Evolutionary analysis of bacterial Non-Homologous End Joining Repair

Mohak Sharda[1,2], Anjana Badrinarayanan[1,*], Aswin Sai Narain Seshasayee[1,*]

[1]National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, Karnataka, India, 560065

[2]The University of Trans-Disciplinary Health Sciences & Technology (TDU), Bengaluru, Karnataka, India, 560064

*For correspondence: anjana@ncbs.res.in; aswin@ncbs.res.in.

Abbreviations: NHEJ- Non Homologous End Joining, DSB- Double Strand Break, DNA- Deoxyribonucleic Acid, rRNA- ribosomal Ribonucleic Acid, LIG- Ligase domain, POL- Polymerase domain, PE- Phosphoesterase domain, CDS- Coding DNA sequence, HAR- Horizontally Acquired Region, HGT- Horizontal Gene Transfer, AIC- Akaike Information Criterion, BPIC- Bayesian Predictive Information Criterion, PSRF- Proportional Scale Reduction Factor, ML- Maximum Likelihood, RJMCMC- Reverse Jump Markov Chain Monte Carlo, GS- Genome size, GR- Growth Rate

Keywords: DNA repair, phylogeny, double-strand breaks, non-homologous end joining, recombination, bioinformatics, bacteria, comparative genomics

Abstract

DNA double-strand breaks (DSBs) are a threat to genome stability. In all domains of life, DSBs are faithfully fixed via homologous recombination. However, recombination requires the presence of an uncut copy of duplex DNA that can be used as a template for repair. Alternate to this, in the absence of a template, eukaryotes and some prokaryotes also utilize Non-homologous end joining (NHEJ) repair. This pathway, or variations of it, can be error-prone. However, it avoids the lethality of a DSB, making NHEJ an important form of repair. Although ubiquitously found in eukaryotes, NHEJ is not universally present in bacteria. It is unclear as to why many prokaryotic systems lack this pathway. To understand what could have led to the current distribution of NHEJ in bacteria, we carried out comparative genomics and phylogenetic analysis across ~6000 sequenced genomes. Our results show that this repair trait is sporadically distributed across the phylogeny, with a few clades that are completely devoid of it. Ancestral reconstruction suggests that NHEJ was absent in the eubacterial ancestor, followed by primary independent gains as well as secondary losses and gains, in different clades across the bacterial history. Further, our analysis suggests that this pattern of occurrence is consistent with the correlated evolution of NHEJ with key genome characteristics of genome size and growth rates; NHEJ presence in bacteria is associated with large genome sizes and / or slow growth rates, with the former being the dominant correlate. Given the central role these traits may 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.

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 ionising radiation and reactive oxygen, while generating diversity for natural selection to act on (Tenaillon, Denamur, and Matic 2004), also adversely affect viability and could lead to diseases including cancer (Tomasetti, Li, and Vogelstein 2017; Srinivas et al. 2019). Therefore, most cellular life forms invest in mechanisms that repair damaged DNA including double strand breaks (DSBs).

Two major mechanisms of repair of DNA double strand breaks are homologous recombination and non-homologous end joining (NHEJ). Homologous recombination based repair requires a homologous copy of the DNA around the damage site for repair to occur. In contrast, NHEJ, the subject of the present work, directly ligates the double strand break after detecting and binding the break ends (Aniukwu, Glickman, and Shuman 2008; Bhattarai, Gupta, and Glickman 2014). 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 (Bétermier, Bertrand, and Lopez 2014). 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.

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 (Aravind and Koonin 2001; Doherty, Jackson, and Weller 2001), and its prevalence still remains to be systematically elucidated. Studies in bacteria have found that NHEJ can participate in repair, and also contribute to mutagenesis such as in stationary phase (Bhattarai, Gupta, and Glickman 2014; Paris et al. 2015). Experiments in *Mycobacterium* and *Pseudomonas* have shown that bacterial NHEJ repair machinery consists primarily of two proteins – the homodimeric Ku and the three-domain LigD harbouring ligase (LIG), polymerase (POL) and phosphoesterase (PE) domains (Aniukwu, Glickman, and Shuman 2008; Stephanou et al. 2007; Zhu and Shuman 2010; Della et al. 2004). In contrast, *Bacillus subtilis* encodes Ku along with a two-domain (LIG and POL) LigD (Moeller et al. 2007). In exceptional cases, LigC, which contains only the LIG domain, can carry out repair in the absence of LigD (Bhattarai, Gupta, and Glickman 2014; Matthews and Simmons 2014; Aniukwu, Glickman, and Shuman 2008). Though *Escherichia coli* does not encode NHEJ, expression

of *M. tuberculosis* Ku and LigD renders E. coli NHEJ proficient (Malyarchuk et al. 2007).

Studies in the early 2000s, with a small number of genomes, suggested that the distribution of NHEJ in bacteria could be patchy (Bowater and Doherty 2006). Collectively, what does it mean for bacteria like *Escherichia coli* to not code for NHEJ and others like *Mycobacterium tuberculosi*s to harbour 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 bioinformatic sequence searches of Ku and LigD domains in the genomic sequences of ~6000 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.

Methods

Data

All genome information files for ~6000 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).

Identification of NHEJ repair proteins

Initial blastp (Altschul et al. 1990) was run using each of the four protein domain sequences – Ku and LigD (LigD-LIG, LigD-POL and LigD-PE) from Pseudomonas aeruginosa PAO1 as the query sequence against the UniProtKB database (“UniProt: A Worldwide Hub of Protein Knowledge” 2019) with an E-value cut-off of 0.0001. The top 250 full length sequence hits were downloaded from Uniprot for both Ku and LigD domains respectively. A domain multiple sequence alignment was made using phmmer -A option with the top 250 hits as the sequence database and *P. aeruginosa* domain sequences as the query. An hmm profile was built using the hmmbuild command for the multiple sequence alignment obtained in the previous step. To find domain homologues, hmmsearch command with an E-value cut-off of 0.0001 was used with the hmm profile as the query against a database of 5973 bacterial genome sequences (Supplementary table 1). These homolog searches were done using HMMER package v3.3 (Finn, Clements, and Eddy 2011). An in-house python script was written to

extract the results and assign organisms with Ku and LigD domains.

Bacteria were assigned into five broad categories, based on the status of NHEJ components – No-NHEJ, Ku-only, LigD-only, conventional-NHEJ and non-conventional NHEJ. Conventional NHEJ was said to be present in bacteria harbouring Ku and LigD having LIG, POL and PE domains in the same protein. Non-conventional NHEJ included bacteria with Ku and at least one of the following: a) all LigD domains present in different combinations in different proteins, b) just the LIG and POL domains present in the same or, c) different proteins and, d) 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 No-NHEJ state if Ku domain wasn't detected either; a Ku-only state if Ku domain was detected. (Figure 1)

Ku and LigD neighbourhood 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 10 genes upstream or downstream inclusive of both strands. The organisation 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.

