. Li YF, Kim ST, Sancar A. Evidence for lack of DNA photoreactivating enzyme in humans. Proc Natl Acad Sci U S A. 1993 May 15;90(10):4389–93.

18. Foster PL. Stress-Induced Mutagenesis in Bacteria. Crit Rev Biochem Mol Biol. 2007;42(5):373–97.

19. Battista JR. Against all odds: the survival strategies of Deinococcus radiodurans. Annu Rev Microbiol. 1997;51:203–24.

20. Slade D, Lindner AB, Paul G, Radman M. Recombination and replication in DNA repair of heavily irradiated Deinococcus radiodurans. Cell. 2009 Mar 20;136(6):1044–55.

21. LeClerc JE, Li B, Payne WL, Cebula TA. High mutation frequencies among Escherichia coli and Salmonella pathogens. Science. 1996 Nov 15;274(5290):1208–11.

22. Matic I, Radman M, Taddei F, Picard B, Doit C, Bingen E, et al. Highly variable mutation rates in commensal and pathogenic Escherichia coli. Science. 1997 Sep 19;277(5333):1833–4.

23. Modrich P, Lahue R. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu Rev Biochem. 1996;65:101–33.

24. Alhmoud JF, Woolley JF, Al Moustafa AE, Malki MI. DNA Damage/Repair Management in Cancers. Cancers (Basel). 2020 Apr 23;12(4):1050.

25. Aravind L, Walker DR, Koonin EV. Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. 1999 Mar 1;27(5):1223–42.

26. Eisen JA, Hanawalt PC. A phylogenomic study of DNA repair genes, proteins, and processes. Mutat Res. 1999 Dec 7;435(3):171–213.

27. Erill I, Campoy S, Barbé J. Aeons of distress: an evolutionary perspective on the bacterial SOS response. FEMS Microbiology Reviews. 2007 Nov 1;31(6):637–56.

28. Rošić S, Amouroux R, Requena CE, Gomes A, Emperle M, Beltran T, et al. Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. Nat Genet. 2018 Mar;50(3):452–9.

29. Bertrand C, Thibessard A, Bruand C, Lecointe F, Leblond P. Bacterial NHEJ: a never ending story. Molecular Microbiology. 2019;111(5):1139–51.

30. Hofstatter PG, Lahr DJG. Complex Evolution of the Mismatch Repair System in Eukaryotes is Illuminated by Novel Archaeal Genomes. J Mol Evol. 2021 Feb 1;89(1):12–8.

31. Crick F. The double helix: a personal view. Nature. 1974 Apr 26;248(5451):766–9.

32. Paris Ü, Mikkel K, Tavita K, Saumaa S, Teras R, Kivisaar M. NHEJ enzymes LigD and Ku participate in stationary-phase mutagenesis in Pseudomonas putida. DNA Repair (Amst). 2015 Jul;31:11–8.

33. Friedberg EC. Out of the shadows and into the light: the emergence of DNA repair. Trends Biochem Sci. 1995 Oct;20(10):381.

34. Woese CR. Bacterial evolution. Microbiol Rev. 1987 Jun;51(2):221–71.

35. Gorbachev AY, Fisunov GY, Izraelson M, Evsyutina DV, Mazin PV, Alexeev DG, et al. DNA repair in Mycoplasma gallisepticum. BMC Genomics. 2013 Oct 23;14:726.

36. Sueoka N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol. 1995 Mar;40(3):318–25.

37. Eyre-Walker A. DNA mismatch repair and synonymous codon evolution in mammals. Mol Biol Evol. 1994 Jan;11(1):88–98.

38. Sharp PM, Shields DC, Wolfe KH, Li WH. Chromosomal location and evolutionary rate variation in enterobacterial genes. Science. 1989 Nov 10;246(4931):808–10.

39. Matic I, Rayssiguier C, Radman M. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. Cell. 1995 Feb 10;80(3):507–15.

40. Sniegowski P. Mismatch repair: origin of species? Curr Biol. 1998 Jan 15;8(2):R59-61.

41. Jukes TH, Cantor CR. CHAPTER 24 - Evolution of Protein Molecules. In: Munro HN, editor. Mammalian Protein Metabolism [Internet]. Academic Press; 1969 [cited 2022 May 27]. p. 21–132. Available from: https://www.sciencedirect.com/science/article/pii/B9781483232119500097

42. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980 Jun 1;16(2):111–20.

43. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22(2):160–74.

44. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993 May;10(3):512–26.

45. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. undefined [Internet]. 1986 [cited 2022 May 27]; Available from: https://www.semanticscholar.org/paper/Some-probabilistic-and-statistical-problems-in-the-Tavar%C3%A9/55e3359cd05b1903ffd8f633eb5aa6156791b364

46. Felsenstein J. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Systematic Biology. 1973 Sep 1;22(3):240–9.

47. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.

48. Felsenstein J. Phylogenies and the Comparative Method. The American Naturalist. 1985;125(1):1–15.

49. Garland T Jr, Dickerman AW, Janis CM, Jones JA. Phylogenetic Analysis of Covariance by Computer Simulation. Systematic Biology. 1993 Sep 1;42(3):265–92.

50. FRITZ SA, PURVIS A. Selectivity in Mammalian Extinction Risk and Threat Types: a New Measure of Phylogenetic Signal Strength in Binary Traits. Conservation Biology. 2010;24(4):1042–51.

51. Borges R, Machado JP, Gomes C, Rocha AP, Antunes A. Measuring phylogenetic signal between categorical traits and phylogenies. Bioinformatics. 2019 Jun 1;35(11):1862–9.

52. Bollback JP. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics. 2006 Feb 23;7(1):88.

53. Pagel M. Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. Proceedings: Biological Sciences. 1994;255(1342):37–45.

54. Pagel M, Meade A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat. 2006 Jun;167(6):808–25.

55. Tenaillon O, Denamur E, Matic I. Evolutionary significance of stress-induced mutagenesis in bacteria. Trends Microbiol. 2004 Jun;12(6):264–70.

56. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science. 2017 Mar 24;355(6331):1330–4.

57. Srinivas US, Tan BWQ, Vellayappan BA, Jeyasekharan AD. ROS and the DNA damage response in cancer. Redox Biol. 2019 Jul;25:101084.

58. Aniukwu J, Glickman MS, Shuman S. The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends. Genes Dev. 2008 Feb 15;22(4):512–27.

59. Bhattarai H, Gupta R, Glickman MS. DNA ligase C1 mediates the LigD-independent nonhomologous end-joining pathway of Mycobacterium smegmatis. J Bacteriol. 2014 Oct;196(19):3366–76.

60. Bétermier M, Bertrand P, Lopez BS. Is non-homologous end-joining really an inherently error-prone process? PLoS Genet. 2014 Jan;10(1):e1004086.

61. Bowater R, Doherty AJ. Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. PLoS Genet. 2006 Feb;2(2):e8.

62. Della M, Palmbos PL, Tseng HM, Tonkin LM, Daley JM, Topper LM, et al. Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. Science. 2004 Oct 22;306(5696):683–5.

63. Stephanou NC, Gao F, Bongiorno P, Ehrt S, Schnappinger D, Shuman S, et al. Mycobacterial nonhomologous end joining mediates mutagenic repair of

chromosomal double-strand DNA breaks. J Bacteriol. 2007 Jul;189(14):5237–46.

64. Zhu H, Shuman S. Gap filling activities of Pseudomonas DNA ligase D (LigD) polymerase and functional interactions of LigD with the DNA end-binding Ku protein. J Biol Chem. 2010 Feb 12;285(7):4815–25.

65. Moeller R, Stackebrandt E, Reitz G, Berger T, Rettberg P, Doherty AJ, et al. Role of DNA repair by nonhomologous-end joining in Bacillus subtilis spore resistance to extreme dryness, mono- and polychromatic UV, and ionizing radiation. J Bacteriol. 2007 Apr;189(8):3306–11.

66. Malyarchuk S, Wright D, Castore R, Klepper E, Weiss B, Doherty AJ, et al. Expression of Mycobacterium tuberculosis Ku and Ligase D in Escherichia coli results in RecA and RecB-independent DNA end-joining at regions of microhomology. DNA Repair (Amst). 2007 Oct 1;6(10):1413–24.

67. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

68. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013 Oct 1;29(19):2487–9.

69. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol. 2000 Apr;66(4):1328–33.

70. Gyorfy Z, Draskovits G, Vernyik V, Blattner FF, Gaal T, Posfai G. Engineered ribosomal RNA operon copy-number variants of E. coli reveal the evolutionary trade-offs shaping rRNA operon number. Nucleic Acids Res. 2015 Feb 18;43(3):1783–94.

71. Vernikos GS, Parkhill J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics. 2006 Sep 15;22(18):2196–203.

72. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002 Jun;51(3):492–508.

73. Bansal MS, Kellis M, Kordi M, Kundu S. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. Bioinformatics. 2018 Sep 15;34(18):3214–6.

74. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189–91.

75. Konishi T, Matsukuma S, Fuji H, Nakamura D, Satou N, Okano K. Principal Component Analysis applied directly to Sequence Matrix. Sci Rep. 2019 Dec 17;9(1):19297.

76. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010 Jul 13;10:210.

77. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–74.

78. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587–9.

79. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018 Feb 1;35(2):518–22.

80. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010 May;59(3):307–21.

81. Revell LJ, Graham Reynolds R. A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. Evolution. 2012 Sep;66(9):2697–707.

82. Elliot MG, Mooers AØ. Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. BMC Evol Biol. 2014 Nov 28;14:226.

83. Freckleton RP, Harvey PH, Pagel M. Phylogenetic analysis and comparative data: a test and review of evidence. Am Nat. 2002 Dec;160(6):712–26.

84. Blomberg SP, Garland T, Ives AR. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution. 2003 Apr;57(4):717–45.

85. Ives AR, Garland T. Phylogenetic logistic regression for binary dependent variables. Syst Biol. 2010 Jan;59(1):9–26.

86. McGovern S, Baconnais S, Roblin P, Nicolas P, Drevet P, Simonson H, 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. 2016 Jun 2;44(10):4785–806.

87. Kobayashi H, Simmons LA, Yuan DS, Broughton WJ, Walker GC. Multiple Ku orthologues mediate DNA non-homologous end-joining in the free-living form and during chronic infection of Sinorhizobium meliloti. Mol Microbiol. 2008 Jan;67(2):350–63.

88. Bartlett EJ, Brissett NC, Doherty AJ. Ribonucleolytic resection is required for repair of strand displaced nonhomologous end-joining intermediates. Proc Natl Acad Sci U S A. 2013 May 28;110(22):E1984-1991.

89. Brissett NC, Martin MJ, Bartlett EJ, Bianchi J, Blanco L, Doherty AJ. Molecular basis for DNA double-strand break annealing and primer extension by an NHEJ DNA polymerase. Cell Rep. 2013 Nov 27;5(4):1108–20.

90. Eisen JA. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J Mol Evol. 1995 Dec;41(6):1105–23.

91. Lang JM, Darling AE, Eisen JA. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PLoS One. 2013;8(4):e62510.

92. Kanhere A, Vingron M. Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes. BMC Evol Biol. 2009 Jan 10;9:9.

93. Weissman JL, Fagan WF, Johnson PLF. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. PLoS Genet. 2019 Nov;15(11):e1008493.

94. Matthews LA, Simmons LA. Bacterial nonhomologous end joining requires teamwork. J Bacteriol. 2014 Oct;196(19):3363–5.

95. Hoff G, Bertrand C, Piotrowski E, Thibessard A, Leblond P. Genome plasticity is governed by double strand break DNA repair in Streptomyces. Sci Rep. 2018 Mar 27;8(1):5272.

96. Hoff G, Bertrand C, Zhang L, Piotrowski E, Chipot L, Bontemps C, et al. Multiple and Variable NHEJ-Like Genes Are Involved in Resistance to DNA Damage in Streptomyces ambofaciens. Front Microbiol. 2016;7:1901.

97. Dupuy P, Gourion B, Sauviac L, Bruand C. DNA double-strand break repair is involved in desiccation resistance of Sinorhizobium meliloti, but is not essential for its symbiotic interaction with Medicago truncatula. Microbiology (Reading). 2017 Mar;163(3):333–42.

98. Dupuy P, Sauviac L, Bruand C. Stress-inducible NHEJ in bacteria: function in DNA repair and acquisition of heterologous DNA. Nucleic Acids Res. 2019 Feb 20;47(3):1335–49.

99. Pitcher RS, Tonkin LM, Daley JM, Palmbos PL, Green AJ, Velting TL, et al. Mycobacteriophage exploit NHEJ to facilitate genome circularization. Mol Cell. 2006 Sep 1;23(5):743–8.

100. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 1997 Apr;44(4):383–97.

101. Sinha KM, Unciuleac MC, Glickman MS, Shuman S. AdnAB: a new DSB-resecting motor-nuclease from mycobacteria. Genes Dev. 2009 Jun 15;23(12):1423–37.

