Unexpectedly high mutation rate of a deep-sea hyperthermophilic anaerobic archaeon

Jiaohao Gu[1,2^], Xiaojun Wang[3^], Xiaopan Ma[1,2], Ying Sun[3], Xiang Xiao[1,2*], Haiwei Luo[3,4*]

[1]State Key Laboratory of Microbial Metabolism, Shanghai Jiao Tong University, Shanghai 200240, China
[2]State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[3]Simon F. S. Li Marine Science Laboratory, School of Life Sciences and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
[4]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen 518000, China

[^]These authors contributed equally to this study.

[*]**Corresponding author:**
Haiwei Luo
School of Life Sciences, The Chinese University of Hong Kong
Hong Kong SAR
Phone: (+852) 39436121
E-mail: hluo2006@gmail.com

Xiang Xiao
School of Life Sciences and Biotechnology, Shanghai Jiao Tong University
Shanghai, China
Phone: (+86) 21 34207206
E-mail: zjxiao2018@sjtu.edu.cn

Running Title: Hyperthermophile Evolution in Deep Ocean

Key words: deep sea hydrothermal vent, hyperthermophile, *Thermococcus*, mutation rate, genome reduction

Summary

Deep sea hydrothermal vents resemble the early Earth [1] , and thus the dominant *Thermococcaceae* inhabitants, which occupy an evolutionarily basal position of the archaeal tree [2,3] and take an obligate anaerobic hyperthermophilic free-living lifestyle [4], are likely excellent models to study the evolution of early life. Here, we determined that unbiased mutation rate of a representative species, *Thermococcus eurythermalis* [5], exceeded that of all known free-living prokaryotes by 1-2 orders of magnitude, and thus rejected the long-standing hypothesis that low mutation rates were selectively favored in hyperthermophiles [6–8]. We further sequenced multiple and diverse isolates of this species and calculated that *T. eurythermalis* has a lower effective population size than other free-living prokaryotes by 1-2 orders of magnitude. These data are well explained by the "drift-barrier" model [9], indicating that the high mutation rate of this species is not selectively favored but instead driven by random genetic drift. The availability of these unusual data has far-reaching implications for prokaryote genome evolution. For example, a synthesis of additional 29 species with unbiased mutation rate data across bacteria and archaea enabled us to conclude that genome reduction across prokaryotes is universally driven by increased mutation rate and random genetic drift. Taken together, exceptionally high mutation rate and low effective population size likely feature the early life in hot and anoxic marine habitats, which are indispensable in synthesizing the universal rule of genome evolution across prokaryotes and the Earth history.

Main

One theory for the origin of life is that the last universal common ancestor (LUCA) was an anaerobic hyperthermophilic organism inhabiting the deep sea hydrothermal vents, as these environments display a few characteristics paralleling the early Earth [1]. While hydrothermal vents vary with chemical parameters, they all share a high temperature zone near the black chimney with anaerobic fluid from it. A few studies showed that hyperthermophiles can change metabolic strategy in response to heat stress [10], but little is known whether they have a high intrinsic (i.e., not selected by environmental pressure) rate to change their genetic background information and whether this intrinsic potential itself is a result of selection shaped by these unique habitats.

A previous population genomic analysis showed that protein sequences are under greater functional constraints in thermophiles than in mesophiles, suggesting that mutations are functionally more deleterious in thermophiles than in mesophiles [11]. This explanation is also supported by experimental assays showing nearly neutral mutations in temperate conditions become strongly deleterious at high temperature [6]. Furthermore, fluctuation tests on a hyperthermophilic archeaon *Sulfolobus acidocaldarius* [7] and a hyperthermophilic bacterium *Thermus thermophilus* [8] consistently showed that hyperthermophiles have much lower mutation rate compared to mesophiles. This appears to support the hypothesis that selection favors high replication fidelity at high temperature [6].

Nevertheless, mutation rates measured using fluctuation experiments based on reporter loci are known to be biased, since the mutation rate of the organism is extrapolated from a few specific nonsynonymous mutations enabling survival in an appropriate selective medium, which renders the results susceptible to uncertainties associated with the representativeness of these loci

and to inaccuracies of the assumptions made in extrapolation methods [12–14]. These limitations are avoided by the mutation accumulation (MA) experiment followed by whole-genome sequencing (WGS) of derived lines. In the MA part, multiple independent MA lines initiated from a single progenitor cell each regularly pass through a single-cell bottleneck, usually by transferring on solid medium. As the effective population size (N_e) becomes one, selection is unable to eliminate all but the lethal mutations, rendering the MA/WGS an approximately unbiased method to measure the spontaneous mutation rate [15].