Identification of rRNA sequences

A 16S rRNA sequence database was downloaded from the GRD (Genomic-based 16S ribosomal RNA database) website (https://metasystems.riken.jp/grd/). A multiple sequence alignment and a profile of the database was made using FAMSA v1.1 (Deorowicz, Debudaj-Grabysz, and Gudyś 2016) and hmmbuild (Finn, Clements, and Eddy 2011) respectively. To detect 16S rRNA homologues in our database of 5973 bacteria, we used nhmmer (Wheeler and Eddy 2013) with an E-value cut-off 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.

Calculation of genome sizes, growth rates and G-C content

In-house python scripts were written to calculate the genome sizes, growth rates and G-C content of all

bacteria in our dataset (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 growth rate and the number of rRNA operons harboured in it's genome (Vieira-Silva and Rocha 2010; Roller, Stoddard, and Schmidt 2016; Klappenbach, Dunbar, and Schmidt 2000; Gyorfy et al. 2015). Therefore, as an estimate of growth rate for a given bacteria, rRNA copy number was taken as a proxy.

Genome size randomization analysis

For this analysis, we only considered 920 organisms harbouring conventional NHEJ. We used the total number of coding DNA sequences (CDS) as the proxy for genome size. Since the chances of picking up a gene coming from a larger genome size are more than a smaller genome, we drew 920 genes at random from a pool of all the genes coming from 5973 organisms in our dataset and every time the genome CDS size (genome size) from which the gene came from was noted. For this iteration, the median genome size was calculated. This was repeated for 100 iterations and a distribution of median genome size was obtained. The non-parametric wilcoxon rank sum test was used to compare this random distribution to the median genome size of NHEJ harbouring bacteria.

Horizontal gene transfer analysis

We used Alien Hunter v1.7 (Vernikos and Parkhill 2006) with default options to predict horizontally acquired regions (HARs) - based on their oligonucleotide composition - across bacterial genomes present in our dataset. 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 if both Ku and LigD were present in the predicted HARs.

Phylogenetic tree construction

For the construction of 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 Ns 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 non-conventional NHEJ, we treated all its four sub-categories, as described in the previous sections, separately, when including bacteria at the genus level. A multiple sequence alignment (MSA)

was built using FAMSA 1.1 (Deorowicz, Debudaj-Grabysz, and Gudyś 2016) with default options. The conserved regions relevant for phylogenetic inference were extracted from the MSA using BMGE v1.12 (Criscuolo and Gribaldo 2010). After the manual detection of the alignment, one spurious sequence was removed and the MSA was built again for 969 bacteria (Supplementary table 2). Using IQTREE v1.6.5 (Nguyen et al. 2015), a maximum likelihood based phylogenetic tree was built with the best model chosen as SYM+R10 (LogL= -118152.5876, BIC= 249946.0604). ModelFinder (-m MF option) (Kalyaanamoorthy et al. 2017) was used to choose the best model for the tree construction compared against 285 other models . Branch supports were assessed using both 1000 ultrafast bootstrap approximations (-bb 1000 -bnni option) (Hoang et al. 2018) and SH-like approximate likelihood ratio test (-alrt 1000 option) (Guindon et al. 2010).

A similar approach was used to build a phylogeny of 1403 (Supplementary table 3) organisms, non-redundant at the species level, comprising of just two NHEJ states- 1) No NHEJ and 2) conventional NHEJ.

NHEJ ancestral state reconstruction analysis

To trace the evolutionary history, 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 (Bollback 2006) with 1000 simulations provided as make.simmap() in the R package phytools v0.6-44 (Revell 2012). 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 were 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 AIC weights, ARD model ($w\text{-}AIC_{ARD}$ =1, $w\text{-}AIC_{SYM}$= 0, $w\text{-}AIC_{ER}$= 0) was chosen as the best fit with unequal forward and backward rates for each character state transition. Finally, Q was sampled 1000 times from the posterior probability distribution of Q using MCMC and 1000 stochastic maps were simulated conditioned on each sampled value of Q. This strategy was used to reconstruct ancestral states for both phylogenies comprising of 969 and 1403 organisms as described in the previous sections.

Ancestral states for the continuous trait genome size were reconstructed using fastAnc() employed in phytools v0.6-44 based on the brownian motion model. This model was found to be the better fit model as compared to multiple rate model ($BPIC_{brownian}<BPIC_{stable}$ and PSRF approaching <1.1 well within 1000000 iterations), assessed using StableTraits (Elliot and Mooers 2014). 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 growth rate 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.

NHEJ and genome characteristics phylogenetic comparative analysis

The two genome characteristics – genome size and growth rate – 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 (Freckleton, Harvey, and Pagel 2002) and Blomberg's K (Blomberg, Garland, and Ives 2003) – were used for detecting the impact of shared ancestry, for genome size and growth rate across bacteria, using the phylosig() routine in phytools R package (Revell 2012). A phylogenetic ANOVA, employed in phytools R package v0.6-44 (Revell 2012), was carried out with 1000 simulations and Holm-Bonferonni correction to control for family-wise error rate, based on a method by Garland T. (Garland et al. 1993), to compare the genome characteristics in a phylogenetically controlled manner. Please note that we $\log_{10}$ transformed genome sizes and rRNA copy number for all the analysis.

Correlated evolution analysis

Two relationships – (NHEJ repair and genome size) and (NHEJ repair and growth rate) - were quantified using a statistical framework. To test if changes in genome characteristics occur independently of NHEJ or are these changes more (or less) likely to occur in lineages with (or without) NHEJ , we considered two models of evolution- independent and dependent. In the independent model, both the traits were allowed to evolve separately on a phylogenetic tree i.e non-correlated evolution. In the dependent model, the two traits were evolved in a non-agnostic manner i.e correlated evolution. The NHEJ repair trait had two character states – no NHEJ repair (0) and conventional NHEJ repair (1).

Genome sizes, a continuous trait, were converted into binary state as well. For this, we computed the mean genome size of organisms with 0 or 1 NHEJ repair state, and assigned a "lower" state (0) 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 growth rates (rRNA copy number) into a binary state.

A continuous-time Markov model approach was used to investigate correlated evolution between NHEJ repair and the genome characteristics. First, we used the maximum likelihood (ML) approach (Pagel 1994) to calculate log-likelihoods for the two models of evolution per trait pair – 1) NHEJ repair and genome size; 2) NHEJ repair and growth rate. A likelihood ratio statistic (LR) was calculated for both comparisons, followed by a Chi-square test to assess if 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 8 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 our analysis robust, we used the bayesian rjmcmc (reverse jump markov chain monte carlo) approach to calculate the marginal log-likelihoods of the independent and dependent models of evolution (Pagel and Meade 2006). 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 (Pagel and Meade 2006) 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 5050000 iterations, sampling every 1000th iteration with a burn-in of 50000. For the estimation of marginal likelihood, stepping stone sampler algorithm was used where the number of stones were set to 100 and each stone was allowed to run for 10000 iterations.