102. Gupta R, Barkan D, Redelman-Sidi G, Shuman S, Glickman MS. Mycobacteria exploit three genetically distinct DNA double-strand break repair pathways. Mol Microbiol. 2011 Jan;79(2):316–30.

103. Zhang X, Chen W, Zhang Y, Jiang L, Chen Z, Wen Y, et al. Deletion of ku homologs increases gene targeting frequency in Streptomyces avermitilis. J Ind Microbiol Biotechnol. 2012 Jun;39(6):917–25.

104. Kushwaha AK, Grove A. C-terminal low-complexity sequence repeats of Mycobacterium smegmatis Ku modulate DNA binding. Biosci Rep. 2013 Jan 24;33(1):175–84.

105. Woese CR, Maniloff J, Zablen LB. Phylogenetic analysis of the mycoplasmas. Proc Natl Acad Sci U S A. 1980 Jan;77(1):494–8.

106. Wolf M, Müller T, Dandekar T, Pollack JD. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. Int J Syst Evol Microbiol. 2004 May;54(Pt 3):871–5.

107. Oshima K, Maejima K, Namba S. Genomic and evolutionary aspects of phytoplasmas. Front Microbiol. 2013;4:230.

108. Ipoutcha T, Tsarmpopoulos I, Talenton V, Gaspin C, Moisan A, Walker CA, et al. Multiple Origins and Specific Evolution of CRISPR/Cas9 Systems in Minimal Bacteria (Mollicutes). Front Microbiol. 2019;10:2701.

109. Biehs R, Steinlage M, Barton O, Juhász S, Künzel J, Spies J, et al. DNA Double-Strand Break Resection Occurs during Non-homologous End Joining in G1 but Is Distinct from Resection during Homologous Recombination. Mol Cell. 2017 Feb 16;65(4):671-684.e5.

110. Shen M, Zhang H, Shen W, Zou Z, Lu S, Li G, et al. Pseudomonas aeruginosa MutL promotes large chromosomal deletions through non-homologous end joining to prevent bacteriophage predation. Nucleic Acids Res. 2018 May 18;46(9):4505–14.

111. Wang ST, Setlow B, Conlon EM, Lyon JL, Imamura D, Sato T, et al. The forespore line of gene expression in Bacillus subtilis. J Mol Biol. 2006 Apr 21;358(1):16–37.

112. Li Z, Wen J, Lin Y, Wang S, Xue P, Zhang Z, et al. A Sir2-like protein participates in mycobacterial NHEJ. PLoS One. 2011;6(5):e20045.

113. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. 2011 Apr;21(4):599–609.

114. Seong HJ, Han SW, Sul WJ. Prokaryotic DNA methylation and its functional roles. J Microbiol. 2021 Mar;59(3):242–8.

115. Xiao W, Samson L. In vivo evidence for endogenous DNA alkylation damage as a source of spontaneous mutation in eukaryotic cells. Proc Natl Acad Sci U S A. 1993 Mar 15;90(6):2117–21.

116. Taverna P, Sedgwick B. Generation of an endogenous DNA-methylating agent by nitrosation in Escherichia coli. J Bacteriol. 1996 Sep;178(17):5105–11.

117. Mielecki D, Wrzesiński M, Grzesiuk E. Inducible repair of alkylated DNA in microorganisms. Mutat Res Rev Mutat Res. 2015 Mar;763:294–305.

118. Fu D, Calvo JA, Samson LD. Balancing repair and tolerance of DNA damage caused by alkylating agents. Nat Rev Cancer. 2012 Jan 12;12(2):104–20.

119. Defais M. The adaptive response in E. coli. Biochimie. 1985 Apr;67(3–4):357–60.

120. Sedgwick B. The Adaptive Response to Alkylation Damage in Escherichia Coli. In: Kappas A, editor. Mechanisms of Environmental Mutagenesis-Carcinogenesis [Internet]. Boston, MA: Springer US; 1990 [cited 2022 May 26]. p. 117–28. Available from: https://doi.org/10.1007/978-1-4615-3808-0_9

121. Kleibl K. Molecular mechanisms of adaptive response to alkylating agents in Escherichia coli and some remarks on O(6)-methylguanine DNA-methyltransferase in other organisms. Mutat Res. 2002 Sep;512(1):67–84.

122. Sedgwick B. Repairing DNA-methylation damage. Nat Rev Mol Cell Biol. 2004 Feb;5(2):148–57.

123. Sedgwick B, Vaughan P. Widespread adaptive response against environmental methylating agents in microorganisms. Mutat Res. 1991 Oct;250(1–2):211–21.

124. Moore MH, Gulbis JM, Dodson EJ, Demple B, Moody PC. Crystal structure of a suicidal DNA repair protein: the Ada O6-methylguanine-DNA methyltransferase from E. coli. EMBO J. 1994 Apr 1;13(7):1495–501.

125. Sedgwick B, Lindahl T. Recent progress on the Ada response for inducible repair of DNA alkylation damage. Oncogene. 2002 Dec 16;21(58):8886–94.

126. He C, Hus JC, Sun LJ, Zhou P, Norman DPG, Dötsch V, et al. A methylation-dependent electrostatic switch controls DNA repair and transcriptional activation by E. coli ada. Mol Cell. 2005 Oct 7;20(1):117–29.

127. Yu B, Edstrom WC, Benach J, Hamuro Y, Weber PC, Gibney BR, et al. Crystal structures of catalytic complexes of the oxidative DNA/RNA repair enzyme AlkB. Nature. 2006 Feb 16;439(7078):879–84.

128. Yang CG, Yi C, Duguid EM, Sullivan CT, Jian X, Rice PA, et al. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. Nature. 2008 Apr 24;452(7190):961–5.

129. Bowles T, Metz AH, O'Quin J, Wawrzak Z, Eichman BF. Structure and DNA binding of alkylation response protein AidB. Proc Natl Acad Sci U S A. 2008 Oct 7;105(40):15299–304.

130. Bowman BR, Lee S, Wang S, Verdine GL. Structure of the E. coli DNA glycosylase AlkA bound to the ends of duplex DNA: a system for the structure determination of lesion-containing DNA. Structure. 2008 Aug 6;16(8):1166–74.

131. Van Houten B, Sancar A. Repair of N-methyl-N'-nitro-N-nitrosoguanidine-induced DNA damage by ABC excinuclease. J Bacteriol. 1987 Feb;169(2):540–5.

132. Maslowska KH, Makiela-Dzbenska K, Fijalkowska IJ. The SOS system: A complex and tightly regulated response to DNA damage. Environ Mol Mutagen. 2019 May;60(4):368–84.

133. Mishina Y, He C. Oxidative dealkylation DNA repair mediated by the mononuclear non-heme iron AlkB proteins. J Inorg Biochem. 2006 Apr;100(4):670–8.

134. Fedeles BI, Singh V, Delaney JC, Li D, Essigmann JM. The AlkB Family of Fe(II)/α-Ketoglutarate-dependent Dioxygenases: Repairing Nucleic Acid Alkylation Damage and Beyond. J Biol Chem. 2015 Aug 21;290(34):20734–42.

135. Bian K, Lenz SAP, Tang Q, Chen F, Qi R, Jost M, et al. DNA repair enzymes ALKBH2, ALKBH3, and AlkB oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine in vitro. Nucleic Acids Res. 2019 Jun 20;47(11):5522–9.

136. Poncin K, Roba A, Jimmidi R, Potemberg G, Fioravanti A, Francis N, et al. Occurrence and repair of alkylating stress in the intracellular pathogen Brucella abortus. Nat Commun. 2019 Oct 24;10(1):4847.

137. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research. 2011 Jul 1;39(suppl_2):W29–37.

138. Mielecki D, Saumaa S, Wrzesiński M, Maciejewska AM, Żuchniewicz K, Sikora A, et al. Pseudomonas putida AlkA and AlkB Proteins Comprise Different Defense Systems for the Repair of Alkylation Damage to DNA – In Vivo, In Vitro, and In Silico Studies. PLoS One. 2013 Oct 2;8(10):e76198.

139. Leiros I, Nabong MP, Grøsvik K, Ringvoll J, Haugland GT, Uldal L, et al. Structural basis for enzymatic excision of N1-methyladenine and N3-methylcytosine from DNA. EMBO J. 2007 Apr 18;26(8):2206–17.

140. Moe E, Hall DR, Leiros I, Monsen VT, Timmins J, McSweeney S. Structure-function studies of an unusual 3-methyladenine DNA glycosylase II (AlkA) from Deinococcus radiodurans. Acta Crystallogr D Biol Crystallogr. 2012 Jun;68(Pt 6):703–12.

141. Grøsvik K, Tesfahun AN, Muruzábal-Lecumberri I, Haugland GT, Leiros I, Ruoff P, et al. The Escherichia coli alkA Gene Is Activated to Alleviate Mutagenesis by an Oxidized Deoxynucleoside. Frontiers in Microbiology [Internet]. 2020 [cited 2022 May 28];11. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2020.00263

142. Loeb LA, Preston BD. Mutagenesis by apurinic/apyrimidinic sites. Annu Rev Genet. 1986;20:201–30.

143. Mello LV, Rigden DJ. A new family of bacterial DNA repair proteins annotated by the integration of non-homology, distant homology and structural bioinformatic methods. FEBS Letters. 2012 Nov 2;586(21):3908–13.

144. Yu B, Hunt JF. Enzymological and structural studies of the mechanism of promiscuous substrate recognition by the oxidative DNA repair enzyme AlkB. Proc Natl Acad Sci U S A. 2009 Aug 25;106(34):14315–20.

145. Holland HD. The oxygenation of the atmosphere and oceans. Philos Trans R Soc Lond B Biol Sci. 2006 Jun 29;361(1470):903–15.

146. Lyons TW, Reinhard CT, Planavsky NJ. The rise of oxygen in Earth's early ocean and atmosphere. Nature. 2014 Feb 20;506(7488):307–15.

147. Prorok P, Grin IR, Matkarimov BT, Ishchenko AA, Laval J, Zharkov DO, et al. Evolutionary Origins of DNA Repair Pathways: Role of Oxygen Catastrophe in the Emergence of DNA Glycosylases. Cells. 2021 Jun 24;10(7):1591.

148. Teo I, Sedgwick B, Kilpatrick MW, McCarthy TV, Lindahl T. The intracellular signal for induction of resistance to alkylating agents in E. coli. Cell. 1986 Apr 25;45(2):315–24.

149. Trewick SC, Henshaw TF, Hausinger RP, Lindahl T, Sedgwick B. Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. Nature. 2002 Sep 12;419(6903):174–8.

150. Begley TJ, Samson LD. AlkB mystery solved: oxidative demethylation of N1-methyladenine and N3-methylcytosine adducts by a direct reversal mechanism. Trends Biochem Sci. 2003 Jan;28(1):2–5.

151. Mitra S. MGMT: A Personal Perspective. DNA Repair (Amst). 2007 Aug 1;6(8):1064–70.

152. Soll JM, Sobol RW, Mosammaparast N. Regulation of DNA Alkylation Damage Repair: Lessons and Therapeutic Opportunities. Trends Biochem Sci. 2017 Mar;42(3):206–18.

153. Drusano GL. Antimicrobial pharmacodynamics: critical interactions of “bug and drug.” Nat Rev Microbiol. 2004 Apr;2(4):289–300.

154. Tudek B, Zdżalik-Bielecka D, Tudek A, Kosicki K, Fabisiewicz A, Speina E. Lipid peroxidation in face of DNA damage, DNA repair and other cellular processes. Free Radic Biol Med. 2017 Jun;107:77–89.

155. Guttenplan JB. N-nitrosamines: bacterial mutagenesis and in vitro metabolism. Mutat Res. 1987 Sep;186(2):81–134.

156. Mackay WJ, Han S, Samson LD. DNA alkylation repair limits spontaneous base substitution mutations in Escherichia coli. J Bacteriol. 1994 Jun;176(11):3224–30.

157. Ulusoy S, Ulusoy HI, Pleissner D, Eriksen NT. Nitrosation and analysis of amino acid derivatives by isocratic HPLC. RSC Adv. 2016 Feb 1;6(16):13120–8.

158. de La Pomélie D, Santé-Lhoutellier V, Gatellier P. Mechanisms and kinetics of tryptophan N-nitrosation in a gastro-intestinal model. Food Chem. 2017 Mar 1;218:487–95.

159. Pullman A, Pullman B. Molecular electrostatic potential of the nucleic acids. Q Rev Biophys. 1981 Aug;14(3):289–380.

160. Kohn KW, Hartley JA, Mattes WB. Mechanisms of DNA sequence selective alkylation of guanine-N7 positions by nitrogen mustards. Nucleic Acids Res. 1987 Dec 23;15(24):10531–49.

161. Richardson FC, Richardson KK. Sequence-dependent formation of alkyl DNA adducts: a review of methods, results, and biological correlates. Mutat Res. 1990 Dec;233(1–2):127–38.