Members of the free-living anaerobic hyperthermophilic archaeal family *Thermococcaceae* are among the dominant microbial lineages in the black-smoker chimney at Guaymas Basin [16] and other deep sea hydrothermal vents [17,18]. This family only contains three genera: *Thermococcus*, *Pyrococcus* and *Palaeococcus*. In this study, the MA/WGS procedure was applied to determine the unbiased spontaneous mutation rate of a representative member *Thermococcus eurythermalis* A501, a conditional pizeophilic archaeon which can grow equally well from 0.1 MPa to 30 MPa at 85℃ [5,19]. The MA lines were propagated at this optimal temperature on plates with gelrite which tolerates high temperature, and the experiment was performed under normal air pressure and in strictly anaerobic condition (Fig. 1A-D). To the best of our knowledge, this is the first report of unbiased mutation rate of a hyperthermophile and an obligate anaerobe.

Our MA experiment allowed accumulation of mutations over 314 cell divisions (after correcting the death rate [20]) in 100 independent lines initiated from a single founder colony and passed through a single cell bottleneck every day. By sequencing genomes of 96 survived lines at the end of the MA experiment, we identified 544 base-substitution mutations over these lines (Table S1), which translates to an average mutation rate (μ) of 85.01×10^{-10} per cell division per

nucleotide site (see Methods). The ratio of nonsynonymous to synonymous mutations did not differ from the ratio of nonsynonymous to synonymous sites in the A501 genome (χ^2 test; $p>0.05$). Likewise, there was no difference of the accumulated mutations between intergenic and protein-coding sites (χ^2 test; $p>0.05$). These are evidence for minimal selective elimination of deleterious mutations during the MA process. In general, the mutations were randomly distributed along the chromosome and the plasmid, though 86 mutations fell into 14 genes which showed significant enrichment of mutations (bootstrap test; $p<0.05$ for each gene) and 52 out of the 86 mutations were found in five genes (TEU_RS04685 and TEU_RS08625-08640 gene cluster) (Fig. 1E, Table S2). These regions may represent either mutational hotspots or that mutations confer selective advantages [21]. The TEU_RS04685 encodes glutaconyl-CoA decarboxylase subunit β which acts as a Na^+ pump, and the TEU_RS08625-08640 encodes a ribose ABC transporter. It remains unknown the molecular mechanism underlying repeated mutations at these loci. Removing these mutations led to a spontaneous mutation rate of 71.57×10^{-10} per cell division per site for *T. eurythermalis* A501. After removing the mutations in these 14 genes, both the accumulated mutations at nonsynonymous sites relative to those at synonymous sites (χ^2 test; $p=0.014$) and the accumulated mutations at intergenic regions relative to protein-coding regions (χ^2 test; $p=0.013$) showed marginally significant differences.

To date, over 20 phylogenetically diverse free-living bacterial species and two archaeal species isolated from various environments have been assayed with MA/WGS, and their mutation rates vary from 0.79×10^{-10} to 97.80×10^{-10} per cell division per site [22]. The only prokaryote that displays a mutation rate (97.80×10^{-10} per cell division per site) comparable to A501 is *Mesoplasma florum* L1 [9], a host-dependent wall-less bacterium with highly reduced genome (~700 genes). Our PCR validation of randomly chosen 20 base-substitution mutations

from two MA lines displaying highest mutation rates and of all nine insertion-deletion (INDEL) mutations involving >10 bp changes across all lines (Table S1) indicates that the calculated high mutation rate does not result from false bioinformatics predictions.

The extremely high mutation rate of *T. eurythermalis* is unexpected. One explanation (also the selection-driven hypothesis) is that high mutation rate may allow the organisms to gain beneficial mutations more rapidly and thus is selectively favored in deep sea hydrothermal vents where physicochemical parameters are highly fluctuating. Alternatively (also the neutral hypothesis), high mutation rate is the result of random genetic drift according to the drift-barrier model [9]. In this model, increased mutation rates are associated with increased load of deleterious mutations, so natural selection favors lower mutation rates. On the other hand, increased improvements of replication fidelity come at an increased cost of investments in DNA repair activities. Therefore, natural selection pushes the replication fidelity to a level that is set by genetic drift, and further improvements are expected to reduce the fitness advantages [9,15]. These two explanations for high mutation rate of *T. eurythermalis* are mutually exclusive, and resolving them requires the calculation of the power of genetic drift, which is inversely proportional to N_e of *T. eurythermalis*.