Phylogenetic logistic regression analysis

A method developed by Ives and Garland (Ives and Garland 2010), was used to carry out the phylogenetic logistic regression analysis provided in the R package phylolm v2.6 (Tung Ho and Ané 2014) as the subroutine phyloglm() with method=logistic_IG10. NHEJ repair was the binary dependent variable with two states - no NHEJ (0) and conventional NHEJ repair (1). The two independent continuous variables were genome size (GS) and growth rate (GR). We tested for three

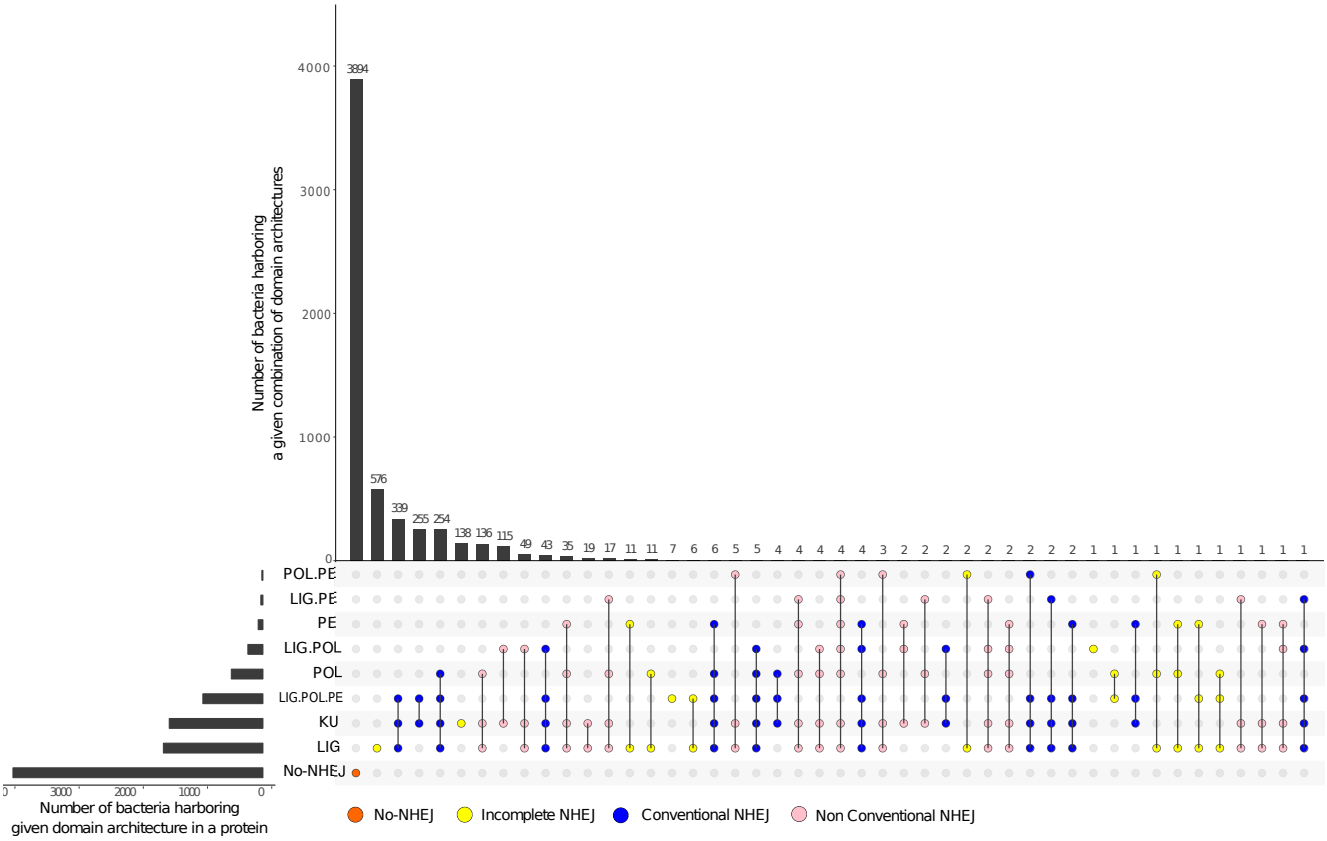

Figure 1: *Distribution of NHEJ components in bacteria. An UpSet plot depicting number of bacteria harboring a certain type of domain architecture per protein and number of bacteria harboring a combination of domain architecture in their respective genomes is shown for 5973 analysed genomes. No NHEJ (orange), Ku only (yellow), LigD only (yellow), Conventional NHEJ+ (blue) and non-conventional NHEJ+ (pink).*

models – 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.

Results

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 Methods). We defined bacteria encoding Ku and the complete, three-domain version of LigD as those harbouring a *conventional* NHEJ system. Organisms lacking the POL