162. Nichols DS, McMeekin TA. Biomarker techniques to screen for bacteria that produce polyunsaturated fatty acids. J Microbiol Methods. 2002 Feb;48(2–3):161–70.

163. Winczura A, Zdżalik D, Tudek B. Damage of DNA and proteins by major lipid peroxidation products in genome stability. Free Radic Res. 2012 Apr;46(4):442–59.

164. Hassan N, Anesio AM, Rafiq M, Holtvoeth J, Bull I, Haleem A, et al. Temperature Driven Membrane Lipid Adaptation in Glacial Psychrophilic Bacteria. Frontiers in Microbiology [Internet]. 2020 [cited 2022 Jun 8];11. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2020.00824

165. Rossi M, Buzzini P, Cordisco L, Amaretti A, Sala M, Raimondi S, et al. Growth, lipid accumulation, and fatty acid composition in obligate psychrophilic, facultative psychrophilic, and mesophilic yeasts. FEMS Microbiology Ecology. 2009 Aug 3;69(3):363–72.

166. Bielski BH, Arudi RL, Sutherland MW. A study of the reactivity of HO2/O2- with unsaturated fatty acids. J Biol Chem. 1983 Apr 25;258(8):4759–61.

167. Zhang Z, Wang J, Wang J, Wang J, Li Y. Estimate of the sequenced proportion of the global prokaryotic genome. Microbiome. 2020 Sep 16;8(1):134.

168. Smits THM. The importance of genome sequence quality to microbial comparative genomics. BMC Genomics. 2019 Aug 20;20(1):662.

169. Young AD, Gillung JP. Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. Systematic Entomology. 2020;45(2):225–47.

170. Arenas M. Trends in substitution models of molecular evolution. Front Genet. 2015 Oct 26;6:319.

171. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol. 1994 May;11(3):459–68.

172. Losos JB, Irschick DJ, Schoener TW. Adaptation and Constraint in the Evolution of Specialization of Bahamian Anolis Lizards. Evolution. 1994;48(6):1786–98.

173. Martins EP. Conducting Phylogenetic Comparative Studies When the Phylogeny Is Not Known. Evolution. 1996;50(1):12–22.

174. Khaledian E, Brayton KA, Broschat SL. A Systematic Approach to Bacterial Phylogeny Using Order Level Sampling and Identification of HGT Using Network Science. Microorganisms. 2020 Feb 24;8(2):E312.

175. Sober E. The Nature of Selection: Evolutionary Theory in Philosophical Focus. University of Chicago Press; 1984.

Chapter 7: Appendices

Supplementary table/files 1-6 can found at the following link:
https://drive.google.com/drive/folders/1nF3VW-KEjGTA6emSWtn6UVfXhL40ojlP?usp=sharing

Supplementary table 7
Contingency tables and Fisher's Exact test results, run on 969 organisms:

!Conventional NHEJ = Non-conventional NHEJ + No-NHEJ
!Non-conventional NHEJ = Conventional NHEJ + No-NHEJ

	Proteobacteria	! Proteobacteria
Conventional NHEJ	98	58
! Conventional NHEJ	368	446

```
Fisher's Exact Test for Count Data

data:  prot.conventional
p-value = 3.803e-05
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.501099      Inf
sample estimates:
odds ratio
  2.046295
```

	Proteobacteria	! Proteobacteria
Non-Conventional NHEJ	15	91
! Non-Conventional NHEJ	451	413

```
Fisher's Exact Test for Count Data

data:  prot.nonconventional
p-value = 1
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.08864445        Inf
sample estimates:
odds ratio
  0.151218
```

	Actinobacteria	! Actinobacteria
Conventional NHEJ	28	128
! Conventional NHEJ	716	98

```
Fisher's Exact Test for Count Data

data:  actino.conventional
p-value = 1
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.01976974        Inf
sample estimates:
odds ratio
0.03013512
```

	Actinobacteria	! Actinobacteria

Non-Conventional NHEJ	44	62
! Non-Conventional NHEJ	82	782

```
Fisher's Exact Test for Count Data

data:  actino.nonconventional
p-value = 2.123e-15
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 4.511277      Inf
sample estimates:
odds ratio
  6.745681
```

	Bacteroidetes	! Bacteroidetes
Conventional NHEJ	19	137
! Conventional NHEJ	74	740

```
Fisher's Exact Test for Count Data

data:  bacteroidetes.conventional
p-value = 0.1468
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.8406495       Inf
sample estimates:
odds ratio
   1.38635
```

	Bacteroidetes	! Bacteroidetes
Non-Conventional NHEJ	0	106
! Non-Conventional NHEJ	93	771

```
Fisher's Exact Test for Count Data

data:  bacteroidetes.nonconventional
p-value = 1
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
   0 Inf
sample estimates:
odds ratio
         0
```

	Firmicutes	! Firmicutes
Conventional NHEJ	5	151
! Conventional NHEJ	122	692

```
Fisher's Exact Test for Count Data

data:  firmicutes.conventional
p-value = 1
```

```
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.07174517        Inf
sample estimates:
odds ratio
 0.1880178
```

	Firmicutes	! Firmicutes
Non-Conventional NHEJ	42	64
! Non-Conventional NHEJ	85	779

```
Fisher's Exact Test for Count Data

data:  firmicutes.nonconventional
p-value = 1.23e-13
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 4.00504     Inf
sample estimates:
odds ratio
  5.997159
```

	Acidobacteria	! Acidobacteria
Conventional NHEJ	3	153
! Conventional NHEJ	3	811

```
Fisher's Exact Test for Count Data

data:  acidobacteria.conventional
p-value = 0.05622
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.947979      Inf
sample estimates:
odds ratio
  5.286889
```

	Acidobacteria	! Acidobacteria
Non-Conventional NHEJ	2	104
! Non-Conventional NHEJ	4	860

```
Fisher's Exact Test for Count Data

data:  acidobacteria.nonconventional
p-value = 0.1325
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.5472948       Inf
sample estimates:
odds ratio
  4.125139
```

	Chlamydiae	! Chlamydiae
Conventional NHEJ	1	155
! Conventional NHEJ	5	809

```
Fisher's Exact Test for Count Data

data:  chlamydiae.conventional
p-value = 0.6518
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.04453014        Inf
sample estimates:
odds ratio
  1.043831
```

	Chlamydiae	! Chlamydiae
Non-Conventional NHEJ	1	105
! Non-Conventional NHEJ	5	859

```
Fisher's Exact Test for Count Data

data:  chlamydiae.nonconventional
p-value = 0.5015
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.06959043        Inf
sample estimates:
odds ratio
  1.635181
```

	Verrucomicrobia	! Verrucomicrobia
Conventional NHEJ	0	156
! Conventional NHEJ	6	808

```
Fisher's Exact Test for Count Data

data:  verrucomicrobia.conventional
p-value = 1
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
   0 Inf
sample estimates:
odds ratio
         0
```

	Verrucomicrobia	! Verrucomicrobia
Non-Conventional NHEJ	1	105
! Non-Conventional NHEJ	5	859

```
Fisher's Exact Test for Count Data

data:  verrucomicrobia.nonconventional
p-value = 0.5015
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.06959043        Inf
sample estimates:
odds ratio
  1.635181
```

List of phyla with all *No-NHEJ* state organisms only:

1) Tenericutes
2) Thermotogae
3) Synergistetes
4) Fusobacteria
5) Chlorobi
6) Deferribacteres
7) Elusimicrobia
8) Ignavibacteriae
9) Fibrobacteres
10) Coprothermobacterota
11) Chrysiogenetes
12) Calditrichaeota
13) Kirtimatiellaeota
14) Candidatus Cloacimonetes

List of phyla with *incomplete NHEJ* or *No-NHEJ* state organisms only:

1) Caldiserica
2) Dictyoglomi
3) Gemmatimonadetes
4) Thermobaculum
5) Deinococcus-Thermus
6) Planctomycetes
7) Aquificae
8) Chloroflexi
9) Cyanobacteria
10) Spirochaetes

Supplementary table 8

	Sub-phyla name	Type of gain	Type of loss	Organism names
	Flavobacteriia	Sequential secondary gain (NHEJ- to LigD to NHEJ+)	-	Maribacter sp. 1_2014MBL_MicDiv; Cellulophaga baltica NN016038; Flavobacterium johnsoniae UW101; Aequorivita sublithincola DSM 14238; Zunongwangia profunda SM-A87; Gramella sp. LPB0144
		Secondary direct gain (NHEJ- to NHEJ+)	-	Chryseobacterium sp. IHB B17019; Flavobacteriaceae bacterium 3519-10
	Common ancestor of Flavobacteriia and Bacteroidia	-	Major Secondary loss of LigD to NHEJ-	-
	Sphingobacteria	Primary sequential gain (LigD to NHEJ+)	-	Mucilaginibacter sp. PAMC 26640 plasmid unnamed; Pedobacter saltans DSM 12145; Sphingobacterium sp. ML3W; Solitalea canadensis DSM 3403
	Chitinophagia	Primary NHEJ sequential gain (LigD to NHEJ+)	-	Niabella ginsenosidivorans strain BS26; Niastella koreensis GR20−10; Flavisolibacter sp. LCS9; Chitinophaga pinensis DSM 2588
	Cytophagia	Primary NHEJ sequential gain (LigD to NHEJ+)	-	Cytophaga hutchinsonii ATCC 33406; Dyadobacter fermentans DSM 18053; Cyclobacterium amurskyense strain KCTC 12363
	Common ancestor of Flavobacteriia, Bacteroidia, Sphingobacteria, Chitinophagia and a subclade of Cytophagia	Major Primary LigD gain	-	-

	Thermoleophilia	Minor Primary direct gain (NHEJ- to NHEJ+)	-	Conexibacter woesei DSM 14684
	Coriobacteriia	Minor Primary direct gain (NHEJ- to NHEJ+)	-	Eggerthella lenta DSM 2243
	Acidimicrobiia	Minor Primary direct gain (NHEJ- to NHEJ+)	-	Ilumatobacter coccineus YM16−304
	Common ancestor of Frankiales, Geodermatophilales, Streptosporangiales, Streptomycetales, Catenulisporales, Glycomycetales, Pseudonocardiales, Propionibacteriales, Corynebacteriales, Micromonosporales, Bifidobacteriales, Micrococcales, Kineosporiales, Nakamurellales, Actinomycetales	Major Primary direct gain	-	
	Frankiales	Primary direct gain		Frankia sp. EuI1c; Frankia sp. EAN1pec
	Geodermatophilales	Primary direct gain		Blastococcus saxobsidens DD2; Geodermatophilus obscurus DSM 43160; Modestobacter marinus str. BC501
	Actinobacteria incertae sedis	Primary direct gain		Thermobispora bispora DSM 43833
	Streptosporangiales	Primary direct gain		Thermomonospora curvata DSM 43183; Streptosporangium roseum DSM 43021
	Streptosporangiales	-	Direct secondary loss	Thermobifida fusca; Nocardiopsis alba ATCC BAA−2165
	Catenulisporales	Primary direct gain		Catenulispora acidiphila DSM 44928
	Streptomycetales	-	Direct secondary loss	Kitasatospora setae KM−6054 DNA
	Streptomycetales	Primary direct gain	-	Streptomyces lydicus strain 103;

				Streptomyces sampsonii strain KJ40
	Pseudonocardiales	Primary direct gain	-	Actinoalloteichus hymeniacidonis strain HPA177(T) (=DSM 45092(T)); Pseudonocardia dioxanivorans CB1190; Kibdelosporangium phytohabitans strain KLBMP1111; Saccharothrix espanaensis DSM; Alloactinosynnema sp. L−07; Actinosynnema mirum DSM 43827; Lentzea guizhouensis strain DHS C013; Saccharopolyspora erythraea NRRL2338; Saccharomonospora viridis DSM 43017; Amycolatopsis japonica strain MG417−CF17;
	Glycomycetales	Primary direct gain	-	Stackebrandtia nassauensis DSM 44728;
	Corynebacteriales	Primary direct gain	-	Rhodococcus fascians D188; Nocardia seriolae strain EM150506; Corynebacterium humireducens NBRC 106098 = DSM 45392; Mycobacterium tuberculosis 49−02; Mycobacterium goodii strain 7B; Amycolicicoccus subflavus DQS3−9A1; Gordonia sp. QH−11; Gordonia sp. KTR9; Tsukamurella paurometabola DSM 20162;
	Corynebacteriales	-	Direct secondary loss	Brevibacterium flavum ZL−1; Corynebacteriales bacterium 1698; Segniliparus rotundus DSM 44985; Corynebacterium pseudotuberculosis strain 36;