A common way to calculate N_e for a prokaryotic population is derived from the equation $\pi_S = 2 \times N_e \times \mu$, where π_S represents the nucleotide diversity at silent (synonymous) sites among randomly sampled members of a panmictic population [23]. We therefore sequenced genomes of another eight *T. eurythermalis* isolates available in our culture collections. Like *T. eurythermalis* A501, these additional isolates were collected from the same cruise but varying at the water depth from 1,987 m to 2,009 m at Guaymas Basin. They differ by only up to 0.135% in the 16S rRNA gene sequence and share a minimum whole-genome average nucleotide identity (ANI) of

95.39% (Table S3), and thus fall within an operationally defined prokaryotic species typically delineated at 95% ANI [24]. Population structure analysis with PopCOGenT [25] showed that these isolates formed a panmictic population and that two of them were repetitive as a result of clonal descent (see Methods). Using the median value of π_S =0.083 across 1,628 single-copy orthologous genes shared by the seven non-repetitive genomes, we calculated the N_e of *T. eurythermalis* to be 5.83×10^6.

Next, we collected the unbiased mutation rate of other prokaryotic species determined with the MA/WGS strategy from the literature [15,26–28]. While the N_e data were also provided from those studies, the isolates used to calculate the N_e were identified based on their membership of either an operationally defined species (e.g., ANI at 95% cutoff) or a phenotypically characterized species (e.g., many pathogens), which often create a bias in calculating N_e [23]. We therefore again employed PopCOGenT to delineate panmictic populations from those datasets and re-calculated N_e accordingly. There was a significant negative linear relationship between μ and N_e on a logarithmic scale (dashed gray line in Fig. 2A [r^2 = 0.83, slope = -0.85, s.e.m. = 0.09, p<0.001]) according to a generalized linear model (GLM) regression. This relationship cannot be explained by shared ancestry, as confirmed by phylogenetic generalized least square (PGLS) regression analysis (solid blue line in Fig. 2A [r^2 = 0.81, slope = -0.81, s.e.m. = 0.09, p<0.001]). The nice fit of *T. eurythermalis* to the regression line validated the drift-barrier hypothesis. This is evidence that the high mutation rate of *T. eurythermalis* is driven by genetic drift rather than by natural selection.

As stated in the drift-barrier theory, high mutation rate is associated with a high load of deleterious mutations. In the absence of back mutations, recombination becomes an essential mechanism in eliminating deleterious mutations [29]. In support of this argument, the

ClonalFrameML analysis [30] shows that members of the *T. eurythermalis* population recombine frequently, with a high ratio of the frequency of recombination to mutation (ρ/θ=0.59) and a high ratio of the effect of recombination to mutation (r/m=5.76). In fact, efficient DNA incorporation to *Thermococcaceae* genomes from external sources has been well documented experimentally [31,32]. A second potentially important mechanism facilitating *T. eurythermalis* adaptation at high temperature is strong purifying selection at the protein sequence level, as protein sequences in thermophiles are generally subject to stronger functional constraints compared to those in mesophiles [11,33].

Our result of the exceptionally high mutation rate of a free-living archaeon is a significant addition to the available collection of the MA/WGS data (Table S4), in which prokaryotic organisms with very high mutation rate have only been known for a host-dependent bacterium (*Mesoplasma florum* L1) with unusual biology (e.g., cell wall lacking). The availability of these two deeply branching (one archaeal versus the other bacterial) organisms adopting opposite lifestyles (one free-living versus the other host-restricted; one hyperthermophilic versus the other mesophilic; one obligate anaerobic versus the other facultative anaerobe), along with other phylogenetically and ecologically diverse prokaryotic organisms displaying low and intermediate mutation rates, provides an unprecedented opportunity to illustrate key mechanisms driving genome size evolution across prokaryotes. First, a negative linear relationship (dashed gray line in Fig. 2B [r^2 = 0.42, slope = -1.43, s.e.m. = 0.31, p<0.001]) between genome size and base-substitution mutation rate is evidence that increased mutation rate drives genome reduction across bacteria and archaea. Second, a positive lineage relationship (dashed gray line in Fig. 2C [r^2 = 0.47, slope = 0.24, s.e.m. = 0.06, p<0.001]) between genome size and N_e supports that random genetic drift drives genome reduction across prokaryotes. These correlations remain

193 robust when the data were analyzed as phylogenetically independent contrasts (blue solid lines in
194 Fig. 2B [r^2 = 0.39, slope = -1.43, s.e.m. = 0.32, p<0.001] and in Fig. 2C [r^2 = 0.45, slope = 0.25,
195 s.e.m. = 0.06, p<0.001]). These two mechanisms for genome reduction each were proposed for
196 both free-living [34,35] and host-dependent [36,37] bacteria. In addition, previous studies which
197 intended to illustrate the universal mechanisms for genome reduction across bacteria (archaea
198 datasets not available by the time of those studies) were reliant on incomplete datasets that
199 lacked data from genome-reduced free-living bacteria [15,38]. Nevertheless, the analysis presented
200 here is the first time that the unbiased spontaneous mutation rate and N_e from a genome-reduced
201 free-living prokaryotic population is included, enabling the generalization of the mechanisms
202 across bacteria and archaea.