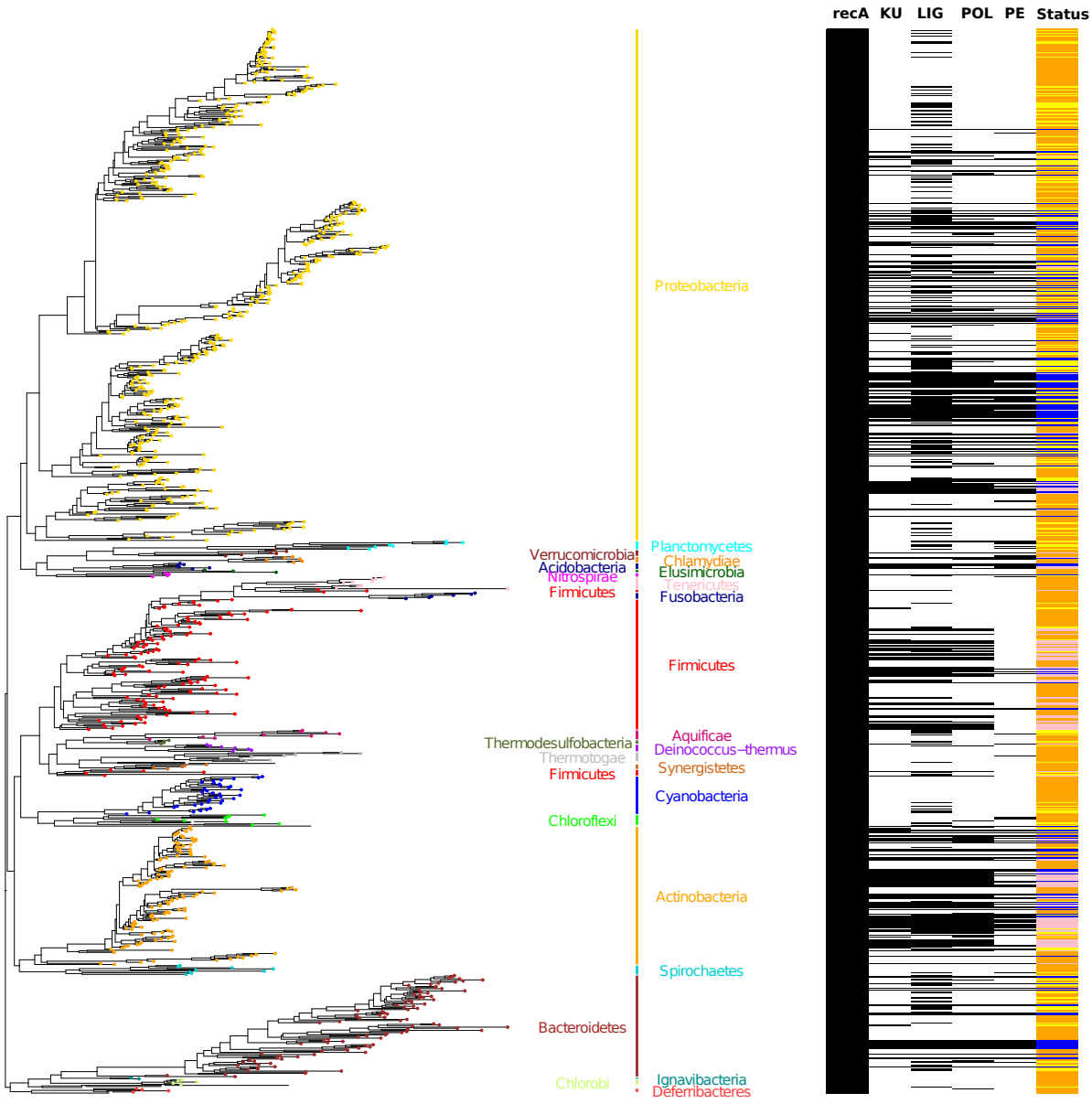

Figure 2: *NHEJ is sporadically distributed across bacteria. 16S rRNA based species phylogenetic tree of 969 bacterial species (left) with presence/absence matrix of RecA, KU, LIG, POL and PE domains (right). Tip labels, representing bacteria, are coloured according to the phylum names and bars (right of the phylogenetic tree) that the given tip belongs to. The last column of the matrix represents NHEJ status corresponding to the species tip mapped to the phylogenetic tree. Status: Orange – No-NHEJ; Yellow – Incomplete NHEJ; Blue – Conventional NHEJ; Pink – Non-Conventional NHEJ.*

and/or the PE domains of LigD, and those encoding these domains in separate proteins, were defined as those carrying *non-conventional* NHEJ (Figure 1).

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 the conventional NHEJ. 75% bacteria harbouring conventional NHEJ coded for Ku and LigD in a 10kb vicinity of each other, with 60% organisms carrying Ku and LigD in an operon. Most bacteria (84%) harbouring NHEJ coded for a single copy of Ku, whereas the remaining coded 2-8 Ku copies in their genomes. 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. (Figure 1; Supplementary table 1).

NHEJ was not restricted to specific bacterial classes (Figure 2), and was found in 10 classes. We found a significant enrichment of *conventional* NHEJ in Proteobacteria (Fisher's Exact test, P = 3.8 x 10^{-5}, odds ratio: 2.04) and Acidobacteria (Fisher's Exact test, P = 5 x 10^{-2}, odds ratio: 5.286). All Bacteroidetes with NHEJ harbour a *conventional* NHEJ, although they were not significantly enriched (Fisher's Exact test, P = 0.14, odds ratio: 1.38). In contrast, non-conventional NHEJ repair was significantly over-represented in Firmicutes (Fisher's Exact test, P =1.23 x 10^{-13}, odds ratio: 6) and Actinobacteria (Fisher's Exact test, P = 2.12 x 10^{-15}, odds ratio: 6.14). Twenty-four phyla did not include any NHEJ positive organism. (Supplementary table 4).

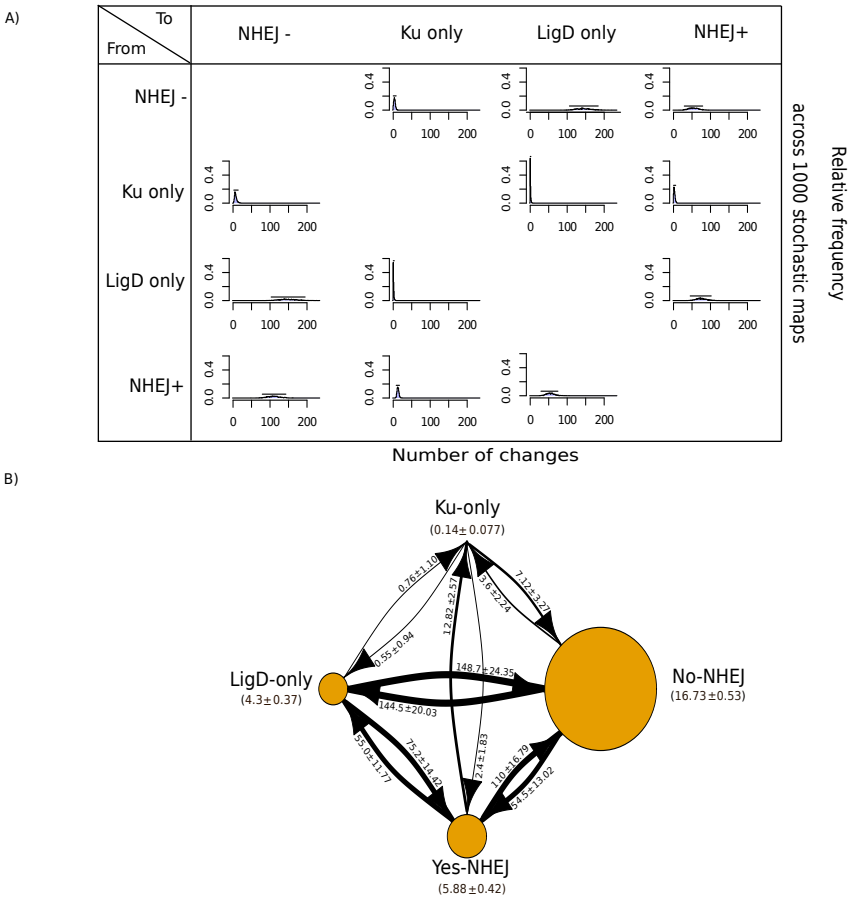

Figure 3: *Transitions to a Ku-only state are rare. A) A matrix depicting relative frequency of number of changes of a state transition type across 1000 stochastic maps. B) A state transition diagram depicting the number of transitions between two given states and the time spent in each state during 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.*