				Mycobacterium leprae TN
	Propionibacteriales	Direct primary gain	-	Microlunatus phosphovorus NM−1 DNA; Nocardioides dokdonensis FR1436; Pimelobacter simplex strain VKM Ac−2033D; Aeromicrobium erythreum strain AR18; Kribbella flavida DSM 17836
	Propionibacteriales	-	Direct secondary loss	Cutibacterium avidum strain DPC 6544; Propionibacterium acnes C1
	Bifidobacteriales	-	Direct secondary loss	Scardovia inopinata JCM 12537 DNA; Parascardovia denticolens DSM 10105 = JCM 12538 DNA; Gardnerella vaginalis ATCC 14018 = JCM 11026 DNA; Bifidobacterium breve 12L
	Micromonosporales	Direct primary gain	-	Salinispora arenicola CNS−205; Salinispora tropica CNB−440; Micromonospora sp. L5; Verrucosispora maris AB−18−032; Actinoplanes friuliensis DSM 7358
	Nakamurellales	Direct primary gain	-	Nakamurella multipartita DSM 44233
	Kineosporiales	Direct primary gain	-	Kineococcus radiotolerans SRS30216 plasmid pKRAD02
	Micrococcales	Direct primary gain	-	Xylanimonas cellulosilytica DSM 15894 ; Isoptericola dokdonensis DS−3; Beutenbergia cavernae DSM 12333; Serinicoccus sp. JLT9; Kytococcus sedentarius DSM 20547; Dermacoccus

				nishinomiyaensis strain M25; Luteipulveratus mongoliensis strain MN07−A0370; Arsenicicoccus sp. oral taxon 190; Intrasporangium calvum DSM 43043; Janibacter terrae strain | F|001
	Micrococcales	-	Direct secondary loss	Jonesia denitrificans DSM 20603; Tropheryma whipplei TW08/27
	Actinomycetales	-	Direct secondary loss	Trueperella pyogenes TP8; Arcanobacterium haemolyticum DSM 20595; Actinobaculum schaalii strain CCUG 27420; Actinomyces sp. Marseille−P2985 strain Marseille−P2985T ; Mobiluncus curtisii ATCC 43063;
	Micrococcales	Direct primary gain	-	Brevibacterium linens strain SMQ−1335; Sanguibacter keddieii DSM 10542; Cellvibrio gilvus ATCC 13127; Cellulomonas fimi ATCC 484; Brachybacterium faecium DSM 4810; Sinomonas atrocyanea strain KCTC 3377; Arthrobacter aurescens TC1 plasmid TC2; Rathayibacter tritici strain NCPPB 1953; Agromyces aureus strain AR33; Microbacterium sp. 1.5R; Microbacterium sp. T11; Cryobacterium arcticum strain PAMC 27867; Curtobacterium sp. BH−2−1−1; Frondihabitans sp. SR6 plasmid 3; Clavibacter

				michiganensis subsp. insidiosus strain R1−1; Leifsonia xyli strain SE134
	Micrococcales	-	Direct secondary loss	Devriesea agamarum genome assembly S2_rcS3_S1_rcS3; Rothia mucilaginosa D| −18 DNA; Kocuria rhizophila DC2201; Dermabacter vaginalis strain AD1−86; Neomicrococcus aestuarii strain B18; Arthrobacter sp. A3; Renibacterium salmoninarum ATCC 33209; Micrococcus luteus NCTC 2665; Microbacterium sp. No. 7; Rathayibacter toxicus strain WAC3373; Microterricola viridarii strain ERGS5:02
	Fimbriimonadia	Direct primary gain	-	Fimbriimonas ginsengisoli Gsoil 348
	Clostridia	Direct primary gain	-	Thermaerobacter marianensis DSM 12885; Candidatus Desulforudis audaxviator MP104C;
	Thermodesulfobac teriales	Direct primary gain	-	Thermodesulfatator indicus DSM 15286
	Clostridia	Direct primary gain	-	Thermosediminibacter oceani DSM 16646; Thermacetogenium phaeum DSM 12270; Carboxydothermus hydrogenoformans Z−2901; Syntrophothermus lipocalidus DSM 12680; Syntrophomonas wolfei subsp. wolfei str. Goettingen G311; Moorella thermoacetica strain DSM 103132; Natranaerobius thermophilus

				JW/NM−WN−LF; Thermoanaerobacter sp. 513; Halothermothrix orenii H 168; Limnochorda pilosa DNA; Symbiobacterium thermophilum IAM 14863 DNA; Clostridium cellulosi genome assembly DG5; Clostridium stercorarium subsp. stercorarium DSM 8532; Clostridium stercorarium subsp. leptospartum DSM 9219; Thermoanaerobacterium thermosaccharolyticum M0795; Desulfotomaculum reducens MI−1; Thermincola sp. JR; Desulfitobacterium dehalogenans ATCC 51507; Desulfitobacterium metallireducens DSM 15288; Syntrophobotulus glycolicus DSM 8271; Heliobacterium modesticaldum Ice1 strain Ice1; Desulfosporosinus meridiei DSM 13257; Dehalobacter restrictus DSM 9455; Pelotomaculum thermopropionicum SI DNA
	Clostridia	-	Direct Secondary loss	Tepidanaerobacter acetatoxydans Re1; Caldicelulosiruptor becscii DSM 6725; Caldicellulosiruptor saccharolyticus DSM 8903; Ethanoligenens harbinense |UAN−3; Ruminococcus bicirculans; Halobacteroides halobius DSM 5150; Acetohalobium

				arabaticum DSM 5501; Halanaerobium praevalens DSM 2228; Flavonifractor plautii strain |L31; Intestinimonas butyriciproducens strain AF211; Oscillibacter valericigenes Sjm18−20; Mahella australiensis 50−1 BON; Ruminiclostridium thermocellum DSM 2360; Clostridium clariflavum DSM 19732; Clostridiales genomosp. BVAB3 str. UPII9−5; Eubacterium limosum strain SA11; Acetobacterium woodii DSM 1030; Desulfotomaculum acetoxidans DSM 771
	Common ancestor of Tissierellia and a sub-clade of Clostridia	-	Direct Secondary loss	
	Clostridia	-	Direct Secondary loss	Clostridium propionicum DSM 1682; Peptoclostridium difficile strain 08ACD0030; Filifactor alocis ATCC 35896; Geosporobacter ferrireducens strain IRF9; Alkaliphilus metalliredigens Q|MF; Lachnoclostridium sp. |L32; Butyrivibrio proteoclasticus B316; Roseburia hominis A2−183
	Clostridia	Direct Secondary gain	-	Blautia sp. |L58; Clostridium saccharolyticum WM1; Clostridium phytofermentans ISDg

	Tissierellia	-	Direct Secondary loss	Levyella sp. Marseille−P3170 strain Marseille−P3170T ; Murdochiella sp. Marseille−P2341 strain Marseille−P2341T ; Parvimonas micra strain KCOM 1535; Finegoldia magna ATCC 29328; Peptoniphilus sp. ING2−D1G; Anaerococcus prevotii DSM 20548
	Negativicutes	-	Direct Secondary loss	Acidaminococcus fermentans DSM 20731; Selenomonas sputigena ATCC 35185; Dialister pneumosintes strain F0677; Megasphaera elsdenii 14−14; Veillonella parvula DSM 2008
		Direct Secondary gain	-	Pelosinus fermentans JBW45
	Bacilli	Direct Primary gain	-	Kyrpidia tusciae DSM 2912; Alicyclobacillus acidocaldarius subsp. acidocaldarius Tc−4−1; Thermobacillus composti KWC4; Bacillus cellulosilyticus DSM 2522; Paenibacillus sp. LPB0068; Paenibacillus polymyxa SC2; Halobacillus halophilus DSM 2266; Terribacillus aidingensis strain MP602; Lentibacillus amyloliquefaciens strain LAM0015; Oceanobacillus iheyensis HTE831; Virgibacillus sp. 6R; Bacillus pseudofirmus OF4; Lysinibacillus sphaericus strain 2362; Solibacillus silvestris strain DSM 12223; Rummeliibacillus stabekisii strain PP9;

		-	Direct Secondary loss	Geobacillus stearothermophilus 10; Paenibacillus sp. BD3526; Salimicrobium jeotgali strain MJ3; Amphibacillus xylanus NBRC 15112; Planococcus rifietoensis strain M8; Anoxybacillus sp. B7M1; Exiguobacterium antarcticum B7; Sporosarcina psychrophila strain DSM 6497; Aneurinibacillus sp. H2; Kurthia sp. 11Kri321; Jeotgalibacillus sp. D5; Leuconostoc carnosum JB16; Oenococcus oeni PSU−1; Weissella koreensis KACC 15510; Pediococcus damnosus strain TMW 2.1534; Aerococcus christensenii strain CCUG28831; Lactobacillus salivarius strain JCM 1046; Marinilactibacillus sp. 15R; Carnobacterium sp. WN1359; Streptococcus constellatus subsp. pharyngis C818; Lactococcus garvieae Lg2 DNA; Enterococcus faecium strain 64/3; Melissococcus plutonius S1; Vagococcus teuberi strain DSM21459T; Tetragenococcus halophilus NBRC 12172 DNA; Listeria monocytogenes strain ATCC 19117; Macrococcus caseolyticus JCSC5402; Salinicoccus halodurans

				strain H3B36; Bacillus cereus subsp. cytotoxis NVH 391−98; Gemella sp. oral taxon 928; Staphylococcus aureus strain CA15
		Direct Secondary gain	-	Brevibacillus laterosporus LMG 15441
	Erysipelotrichia	Direct Secondary gain	-	Erysipelotrichaceae bacterium I46
		-	Direct Secondary loss	Turicibacter sp. H121; Erysipelothrix rhusiopathiae str. Fujisawa DNA; Faecalibaculum rodentium strain Alo17;
	Fusobacteriales	-	Direct Secondary loss	Ilyobacter polytropus DSM 2926; Fusobacterium nucleatum subsp. animalis strain KCOM 1279v; Sebaldella termitidis ATCC 33386; Leptotrichia sp. oral taxon 847; Streptobacillus moniliformis DSM 12112; Sneathia sp. Sn35;
	Mollicutes	-	Direct Secondary loss	Strawberry lethal yellows phytoplasma (CPA) str. NZSb11; Aster yellows witches'−broom phytoplasma AJWB; Maize bushy stunt phytoplasma strain M3; Onion yellows phytoplasma O|−M DNA; Acholeplasma palmae; Acholeplasma oculi; Acholeplasma oculi strain 19L; Mollicutes bacterium HR1; Mycoplasma mycoides subsp. mycoides strain Ben468; Mesoplasma florum L1; Spiroplasma culicicola

				AES−1; Ureaplasma urealyticum serovar 10 str. ATCC 33699
	Nitrospirales	Direct primary gain	-	Nitrospira moscoviensis strain NSP M−1
	Solibacteres	Direct primary gain	-	Solibacter usitatus Ellin6076
	Acidobacteriales	Direct primary gain	-	Candidatus Koribacter versatilis Ellin345; Terriglobus saanensis SP1PR4; Acidobacterium capsulatum ATCC 51196; Granulicella mallensis MP5ACT 8;
	Parachlamydiales	Direct primary gain	-	Protochlamydia naegleriophila genome assembly PNK1; Parachlamydia acanthamoebae UV−7
	Opitutae	Direct primary gain	-	Opitutus terrae PB90−1
	Deltaproteobacteria	Direct primary gain	-	Desulfovibrio africanus str. Walvis Bay; Desulfomonile tiedjei DSM 6799; Geobacter sp. M21; Geobacter uraniireducens Rf4; Myxococcus fulvus 124B02; Myxococcus stipitatus DSM 14675; Archangium gephyra strain DSM 2261; Vulgatibacter incomptus strain DSM 27710; Anaeromyxobacter dehalogenans 2CP−1; Sorangium cellulosum So0157−2; Sorangium cellulosum 'So ce 56'; Chondromyces crocatus strain Cm c5; Sandaracinus amylolyticus strain DSM 53668; Haliangium ochraceum DSM 14365
	Alphaproteobacteria	Direct primary gain	-	Gluconacetobacter diazotrophicus PAl 5;

				Tistrella mobilis KA081020-065; Porphyrobacter neustonensis strain DSM 9434; Citromicrobium sp. JL477; Erythrobacter litoralis strain DSM 8509; Altererythrobacter atlanticus strain 26DY36; Croceicoccus naphthovorans strain PQ-2; Sphingopyxis terrae NBRC 15098 strain 203-1; Blastomonas sp. RAC04; Blastomonas sp. RAC04; Novosphingobium aromaticivorans DSM 12444; Sphingomonas melonis; Sphingomonas sanxanigenens DSM 19645; Sphingobium sp. SYK-6; Caulobacteraceae bacterium OTSz_A_272; Caulobacter sp. K31; Phenylobacterium zucineum HLK1; Brevundimonas sp. GW460-12-10-14-LB 2; Asticcacaulis excentricus CB 48; Bradyrhizobium icense strain LMTR 13; Rhodopseudomonas palustris BisB5; Oligotropha carboxidovorans OM5; Nitrobacter winogradskyi Nb-255; Methylocella silvestris BL2; Beijerinckia indica subsp. indica ATCC 9039; Methylocystis sp. SC2; Starkeya novella DSM 506 anthobacter

				autotrophicus Py2; Azorhizobium caulinodans ORS 571; Rhodoplanes sp. Z2−| C6860; Chelatococcus daeguensis strain TAD1; Bosea vaviloviae strain Vaf18; Mesorhizobium loti MAFF303099; Chelativorans sp. BNC1; Martelella endophytica strain	C6887; Aminobacter aminovorans strain KCTC 2477; Hoeflea sp. IMCC20628 ; Neorhizobium galegae chromid pHAMBI540a; Agrobacterium tumefaciens strain S33; Ensifer adhaerens strain Casida A; Sinorhizobium americanum CCGM7; Ochrobactrum anthropi strain OAB; Filomicrobium sp. ; Hyphomicrobium denitrificans ATCC 51888; Rhodomicrobium vannielii ATCC 17100; Methyloceanibacter caenitepidi; Devosia sp. H5989; Pelagibacterium halotolerans B2; Rhizobium phaseoli strain R650; Shinella sp. HZN7; Parvibaculum lavamentivorans DS−1;
		-	Direct secondary loss	Alpha proteobacterium HIMB59; Novosphingobium pentaromativorans US6−1; Sphingopyxis sp. LPB0140; Zymomonas mobilis subsp. pomaceae ATCC 29192; Maricaulis maris MCS10; Micavibrio aeruginosavorus EPB; Magnetospirillum sp.