203 Whereas our analysis rejected natural selection as a universal mechanism driving genome
204 reduction across prokaryotes (Fig. 2B&C), it does not mean that selection has no role in genome
205 reduction of a particular taxon. In the case of thermophiles, proponents for selection acting to
206 reduce genomes explained that genome size, due to its positive correlation with cell volume, may
207 be an indirect target of selection which strongly favors smaller cell volume [33]. The underlying
208 principle is that high temperature requires cells to increase the lipid content and change the lipid
209 composition of the cell membranes, which consumes a large part of the cellular energy, and thus
210 lower cell volume is selectively favored at high temperature [33]. Our calculations of a relatively
211 small N_e in *T. eurythermalis* does not necessarily contradict with this selective argument, given
212 that the fitness gained by decreasing cell volume and thus reducing genome size is large enough
213 to overcome the power of random genetic drift. On the other hand, our data strongly indicate that
214 neutral forces dictate the genome evolution of *T. eurythermalis*, and are not negligible with
215 regard to its genome reduction process. The significantly more deletion over insertion events (t

test; 95 versus 37 events with $p<0.001$ and 48 versus 20 events with $p<0.05$ before and after removing the 14 genes enriched in mutations, respectively) and the significantly more nucleotides involved in deletions over insertions (t test; 433 versus 138 bp with $p<0.05$ and 386 versus 121 bp with $p<0.001$ before and after removing the 14 genes enriched in mutations, respectively) suggest that the deletion bias, combined with increased chance fixation of deletion mutants due to low N_e, is a potentially important neutral mechanism giving rise to the small genomes of *T. eurythermalis* (2.12 Mbp).

The globally distributed deep sea hydrothermal vents are microbe-driven ecosystems, with no known macroorganisms surviving at the vent fluids. Sample collections, microbial isolations, and laboratory propagations of mutation lines at high temperature are challenging. In the present study, we determined that *T. eurythermalis*, and perhaps *Thermococcaceae* in general, has a highly increased mutation rate and a highly decreased N_e compared to all other known free-living prokaryotic lineages. While it remains to be tested whether this is a common feature among the vents' populations, the present study nevertheless opens a new avenue for investigating the hyperthemophile ecology and evolution in the deep sea. Furthermore, the availability of the *T. eurythermalis* unbiased mutation rate data allows us to draw another major conclusion that the genome reduction processes across bacteria and archaea are largely dictated by increased mutation rate and decreased selection efficiency.

Methods

Sampling, cultivation, and genome sequencing of Thermococcus eurythermalis isolates

Nine *Thermococcus eurythermalis* strains (Table S3) were isolated from samples of Guaymas Basin hydrothermal vents in the cruise number AT 15–55, during 7-17 November 2009

239 [39]. Briefly, samples were stored in the Hungate anaerobic tubes and kept at 4℃. Then the
240 samples were enriched at 85℃ or 95℃ using *Thermococcales* Rich Medium (TRM) medium.
241 Next, enrichment cultures were inoculated on the solid medium prepared with hungate roll-tube
242 technique and incubated at 85℃ or 95℃ under atmosphere pressure. Single colonies were
243 transferred into new TRM medium and purified using roll-tube technique for 3 times and stocks
244 were kept at -80℃. More details of sampling and isolation can be found in a previous paper [39].
245 Among these isolates, the complete genome of the type strain A501 (GCA_000769655.1) was
246 downloaded from the NCBI GenBank database [40], and the rest eight strains were sequenced in
247 the present study. To get enrichment of these eight strains, stocks kept in -80℃ were inoculated
248 into 50 mL anaerobic TRM medium in the serum bottle and cultured in the incubator in 85℃.
249 The liquid medium was supplemented with sulfur and $Na_2S \cdot 9H_2O$. After enrichment, the cells
250 were collected using centrifuge (12,000 rpm, 10min). Genomic DNA of each isolate was
251 extracted using the Magen Hipure Soil DNA Kit and was sequenced using the Illumina Hiseq
252 platform with 2×150 bp paired-end. Raw reads were first processed by Trimmomatic 0.32 [40] to
253 remove adaptors and trim bases of low quality. The draft genome of each isolate was assembled
254 with quality reads using SPAdes v3.10.1 [41] with default parameters.
255
256 *Mutation accumulation experiment*
257 For culture propagation under high temperature, anaerobic high-temperature-tolerant
258 plates were made every day before the transfer. Plates were made using anaerobic *Thermococcus*
259 Rich Medium [38] (TRM) with gelrite (15g $liter^{-1}$). After sterilization, 1.5 mL of a polysulfide
260 solution [42] was added per liter of medium using syringe to make sure a strictly anaerobic
261 condition. The medium was transferred into an anaerobic chamber (COY, Vinyl Anaerobic