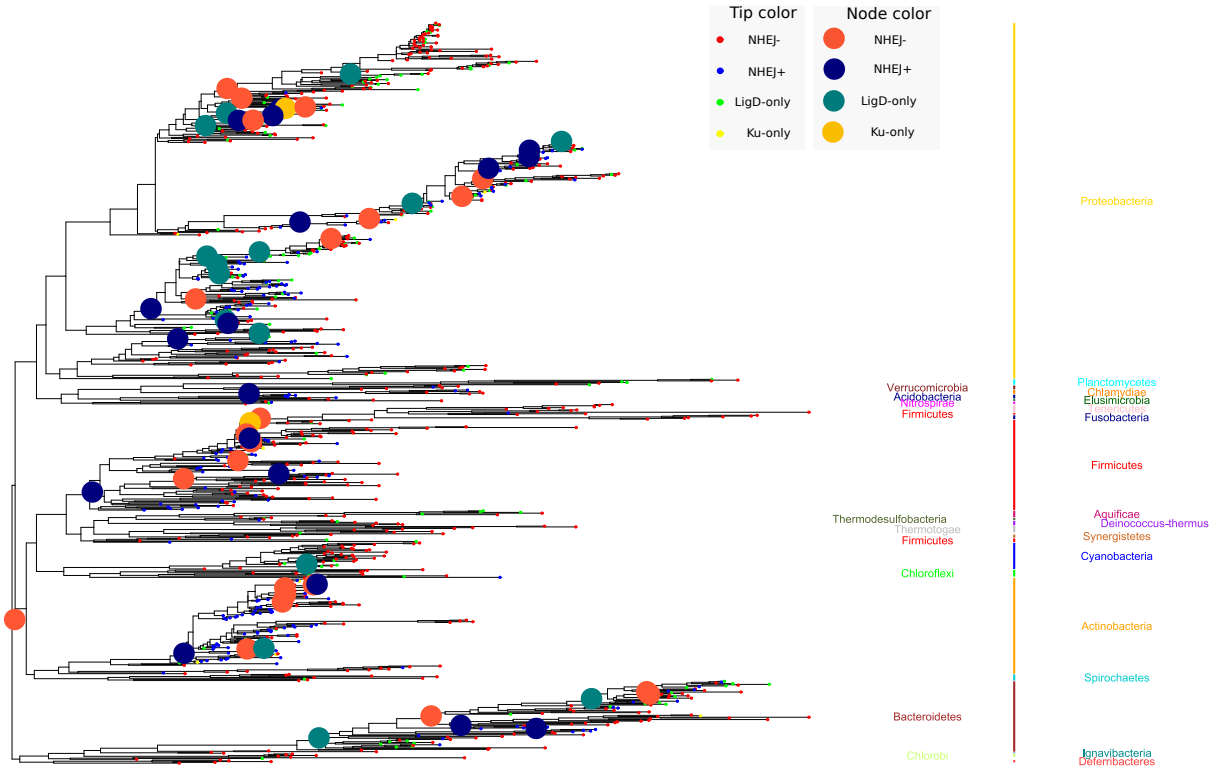

Figure 4: *NHEJ was gained and lost multiple times through evolution. A trace of the evolutionary history of the two-component NHEJ system across bacteria. The tip and node labels are coloured according to the NHEJ states: Red – NHEJ- ; Yellow – Ku-only ; Green – LigD-only ; Blue – NHEJ+ (Conventional and Non-conventional). NHEJ state for nodes is shown only when the posterior probability support is greater than 70%; interpreted as change in NHEJ state at that node as compared to shallower phylogenetic depths. Major primary sequential NHEJ gains occured in Bacteroidetes and sub-clades in Proteobacteria. There was a direct primary gain at actinobacteria at a relatively shallow depth (posterior probability > 0.7), however, it should be noted that a node before the direct gain, supports the acquisition via LigD if cut-offs are relaxed (posterior probability > 0.5). Major direct primary gains occured at the ancestral nodes of Firmicutes, Acidobacteria and different sub-clades in Proteobacteria.*

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 we defined four discrete character states – *Ku only*, *LigD only* (conventional and non-conventional), *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 (Figure 3A) and the distribution of the total time spent in each state (Figure S1). We performed this analysis for a set of 969 genomes in which each genus was represented once for each state (see Methods).

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 5). We assigned a major *primary* gain to an internal ancestral node if (a) all nodes leading to it from the root had *NHEJ-* state; (b) the pp of either *NHEJ+*, *LigD only* or *Ku only* at that ancestral node was >=0.7; (c) a gain of *LigD only* or *Ku only* was followed by a transition to *NHEJ+* and; (d) 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 sub-clades of Proteobacteria (Supplementary table 5). It follows that the common eubacterial ancestor likely did not have NHEJ (Figure 4).

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 (Figure 3B). The most common transition from an *NHEJ-* state was to a *LigD only* state, followed by that to a *NHEJ+* state. Transition from *NHEJ-* to *Ku only* was negligible. In the reverse direction, a one-step loss of both Ku and LigD is the most likely. However, *Ku only* state is rare.

A one-step transition from *NHEJ-* to *NHEJ+* is likely through horizontal gene transfer. 60 bacterial genomes belonging to the phyla Alpha (in particular the Rhizobiales) - and Beta-proteobacteria, and Streptomycetales carried their NHEJ components on plasmids (Supplementary table1). However, based on abnormal word usage statistics (see Methods), we couldn't 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 may be an underestimate (Figure 4).

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

NHEJ occurrence is associated with genome size, growth rate and G-C content

Recently, Ku-encoding organisms were shown to have higher genomic G+C content (Weissman, Fagan, and Johnson 2019). 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. held true for *NHEJ+* states as defined in our study (Figure 5A, Supplementary S4-S5, S7C) (Weissman, Fagan, and Johnson 2019). Along with this, we tested two additional characteristics: genome size and growth rate (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