				M−1; Rhodospirillum rubrum ATCC 11170; Haematospirillum jordaniae strain H5569; Thalassospira xiamenensis M−5 = DSM 17429; Liberibacter crescens BT-1; Brucella abortus strain 63 75; Bartonella tribocorum; Magnetospira sp. QH−2
		Sequential secondary gain	-	Defluviimonas alba strain cai42; Rhodobacter sphaeroides ATCC 17025; Sulfitobacter sp. AM1-D1; Celeribacter indicus strain P73
	Gammaproteobacteria	Direct primary gain	-	Luteibacter rhizovicinus strain LJ96T; Lysobacter capsici strain 55; Lysobacter antibioticus strain ATCC 29479; Dokdonella koreensis DS−123; Dyella jiangningensis strain SBZ 3−12; Pseudoxanthomonas spadix BD−a59; Stenotrophomonas maltophilia D457; Xanthomonas gardneri strain ICMP 7383;
		-	Direct secondary loss	Xylella fastidiosa 9a5c plasmid pXF51; Xanthomonas albilineans str. GPE PC73;
	Betaproteobacteria	Direct primary gain	-	Nitrosospira briensis C−128; Methylovorus glucosetrophus SIP3-4; Massilia sp. WG5 plasmid unnamed 2
		Sequential primary gain	-	Janthinobacterium sp. 1_2014MBL_MicDiv; Pandoraea apista strain DSM 16535; Pandoraea sputorum strain DSM 21091;Collimonas

				fungivorans Ter331; Herbaspirillum seropedicae SmR1; Paraburkholderia caribensis strain Bcrs1W; Burkholderia pseudomallei Pasteur 52237; Burkholderia pseudomallei strain Burk178-Type2; Azoarcus sp. KH32C plasmid pAZKH DNA; Achromobacter xylosoxidans genome assembly NCTC10807; Bordetella flabilis strain AU10664; Bordetella holmesii ATCC 51541; Advenella kashmirensis WT001; Polyangium brachysporum strain DSM 7029; Rubrivivax gelatinosus IL144 DNA; Delftia sp. HK171; Mitsuaria sp. 7; Roseateles depolymerans strain KCTC 42856; Acidovorax ebreus TPS|; Acidovorax citrulli AAC00-1; Ramlibacter tataouinensis TTB310; Variovorax paradoxus EPS; Polaromonas sp. JS666; Hydrogenophaga sp. PBC;
		Direct tertiary gain	-	Cupriavidus basilensis strain 4G11; Ralstonia mannitolilytica strain SN82F48; Thiobacillus denitrificans ATCC 25259;
		-	Direct secondary loss	Basilea psittacipulmonis DSM 24701; Alcaligenes faecalis strain ZD02; Castellaniella; defragrans 65Phen; Bordetella pertussis strain E476; Pusillimonas sp. T7−7; Taylorella equigenitalis ATCC 35865;

	Gammaproteobacteria	Direct primary gain	-	Legionella hackeliae genome assembly LHA; Tatlockia micdadei genome assembly LMI; Coxiella burnetii CbuG_Q212; Pseudomonas syringae pv. tomato str. DC3000; Pseudomonas stutzeri DSM 4166; Halomonas chromatireducens strain AGD 8−3
		-	Direct secondary loss	Mutant Legionella pneumophila subsp. pneumophila str.; Hextuple_3a Legionella pneumophila str. Corby;
		Direct secondary gain	-	Pseudomonas aeruginosa strain PA_D16

Note: This table doesn't contain *NHEJ-* to *LigD/Ku* events that do not culminate to NHEJ+ eventually and *LigD/Ku* to *NHEJ-* events. Other organisms not listed in the table do not code for NHEJ by the virtue of the eubacterial ancestor.

Figure S1

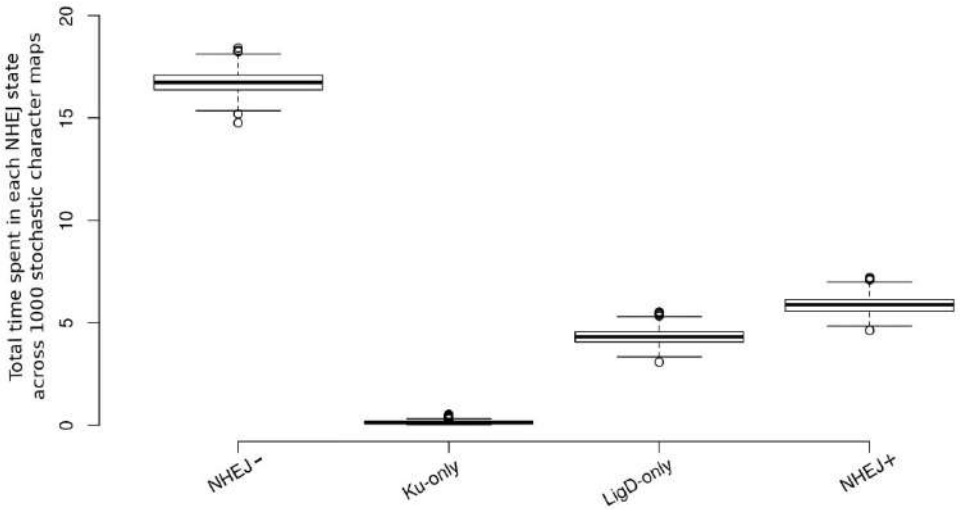

Total time spent in each NHEJ state
across 1000 stochastic character maps
0
5
10
15
20
NHEJ−
Ku-only
LigD-only
NHEJ+

Figure S2

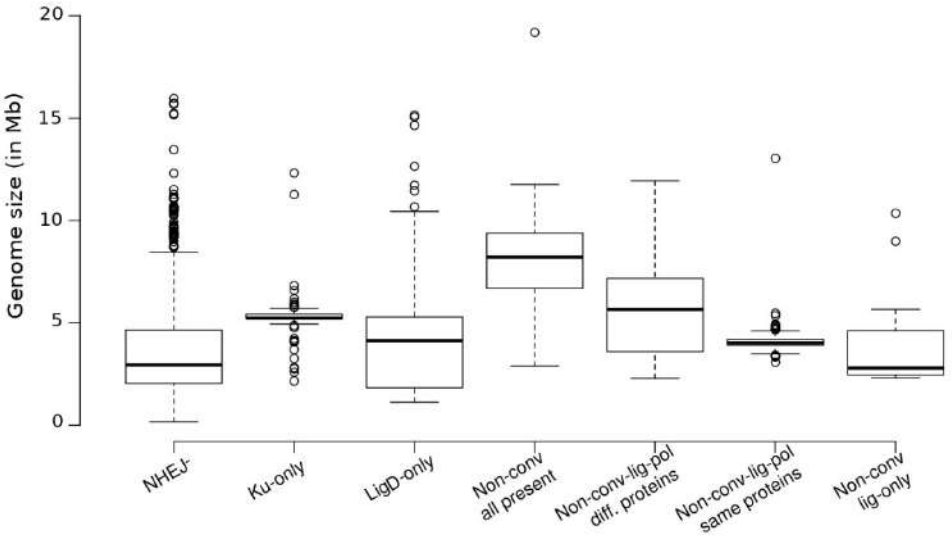

Genome size (in Mb)
20
15
10
5
0
NHEJ-
Ku-only
LigD-only
Non-conv all present
Non-conv-lig-pol diff. proteins
Non-conv-lig-pol same proteins
Non-conv lig-only

Figure S3

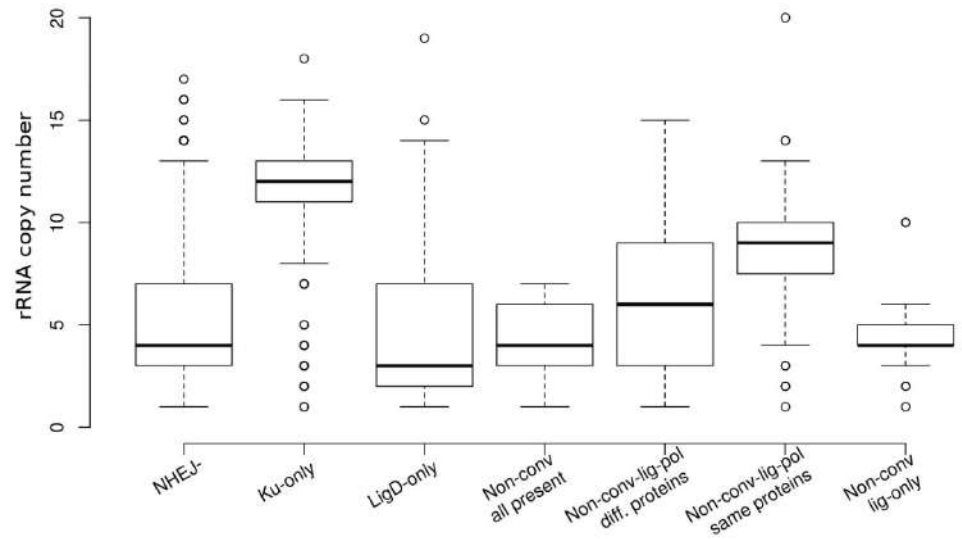

rRNA copy number
0
5
10
15
20
NHEJ-
Ku-only
LigD-only
Non-conv all present
Non-conv-lig-pol diff. proteins
Non-conv-lig-pol same proteins
Non-conv lig-only

Figure S4

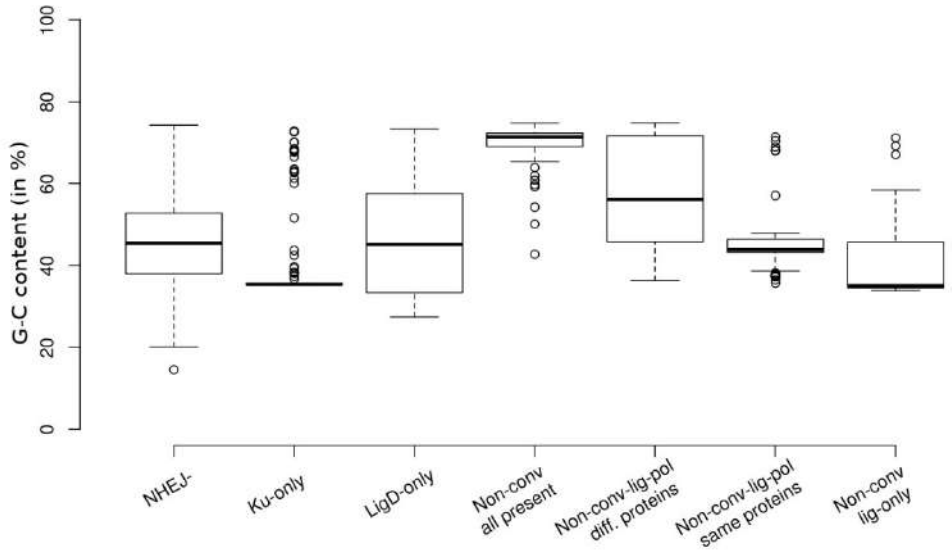

G-C content (in %)
0
20
40
60
80
100
NHEJ-
Ku-only
LigD-only
Non-conv all present
Non-conv-lig-pol diff. proteins
Non-conv-lig-pol same proteins
Non-conv lig-only

Figure S5

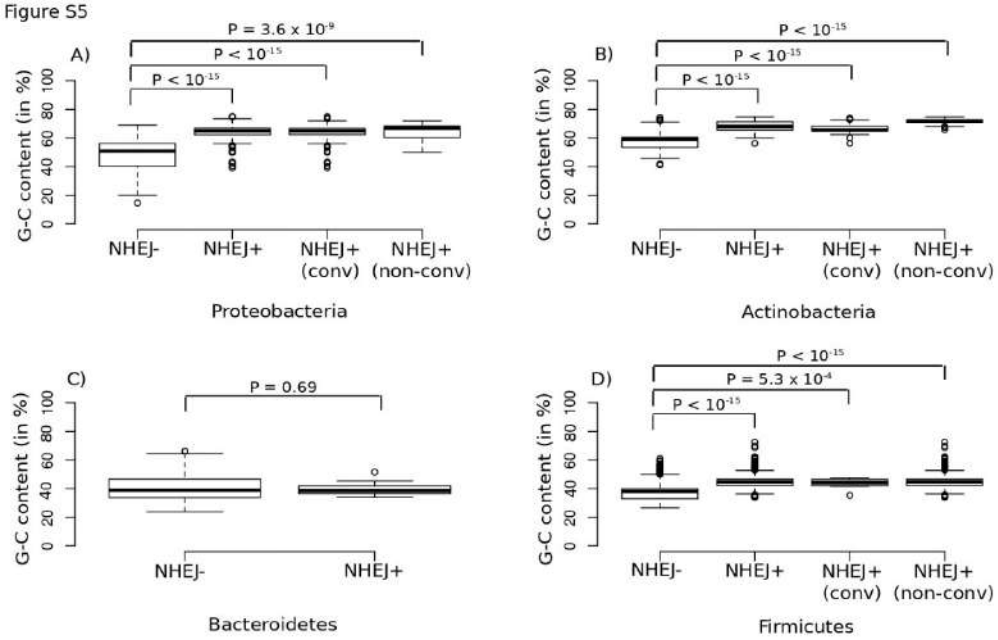

A)
P = 3.6 x 10-9
P < 10-15
P < 10-15
G-C content (in %)
0 20 40 60 80 100
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Proteobacteria
B)
P < 10-15
P < 10-15
P < 10-15
G-C content (in %)
0 20 40 60 80 100
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Actinobacteria
C)
P = 0.69
G-C content (in %)
0 20 40 60 80 100
NHEJ-
NHEJ+
Bacteroidetes
D)
P < 10-15
P = 5.3 x 10-4
P < 10-15
G-C content (in %)
0 20 40 60 80 100
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Firmicutes

Figure S6

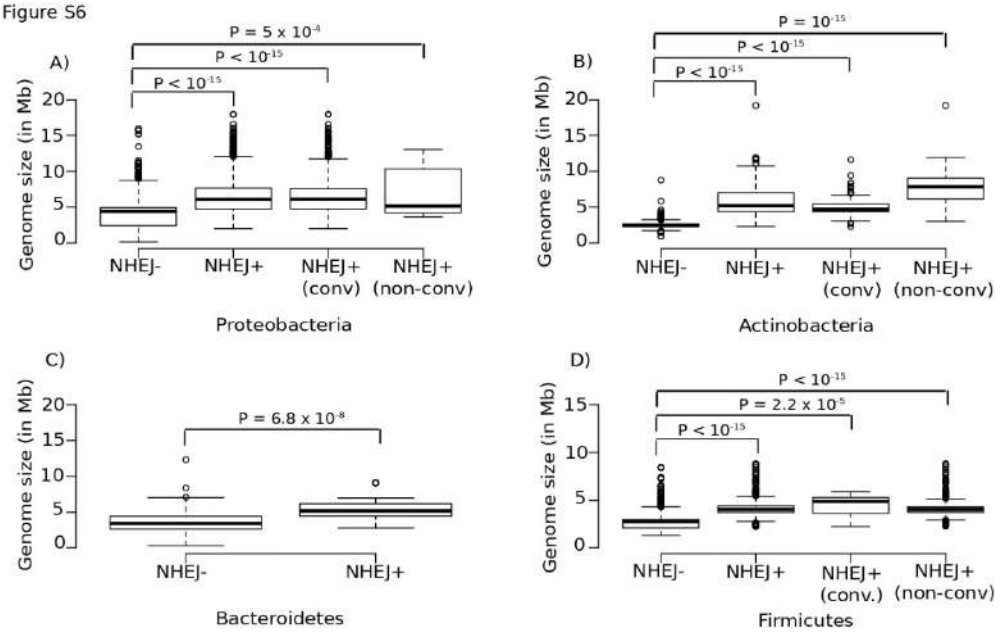

A)
P = 5 x 10^-4
P < 10^-15
P < 10^-15
Genome size (in Mb)
20
15
10
5
0
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Proteobacteria
B)
P = 10^-15
P < 10^-15
P < 10^-15
Genome size (in Mb)
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Actinobacteria
C)
P = 6.8 x 10^-8
Genome size (in Mb)
NHEJ-
NHEJ+
Bacteroidetes
D)
P < 10^-15
P = 2.2 x 10^-5
P < 10^-15
Genome size (in Mb)
NHEJ-
NHEJ+
NHEJ+ (conv.)
NHEJ+ (non-conv)
Firmicutes

Figure S7

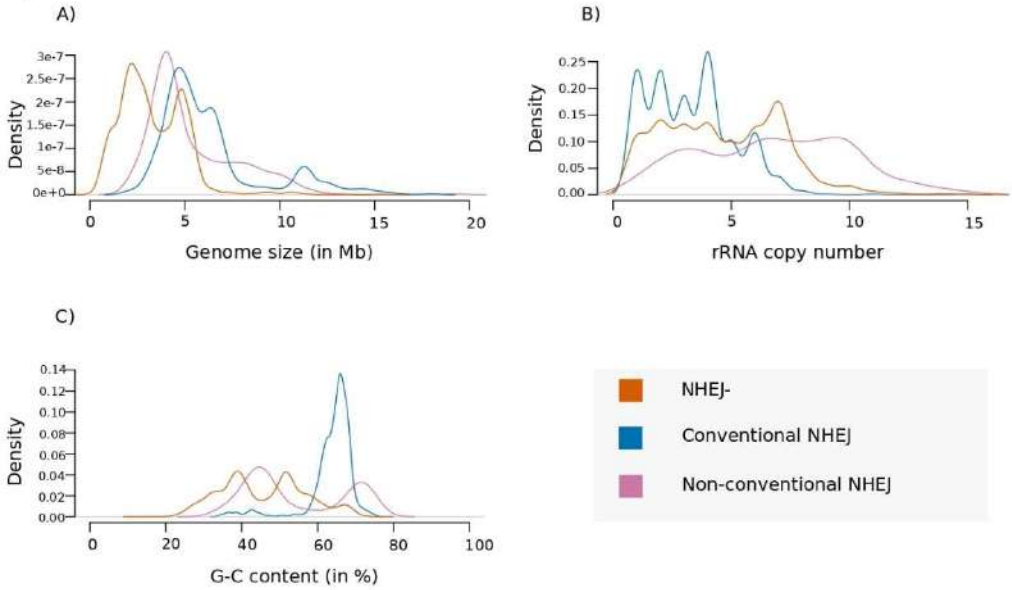
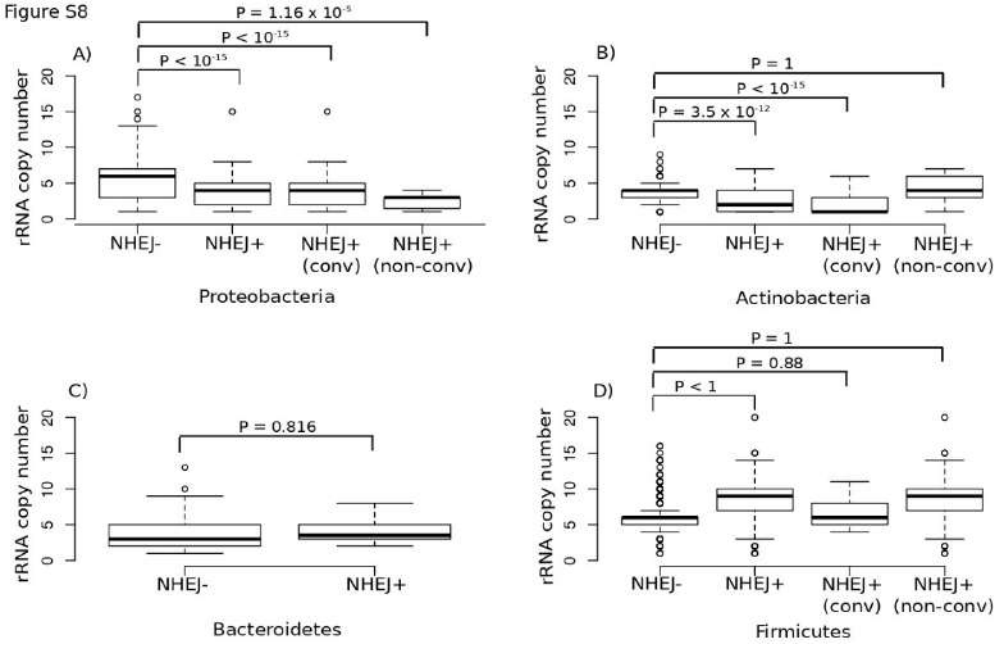

A)
Density
3e-7
2.5e-7
2e-7
1.5e-7
1e-7
5e-8
0e+0
0
5
10
15
20
Genome size (in Mb)
B)
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
rRNA copy number
C)
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
G-C content (in %)
NHEJ-
Conventional NHEJ
Non-conventional NHEJ

Figure S8


A)
P = 1.16 x 10^-5
P < 10^-15
P < 10^-15
rRNA copy number
0 5 10 15 20
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Proteobacteria
B)
P = 1
P < 10^-15
P = 3.5 x 10^-12
rRNA copy number
0 5 10 15 20
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Actinobacteria
C)
P = 0.816
rRNA copy number
0 5 10 15 20
NHEJ-
NHEJ+
Bacteroidetes
D)
P = 1
P = 0.88
P < 1
rRNA copy number
0 5 10 15 20
NHEJ-
NHEJ+
NHEJ+ (conv)
NHEJ+ (non-conv)
Firmicutes

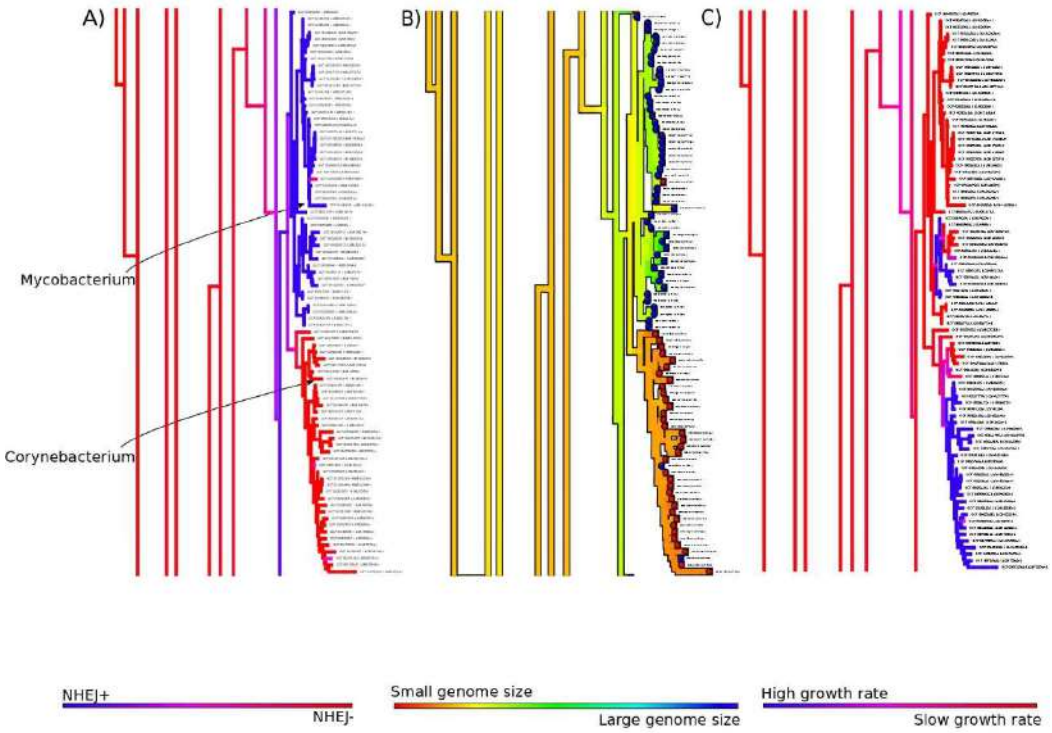
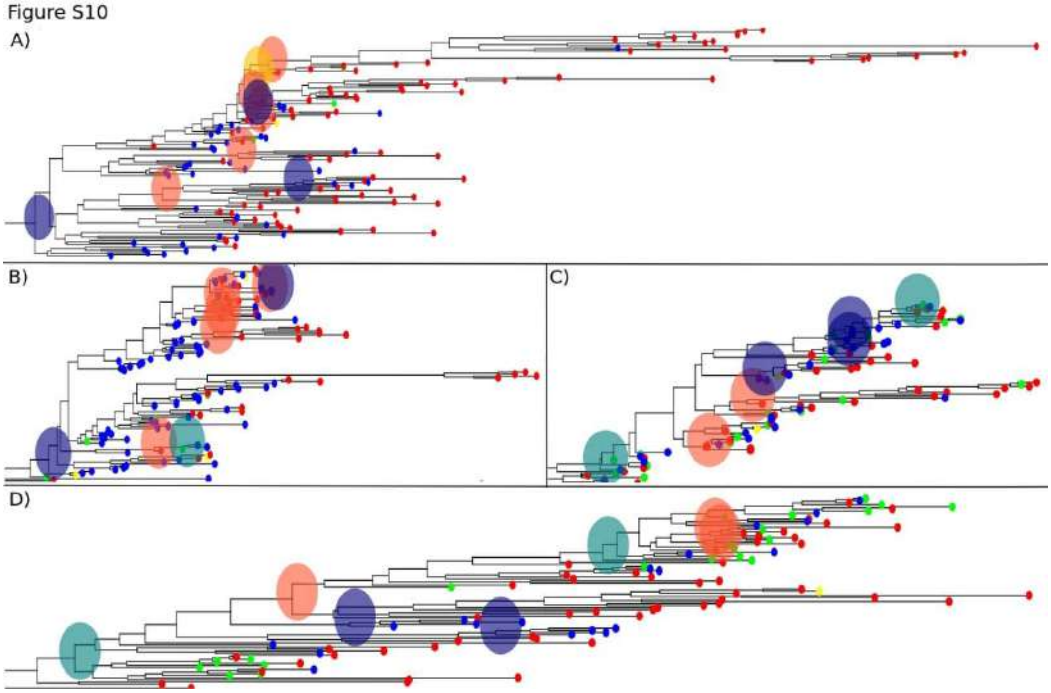