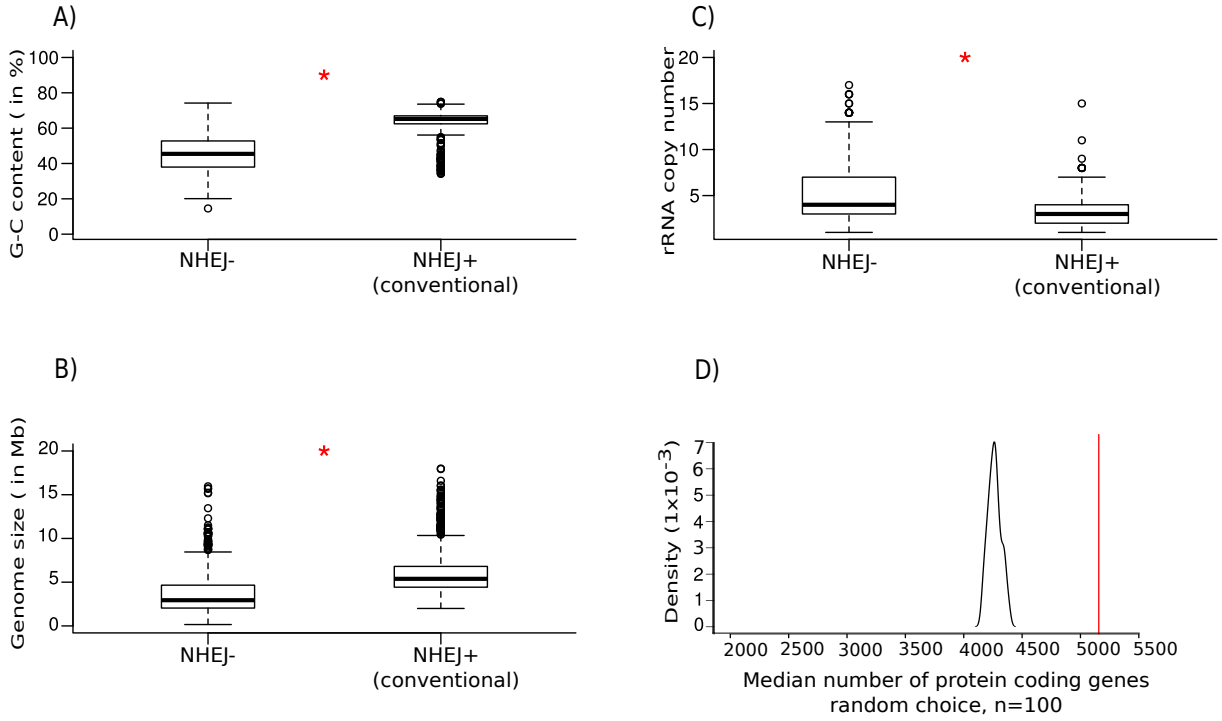

Figure 5: *NHEJ presence and absence is associated with genome size, growth rate and G-C content. A) Boxplot comparing the distribution of G-C content between NHEJ– and conventional NHEJ+ bacteria. B) Boxplot comparing the distribution of genome size between NHEJ- and conventional NHEJ+ bacteria. C) Boxplot comparing the distribution of rRNA copy number between NHEJ- and NHEJ+ bacteria. D) A density distribution plot depicting the distribution of genome size (median number of protein coding sequences) expected by a random distribution where the probability of having NHEJ is linearly proportional to genome size (black) and the median genome size of organisms harboring NHEJ (red).*

fidelity recombination-based repair. We restricted these analyses to conventional NHEJ-harbouring bacteria as a proxy for repair proficiency, and compared them to *NHEJ*- genomes. Data including non-conventional NHEJ are shown in Supplementary Figures S2 and S3.

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, $P < 10^{-15}$; Figure 5B, S7A), and significantly larger than that expected by a random distribution in which the probability of having NHEJ is linearly proportional to genome size (Wilcoxon rank sum test, $P < 10^{-15}$, across 100 simulations; Figure 5D). This relationship was found to be true within the phylum Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes as well (Supplementary figure S6).

In addition, bacteria harbouring NHEJ were found to have significantly fewer rRNA copies (median = 3), and by inference slower growth rates, than bacteria without NHEJ (median = 4; Wilcoxon rank-sum

Table 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	logLikelihood (Independent model)	Marginal logLikelihood (Independent model)	logLikelihood (Dependent model)	Marginal logLikelihood (Dependent model)	Likelihood Ratio Test (LRT) statistics	Bayes factor (>2 = better fit)
Maximum likelihood	NHEJ state and Genome size	-1059.97	-	-1010.002	-	LR=99.94 Pvalue<0.001	-
Bayesian rjmcmc	NHEJ state and Genome size	-	-1110.14	-	-1042.25	-	135.78
Maximum likelihood	NHEJ state and Growth rate	-975.301	-	-950.01	-	LR=25.29 Pvalue<0.001	-
Maximum likelihood	NHEJ state and Growth rate	-	-1051.529	-	-985.579	-	131.9

Table 2: Phylogenetic logistic regression for three models, based on a phylogenetic tree of 1403 species of bacteria harbouring either no NHEJ or conventional NHEJ.

Model	AIC value	Pen.logLik	Alpha	Coefficient estimate (CE)	P-value
NHEJ ~ Genome size (GS)	809.5	-383.3	13.88	6.7×10^{-7}	10^{-15}
NHEJ ~ Growth rate (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}

test; $P < 10^{-15}$; Figure 5C). While 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 figure S7B). 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 figure 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 genome sizes and growth rates in the dataset (see Methods). These measures suggest that phylogenetic coherence is significantly greater than random expectations for both the genome characteristics (Supplementary Table 6). Therefore, the distributions of genome size between bacteria with different NHEJ status were compared while accounting for the statistical non-independence of closely related taxa (see Methods). We found a significant difference in $\log_{10}$(genome sizes) 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 *non-conventional 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 maximum likelihood as well as Bayesian approaches to test if the observed associations of conventional NHEJ individually with large genomes and slow growth rates is indicative of dependent or independent evolution of these traits on the phylogenetic tree (see Methods). Both suggested that the phylogenetic data fit models of evolution in which conventional NHEJ presence or absence and genome size or growth rate are evolving in a correlated manner (Table 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 genome size is the stronger correlate (see Methods; Table 2)

As a case study where the association of both genome size and growth rate 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 genome size in the ancestor of Corynebacteriales, followed by an NHEJ gain. While *Mycobacterium* retained NHEJ, *Corynebacterium* lost the machinery along with a decrease in genome size. Using a similar analysis, growth rate mapped to this sub-clade revealed an increase in rRNA copy number in *Corynebacterium*. (Figure S9; see methods)

Discussion

Taking advantage of the availability of a large number of genomes, we confirm the non-ubiquitous nature of NHEJ across the bacterial domain with 22% bacteria coding for it. At the taxa level, while some phyla retain this sporadicity, others are devoid of an NHEJ repair. A trace of the evolutionary history suggested that NHEJ was most likely absent in the eubacterial ancestor. It was instead gained independently multiple times in different bacterial lineages. However, these primary gains were not sufficient to stabilize the repair states in sub-clades, since a large number of secondary losses and gains were observed across the phylogeny.

Our analysis suggests that there are two common methods to arrive at an *NHEJ+* state- a) the most

common way to gain NHEJ was by the acquisition of LigD followed by Ku and b) 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 ($pp_{Proteobacterial\text{-}subclade}$=0.875 and $pp_{Bacteroidetes}$=0.7; Figure S10C and S10D). 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 horizontal gene transfer (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 – (a) the regions getting horizontally transferred have the same compositional biases in the donor and the recipient cells and/or (b) 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; Figure S10A and S10B).

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 non-functional in repair or in some cases could even block the access of break ends for recombination-based repair (Sinha et al. 2009; Gupta et al. 2011; Zhang et al. 2012). In 138 organisms, where Ku is retained in the absence of LigD, it's 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 (Aniukwu, Glickman, and Shuman 2008).

To understand the selection pressures that might play a role in shaping the evolutionary pattern of NHEJ, we studied the genome characteristics associated with this DNA repair. Under the event of a DSB, the unavailability of a template would prevent the highly accurate homologous recombination repair to act. Therefore, one might expect a higher selection pressure of maintaining NHEJ in bacteria with larger genome sizes and slower growth rates. In line with our hypothesis, we found that the organisms possessing conventional NHEJ tend to have significantly larger genome size and slower

growth rate as compared to those that are devoid of it. A phylogenetically controlled test suggested that this association held true for genome size and not growth rate. We note that phyloANOVA used here for hypothesis testing assumes normality. However, our data for genome size and rRNA copy number 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, Maximum Likelihood and Bayesian inferences showed that there is a correlated evolution of NHEJ with genome size and growth rate along the phylogenetic tree. Considering the two variables together, we found that genome size is the better correlate of the *NHEJ+* state. Therefore, we conclude that the *NHEJ+* state is strongly associated with genome size and to a much smaller extent with growth rate throughout its evolution.

Our study highlights the evolutionary trajectory of NHEJ and central characteristics that may have determined it's sporadic distribution. DSB repair, including NHEJ, has been implicated in shaping bacterial genomes through mutagenesis (Paris et al. 2015), HGT (Popa et al. 2011) and their effect on genomic GC content (Weissman, Fagan, and Johnson 2019). 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.

References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Aniukwu, Jideofor, Michael S. Glickman, and Stewart Shuman. 2008. "The Pathways and Outcomes of Mycobacterial NHEJ Depend on the Structure of the Broken DNA Ends." *Genes & Development* 22 (4): 512–27. https://doi.org/10.1101/gad.1631908.

Aravind, L., and Eugene V. Koonin. 2001. "Prokaryotic Homologs of the Eukaryotic DNA-End-Binding Protein Ku, Novel Domains in the Ku Protein and Prediction of a Prokaryotic Double-Strand Break Repair System." *Genome Research* 11 (8): 1365–74. https://doi.org/10.1101/gr.181001.

Bétermier, Mireille, Pascale Bertrand, and Bernard S. Lopez. 2014. "Is Non-Homologous End-Joining Really an Inherently Error-Prone Process?" *PLOS Genetics* 10 (1): e1004086. https://doi.org/10.1371/journal.pgen.1004086.

Bhattarai, Hitesh, Richa Gupta, and Michael S. Glickman. 2014. "DNA Ligase C1 Mediates the LigD-Independent Nonhomologous End-Joining Pathway of Mycobacterium Smegmatis." *Journal of Bacteriology* 196 (19): 3366–76. https://doi.org/10.1128/JB.01832-14.

Blomberg, Simon P., Theodore Garland, and Anthony R. Ives. 2003. "Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits Are More Labile." *Evolution* 57 (4): 717–45. https://doi.org/10.1111/j.0014-3820.2003.tb00285.x.

Bollback, Jonathan P. 2006. “SIMMAP: Stochastic Character Mapping of Discrete Traits on Phylogenies.” *BMC Bioinformatics* 7 (1): 88. https://doi.org/10.1186/1471-2105-7-88.

Bowater, Richard, and Aidan J. Doherty. 2006. “Making Ends Meet: Repairing Breaks in Bacterial DNA by Non-Homologous End-Joining.” *PLOS Genetics* 2 (2): e8. https://doi.org/10.1371/journal.pgen.0020008.

Criscuolo, Alexis, and Simonetta Gribaldo. 2010. “BMGE (Block Mapping and Gathering with Entropy): A New Software for Selection of Phylogenetic Informative Regions from Multiple Sequence Alignments.” *BMC Evolutionary Biology* 10 (1): 210. https://doi.org/10.1186/1471-2148-10-210.

Della, Marina, Phillip L. Palmbos, Hui-Min Tseng, Louise M. Tonkin, James M. Daley, Leana M. Topper, Robert S. Pitcher, Alan E. Tomkinson, Thomas E. Wilson, and Aidan J. Doherty. 2004. “Mycobacterial Ku and Ligase Proteins Constitute a Two-Component NHEJ Repair Machine.” *Science* 306 (5696): 683–85. https://doi.org/10.1126/science.1099824.

Deorowicz, Sebastian, Agnieszka Debudaj-Grabysz, and Adam Gudyś. 2016. “FAMSA: Fast and Accurate Multiple Sequence Alignment of Huge Protein Families.” *Scientific Reports* 6 (1): 1–13. https://doi.org/10.1038/srep33964.

Doherty, Aidan J., Stephen P. Jackson, and Geoffrey R. Weller. 2001. “Identification of Bacterial Homologues of the Ku DNA Repair Proteins.” *FEBS Letters* 500 (3): 186–88. https://doi.org/10.1016/S0014-5793(01)02589-3.

Elliot, Michael G, and Arne Ø Mooers. 2014. “Inferring Ancestral States without Assuming Neutrality or Gradualism Using a Stable Model of Continuous Character Evolution.” *BMC Evolutionary Biology* 14 (November). https://doi.org/10.1186/s12862-014-0226-8.

Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. “HMMER Web Server: Interactive Sequence Similarity Searching.” *Nucleic Acids Research* 39 (suppl_2): W29–37. https://doi.org/10.1093/nar/gkr367.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. “Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence.” *The American Naturalist* 160 (6): 712–26. https://doi.org/10.1086/343873.

Garland, Theodore, Allan W. Dickerman, Christine M. Janis, and Jason A. Jones. 1993. “Phylogenetic Analysis of Covariance by Computer Simulation.” *Systematic Biology* 42 (3): 265–92. https://doi.org/10.1093/sysbio/42.3.265.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. “New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.” *Systematic Biology* 59 (3): 307–21. https://doi.org/10.1093/sysbio/syq010.

Gupta, Richa, Daniel Barkan, Gil Redelman-Sidi, Stewart Shuman, and Michael S. Glickman. 2011. “Mycobacteria Exploit Three Genetically Distinct DNA Double-Strand Break Repair Pathways.” *Molecular Microbiology* 79 (2): 316–30. https://doi.org/10.1111/j.1365-2958.2010.07463.x.

Gyorfy, Zsuzsanna, Gabor Draskovits, Viktor Vernyik, Frederick F. Blattner, Tamas Gaal, and Gyorgy Posfai. 2015. “Engineered Ribosomal RNA Operon Copy-Number Variants of E. Coli Reveal the Evolutionary Trade-Offs Shaping RRNA Operon Number.” *Nucleic Acids Research* 43 (3): 1783–94. https://doi.org/10.1093/nar/gkv040.

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. “UFBoot2: Improving the Ultrafast Bootstrap Approximation.” *Molecular Biology and Evolution* 35 (2): 518–22. https://doi.org/10.1093/molbev/msx281.

Ives, Anthony R., and Theodore Garland. 2010. “Phylogenetic Logistic Regression for Binary Dependent Variables.” *Systematic Biology* 59 (1): 9–26. https://doi.org/10.1093/sysbio/syp074.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates.” *Nature Methods* 14 (6): 587–89. https://doi.org/10.1038/nmeth.4285.

Klappenbach, Joel A., John M. Dunbar, and Thomas M. Schmidt. 2000. "RRNA Operon Copy Number Reflects Ecological Strategies of Bacteria." *Applied and Environmental Microbiology* 66 (4): 1328–33. https://doi.org/10.1128/AEM.66.4.1328-1333.2000.