A)
B)
C)
Mycobacterium
Corynebacterium
NHEJ+
NHEJ-
Small genome size
Large genome size
High growth rate
Slow growth rate

Figure S10



Supplementary figure legends

Figure S1: Boxplots representing the distribution of total time spent in a given NHEJ state over a phylogenetic tree across 1000 stochastic character maps (Y-axis). X-axis represents the four NHEJ states: *NHEJ-*, *Ku only*, *LigD only* and *NHEJ+*.

Figure S2: Boxplots representing distribution of genome sizes in bacteria with different combinations of NHEJ machinery present: 1) *NHEJ-*, 2) *Ku only*, 3) *LigD only*, 4) *Non-conventional NHEJ+* (all domains present excluding the conventional LigD), 5) *Non-conventional NHEJ+* (LIG and POL domain present in different proteins and no PE domain), 6) *Non-conventional NHEJ+* (LIG and POL domain present in the same protein and no PE domain) and 7) *Non-conventional NHEJ+* (LIG domain present and no POL domain)

Figure S3: Boxplots representing distribution of rRNA copy numbers in bacteria with different combinations of NHEJ machinery present: 1) *NHEJ-*, 2) *Ku only*, 3) *LigD only*, 4) *Non-conventional NHEJ+* (all domains present excluding the conventional LigD), 5) *Non-conventional NHEJ+* (LIG and POL domain present in different proteins and no PE domain), 6) *Non-conventional NHEJ+* (LIG and POL domain present in the same protein and no PE domain) and 7) *Non-conventional NHEJ+* (LIG domain present and no POL domain)

Figure S4: Boxplots representing distribution of G-C content in bacteria with different combinations of NHEJ machinery present: 1) *NHEJ-*, 2) *Ku only*, 3) *LigD only*, 4) *Non-conventional NHEJ+* (all domains present excluding the conventional LigD), 5) *Non-conventional NHEJ+* (LIG and POL domain present in different proteins and no PE domain), 6) *Non-conventional NHEJ+* (LIG and POL domain present in the same protein and no PE domain) and 7) *Non-conventional NHEJ+* (LIG domain present and no POL domain)

Figure S5: Boxplots representing distribution of G-C content in bacteria with different NHEJ states - *NHEJ-* , *conventional NHEJ+* and *non-conventional NHEJ+* across four phyla. G-C content is significantly higher in organisms with NHEJ as compared to those that lack it in three phyla – A) Proteobacteria, B) Actinobacteria and D) Firmicutes. There is no significance difference in G-C content in C) Bacteroidetes.

Figure S6: Boxplots representing distribution of genome size in bacteria with different NHEJ states - *NHEJ-* , *conventional NHEJ+* and *non-conventional NHEJ+* across four phyla. Genome size is significantly higher in organisms with NHEJ as compared to those that lack it in all four phyla – A) Proteobacteria, B) Actinobacteria, C) Bacteroidetes and D) Firmicutes.

Figure S7: Density plots representing the distribution of A) Genome size, B) rRNA copy number and C) G-C content across bacteria with different NHEJ states- *NHEJ-* (orange), *Conventional NHEJ+* (blue) and *Non-conventional NHEJ+* (pink).

Figure S8: Boxplots representing distribution of rRNA copy number in bacteria with different NHEJ states - *NHEJ-* , *conventional NHEJ+* and *non-conventional NHEJ+* across four phyla. rRNA copy number is significantly lower in organisms with NHEJ as compared to those that lack it in two phyla – A) Proteobacteria, B) Actinobacteria. There is no significant difference in rRNA copy number distributions in C) Bacteroidetes. In D) Firmicutes, organisms harboring NHEJ (conventional and non-

conventional) tend to have significantly higher growth rate than organisms without NHEJ. However, when NHEJ + organisms are segregated into conventional and non-conventional, we find no significant difference in rRNA copy numbers between *NHEJ–* and *conventional NHEJ+* Firmicutes.

Figure S9: A) Density map of phylogenetic ancestral reconstruction of conventional NHEJ presence (blue) and absence (red), a binary trait, represented as posterior probability across 1000 stochastic maps in two Corynebacteriales sub-clades: Mycobacterium and Corynebacterium. B) Density map of phylogenetic ancestral reconstruction of genome size, a continuous trait, using the Brownian motion model. The warmer colours represent smaller genome size and colder colours represent larger genome sizes. The tip colours represent conventional NHEJ presence (blue) and absence (red) in the contemporaneous species. C) Density map of phylogenetic ancestral reconstruction of the growth rate, high (blue) and slow (red), a binary trait, represented as posterior probability across 1000 stochastic maps.

Figure S10: A) Direct primary NHEJ gain at the ancestral node of Firmicutes, Fusobacteria and Tenericutes (shallowest phylogenetic depth). B) Direct primary NHEJ gain at the ancestral node of Actinobacteria. C) Sequential primary gain of LigD followed by Ku to yield NHEJ in sub-clade of Proteobacteria. D) Sequential primary gain of LigD in the ancestral node of Bacteroidetes, followed by the complete gain of NHEJ by the acquisition of Ku in different sub-clades. Tip and Node colours- Red: *NHEJ-* , Blue: *NHEJ+* , Green: *LigD only* , Yellow: *Ku only*.

Figure S11

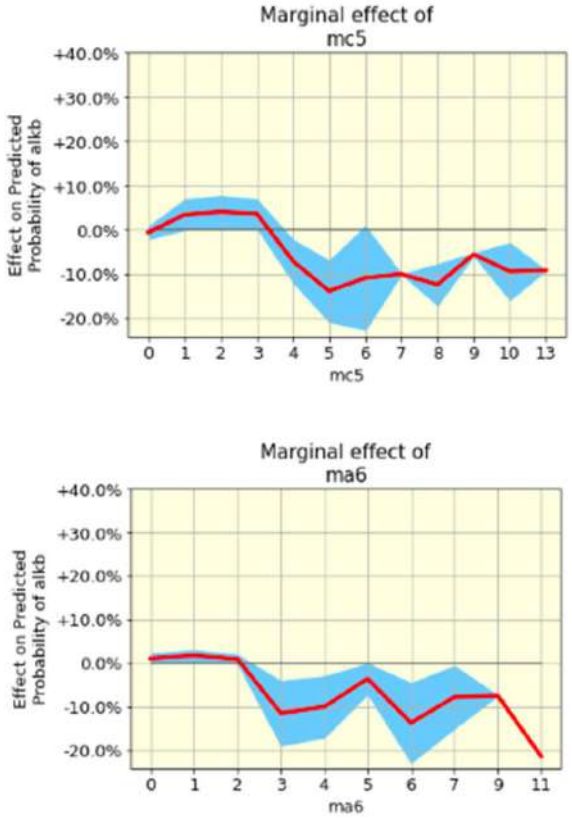

Marginal effect of
mc5
+40.0%
+30.0%
+20.0%
+10.0%
0.0%
-10.0%
-20.0%
Effect on Predicted
Probability of alkb
0 1 2 3 4 5 6 7 8 9 10 13
mc5
Marginal effect of
ma6
+40.0%
+30.0%
+20.0%
+10.0%
0.0%
-10.0%
-20.0%
Effect on Predicted
Probability of alkb
0 1 2 3 4 5 6 7 9 11
ma6