Malyarchuk, Svitlana, Douglas Wright, Reneau Castore, Emily Klepper, Bernard Weiss, Aidan J. Doherty, and Lynn Harrison. 2007. "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* 6 (10): 1413–24. https://doi.org/10.1016/j.dnarep.2007.04.004.

Matthews, Lindsay A., and Lyle A. Simmons. 2014. "Bacterial Nonhomologous End Joining Requires Teamwork." *Journal of Bacteriology* 196 (19): 3363–65. https://doi.org/10.1128/JB.02042-14.

Moeller, Ralf, Erko Stackebrandt, Günther Reitz, Thomas Berger, Petra Rettberg, Aidan J. Doherty, Gerda Horneck, and Wayne L. Nicholson. 2007. "Role of DNA Repair by Nonhomologous-End Joining in Bacillus Subtilis Spore Resistance to Extreme Dryness, Mono- and Polychromatic UV, and Ionizing Radiation." *Journal of Bacteriology* 189 (8): 3306–11. https://doi.org/10.1128/JB.00018-07.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Pagel, Mark. 1994. "Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255 (1342): 37–45. https://doi.org/10.1098/rspb.1994.0006.

Pagel, Mark, and Andrew Meade. 2006. "Bayesian Analysis of Correlated Evolution of Discrete Characters by Reversible-Jump Markov Chain Monte Carlo." *The American Naturalist* 167 (6): 808–25. https://doi.org/10.1086/503444.

Paris, Ülvi, Katren Mikkel, Kairi Tavita, Signe Saumaa, Riho Teras, and Maia Kivisaar. 2015. "NHEJ Enzymes LigD and Ku Participate in Stationary-Phase Mutagenesis in Pseudomonas Putida." *DNA Repair* 31 (July): 11–18. https://doi.org/10.1016/j.dnarep.2015.04.005.

Popa, Ovidiu, Einat Hazkani-Covo, Giddy Landan, William Martin, and Tal Dagan. 2011. "Directed Networks Reveal Genomic Barriers and DNA Repair Bypasses to Lateral Gene Transfer among Prokaryotes." *Genome Research* 21 (4): 599–609. https://doi.org/10.1101/gr.115592.110.

Revell, Liam J. 2012. "Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things)." *Methods in Ecology and Evolution* 3 (2): 217–23. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Roller, Benjamin R. K., Steven F. Stoddard, and Thomas M. Schmidt. 2016. "Exploiting RRNA Operon Copy Number to Investigate Bacterial Reproductive Strategies." *Nature Microbiology* 1 (11): 1–7. https://doi.org/10.1038/nmicrobiol.2016.160.

Sinha, Krishna Murari, Mihaela-Carmen Unciuleac, Michael S. Glickman, and Stewart Shuman. 2009. "AdnAB: A New DSB-Resecting Motor–Nuclease from Mycobacteria." *Genes & Development* 23 (12): 1423–37. https://doi.org/10.1101/gad.1805709.

Srinivas, Upadhyayula Sai, Bryce W. Q. Tan, Balamurugan A. Vellayappan, and Anand D. Jeyasekharan. 2019. "ROS and the DNA Damage Response in Cancer." *Redox Biology*, Redox Regulation of Cell State and Fate, 25 (July): 101084. https://doi.org/10.1016/j.redox.2018.101084.

Stephanou, Nicolas C., Feng Gao, Paola Bongiorno, Sabine Ehrt, Dirk Schnappinger, Stewart Shuman, and Michael S. Glickman. 2007. "Mycobacterial Nonhomologous End Joining Mediates Mutagenic Repair of Chromosomal Double-Strand DNA Breaks." *Journal of Bacteriology* 189 (14): 5237–46. https://doi.org/10.1128/JB.00332-07.

Tenaillon, Olivier, Erick Denamur, and Ivan Matic. 2004. “Evolutionary Significance of Stress-Induced Mutagenesis in Bacteria.” *Trends in Microbiology* 12 (6): 264–70. https://doi.org/10.1016/j.tim.2004.04.002.

Tomasetti, Cristian, Lu Li, and Bert Vogelstein. 2017. “Stem Cell Divisions, Somatic Mutations, Cancer Etiology, and Cancer Prevention.” *Science* 355 (6331): 1330–34. https://doi.org/10.1126/science.aaf9011.

Tung Ho, Lam si, and Cécile Ané. 2014. “A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models.” *Systematic Biology* 63 (3): 397–408. https://doi.org/10.1093/sysbio/syu005.

“UniProt: A Worldwide Hub of Protein Knowledge.” 2019. *Nucleic Acids Research* 47 (D1): D506–15. https://doi.org/10.1093/nar/gky1049.

Vernikos, Georgios S., and Julian Parkhill. 2006. “Interpolated Variable Order Motifs for Identification of Horizontally Acquired DNA: Revisiting the Salmonella Pathogenicity Islands.” *Bioinformatics* 22 (18): 2196–2203. https://doi.org/10.1093/bioinformatics/btl369.

Vieira-Silva, Sara, and Eduardo P. C. Rocha. 2010. “The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics.” *PLOS Genetics* 6 (1): e1000808. https://doi.org/10.1371/journal.pgen.1000808.

Weissman, Jake L., William F. Fagan, and Philip L. F. Johnson. 2019. “Linking High GC Content to the Repair of Double Strand Breaks in Prokaryotic Genomes.” *PLOS Genetics* 15 (11): e1008493. https://doi.org/10.1371/journal.pgen.1008493.

Wheeler, Travis J., and Sean R. Eddy. 2013. “Nhmmer: DNA Homology Search with Profile HMMs.” *Bioinformatics* 29 (19): 2487–89. https://doi.org/10.1093/bioinformatics/btt403.

Zhang, Xiaojuan, Wei Chen, Yang Zhang, Libin Jiang, Zhi Chen, Ying Wen, and Jilun Li. 2012. “Deletion of Ku Homologs Increases Gene Targeting Frequency in Streptomyces Avermitilis.” *Journal of Industrial Microbiology & Biotechnology* 39 (6): 917–25. https://doi.org/10.1007/s10295-012-1097-x.

Zhu, Hui, and Stewart Shuman. 2010. “Gap Filling Activities of Pseudomonas DNA Ligase D (LigD) Polymerase and Functional Interactions of LigD with the DNA End-Binding Ku Protein.” *Journal of Biological Chemistry* 285 (7): 4815–25. https://doi.org/10.1074/jbc.M109.073874.

Acknowledgment

We thank Saurabh Mahajan for discussions. We thank Supriya Khedkar for comments on the manuscript.

Funding

This work is supported by a DBT/Wellcome Trust India Alliance Intermediate Fellowship IA/I/16/2/502711 to ASNS and a Human Frontier of Science Career Development Award to AB. MS is supported by a DBT-JRF fellowship DBT/JRF/BET-16/I/2016/AL/86-466 from the Department of Biotechnology, Government of India.