Figure S12

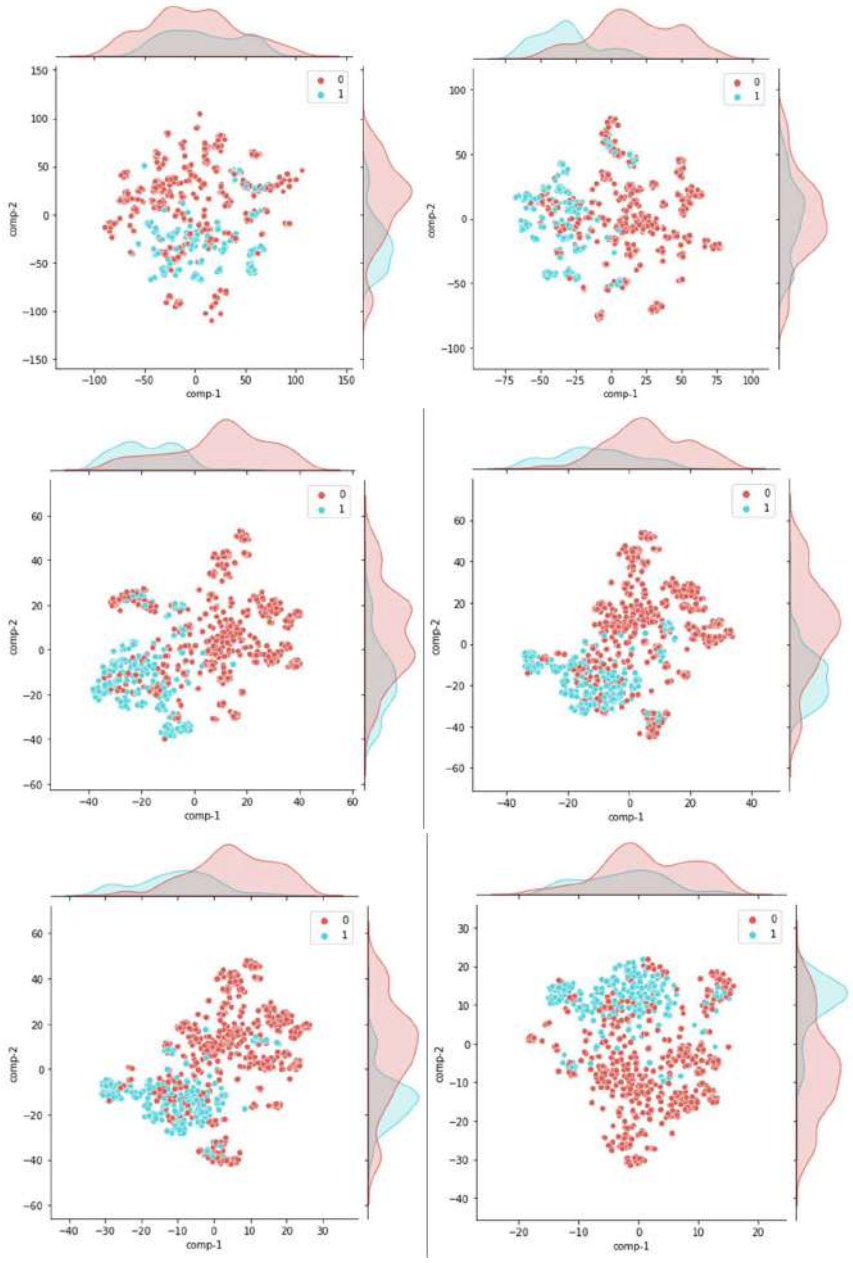


Figure S13

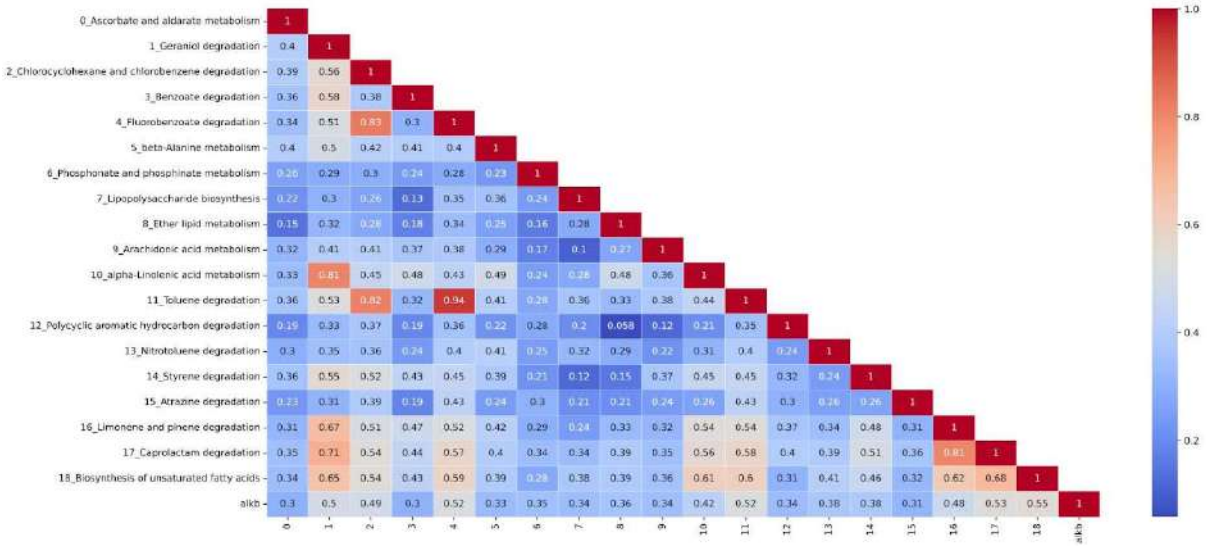

0_Ascorbate and aldarate metabolism
1_Geraniol degradation
2_Chlorocyclohexane and chlorobenzene degradation
3_Benzoate degradation
4_Fluorobenzoate degradation
5_beta-Alanine metabolism
6_Phosphonate and phosphinate metabolism
7_Lipopolysaccharide biosynthesis
8_Ether lipid metabolism
9_Arachidonic acid metabolism
10_alpha-Linolenic acid metabolism
11_Toluene degradation
12_Polycyclic aromatic hydrocarbon degradation
13_Nitrotoluene degradation
14_Styrene degradation
15_Atrazine degradation
16_Limonene and pinene degradation
17_Caprolactam degradation
18_Biosynthesis of unsaturated fatty acids
alkb
1.0
0.8
0.6
0.4
0.2

Figure S14

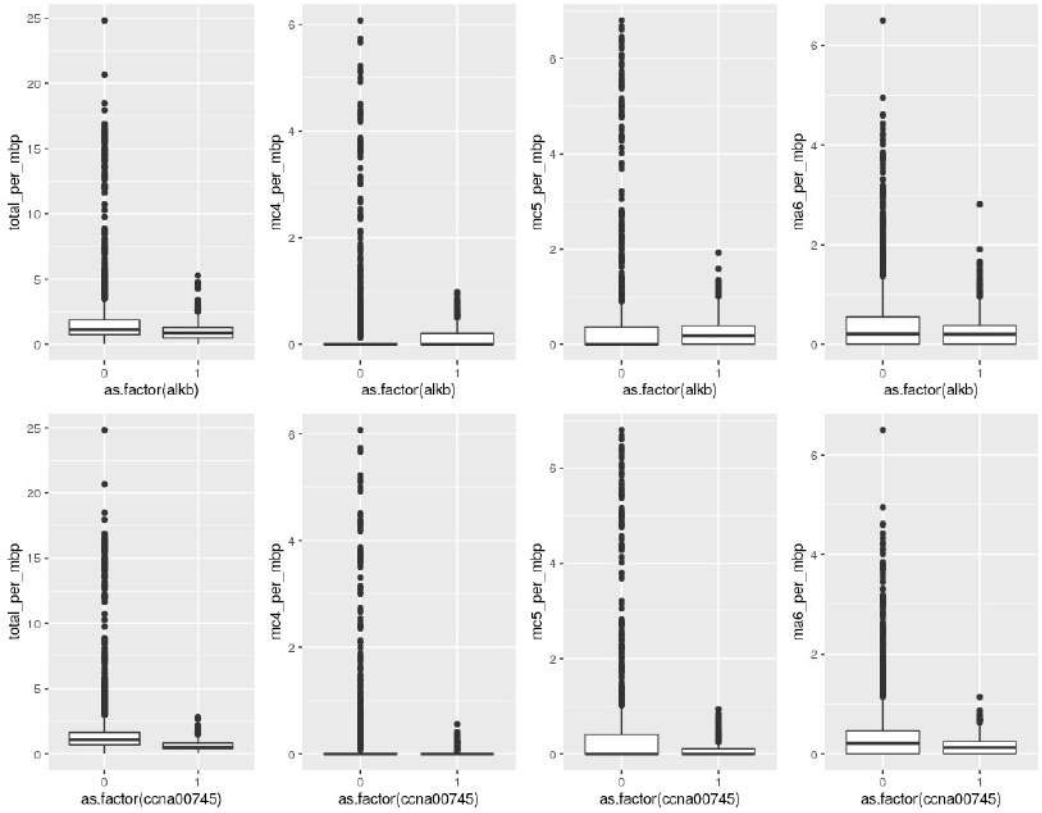
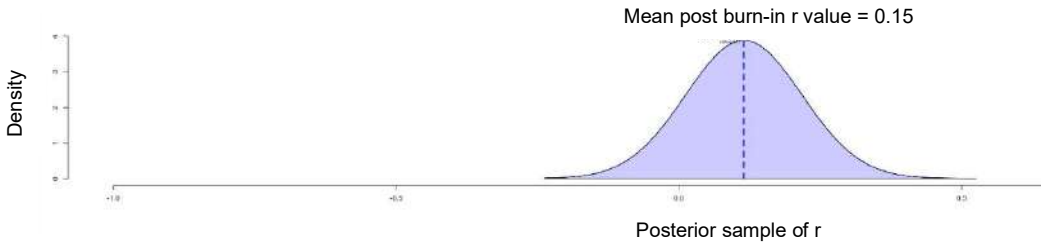

total_per_mbp
mc4_per_mbp
mc5_per_mbp
ma6_per_mbp
as.factor(alkb)
as.factor(ccna00745)

Figure S15

A)



B)

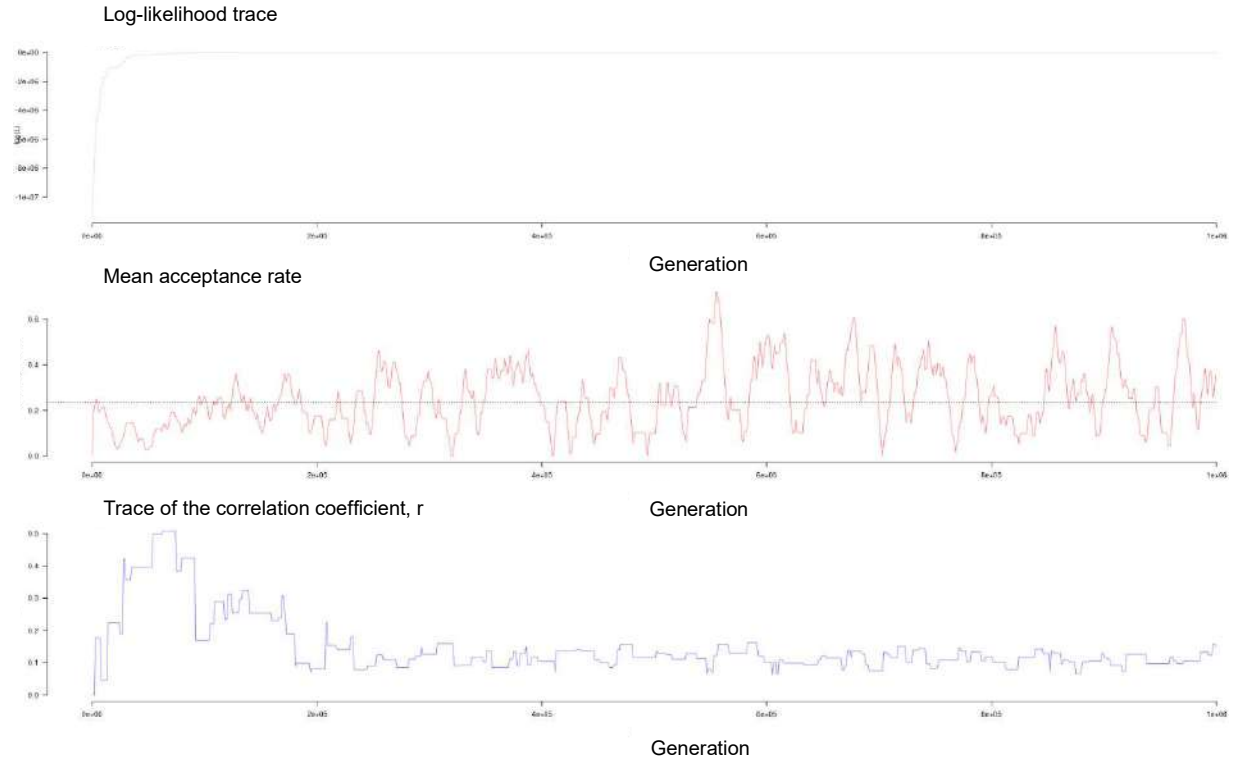
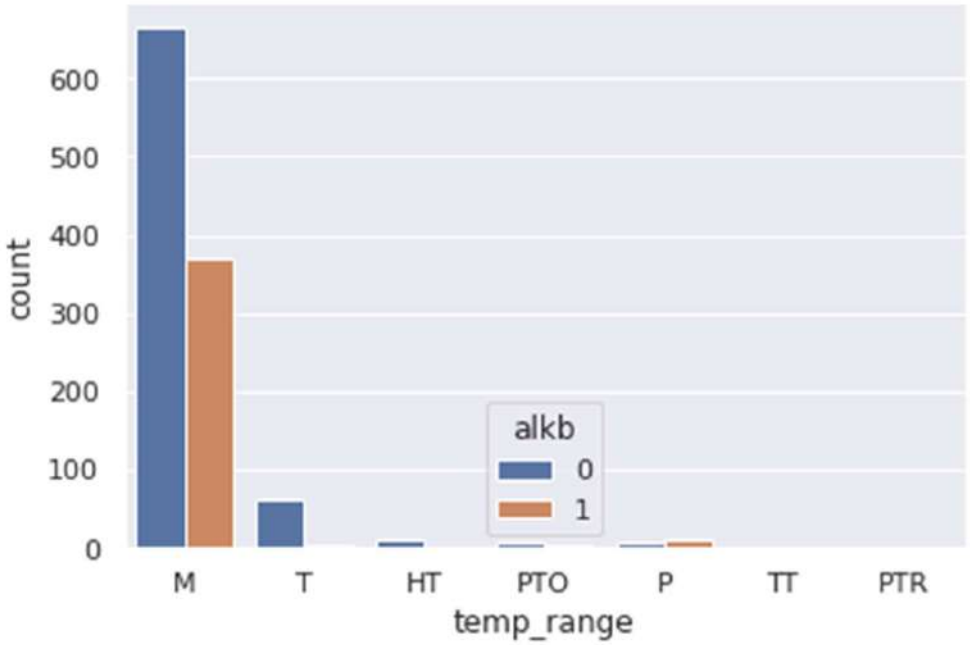


Figure S16


600
500
400
300
200
100
0
count
alkb
0
1
M
T
HT
PTO
P
TT
PTR
temp_range

Chapter 4
Supplementary figure legends

Figure S11 Marginal effect of number of type II 5-cytosine methyltransferases and 6-adenine methyltransferases on predicted probability of *alkB* in bacteria as per the SHAP representation of the XGBoost model.

Figure S12 t-SNE visualisations are dependent on hyperparameter choices. Here shown are visualisations for different values of the perplexity hyperparameter (*left to right*) – 5, 10, 30, 40, 50, 100.

Figure S13 Multicollinearity among 18 metabolic pathways that are correlated with *alkB* with a phi correlation coefficient of >= 0.3

Figure S14 Distribution of type II methyltransferases – 4-methylcytosine, 5-methylcytosine and 6-methyladenine, between organisms harboring and lacking alkB (first row) and those that harbour and lack CCNA00745/COG3826 (second row)

Figure S15 Bayesian analysis of correlated evolution of alkB and presence and absence of type II methyltransferases, assuming a threshold model of evolution. A) Posterior distribution of correlation coefficient, ***r***. B) Bayesian MCMC diagnostics for ***r*** support.

Figure S16 Bar plots showing the distribution of temperature profiles of bacteria sampled in the dataset used for the study to understand alkB presence and absence. M: Mesophile, T: Thermophile, HT: Hyperthermophile, PTO: Psychrotolerant, P: Psychrophile, TT: Thermotolerant, PTR: Psychrotrophic