Chamber) immediately, preventing it from cooling. This is because gelrite used for making plates becomes solidified soon after it become cooler. Plates were made in the chamber.

The mutation accumulation (MA) experiment started from a single founder colony of *Thermococcus eurythermalis* A501. It was transferred to new plates to form 100 independent lines. Plates were put into an anaerobic jar (GeneScience), which were together moved to an incubator. After incubation at 85℃ under normal air pressure (optimal growth pressure from 0.1-30 MPa) for one day, the jar was transferred back into the anaerobic chamber. Plates were then taken out. This was the initiation of the MA process. Caution was taken to ensure a strictly anaerobic condition maintained throughout the experiment. A single/tiny (< 1 mm) colony of each line was carefully picked and transferred onto a new plate. Then the new plates were put back into the anaerobic jar for incubation. The single cell bottleneck of the MA process occurred during every transfer.

The MA propagation was completed following 20 transfers, and four MA lines were lost during the MA process. A single colony on each plate was transferred into 5 mL anaerobic TRM medium in the anaerobic chamber. The liquid medium was supplemented with sulfur and $Na_2S{\cdot}9H_2O$. After incubation at 85℃ for one day, stocks of each line were kept at -80℃. Genomic DNA of each survived MA line was extracted using the Magen Hipure Soil DNA Kit, and was sequenced using the same platform mentioned above. A sequencing coverage depth of ~433× with an average library fragment size of ~470 bp was obtained for each line.

Generation time estimation with correction for cell death rate

To estimate the generation time, a whole single colony was cut from 10 randomly selected MA lines. The selected 10 colonies each were moved into 5 mL anaerobic TRM

285 medium supplemented with $Na_2S \cdot 9H_2O$. After dilution and re-plating, live cell density (*d*) was
286 measured with viable cell counts. The live and dead cell staining was done to correct the total
287 cell density for each colony. Briefly, to obtain the sufficient cell density for staining, ten single
288 colonies were cut from every MA line selected above. Live and dead bacterial staining kit
289 (Yeasen Biotech Co.) was used in this study. The kit was tested to be effective in archaea. The
290 cells were put into 350 μL anaerobic TRM medium supplemented with $Na_2S \cdot 9H_2O$. After
291 centrifuge with 10,000 g for 10 min, cells were resuspended in 50 μL medium. Cell staining was
292 done following the protocol of the kit. Fluorescence microscope (Nikon) was used to
293 differentiate between live and dead cells. The ratio of live cells to total cells (r) was 0.942 (±
294 0.095) (Table S5). The number of cell divisions per transfer (*D*) was corrected by:

295 $$D = \log_2(\frac{d}{r})$$

296 where *d* is the live cell density and *r* is the ratio of live cells in total cells. The total number of
297 generations that each MA line went through was the multiplication of average number of cell
298 divisions per transfer and the total number of transfers. Since each MA line underwent 20
299 transfers with an average of 15.72 ± 1.76 cell divisions per transfer, there were a total of 314.4 ±
300 35.2 generations for each MA line.

301

302 *Mutation calling and mutation rate determination*

303 Raw reads were first processed by Trimmomatic 0.32 [43] to remove adaptors and trim low-
304 quality bases. Then the paired-end reads of 96 MA lines were individually mapped to the *T.*
305 *eurythermalis* A501 reference genome using two different mappers: BWA-mem [44] and
306 NOVOALIGN v2.08.02 (www.novocraft.com). The resulting pileup files were converted to
307 SAM format with SAMTOOLS [45].

The above mapping results were processed by Picard MarkDuplicates (http://broadinstitute.github.io/picard/) to remove duplicate reads which may arise during sample preparation like PCR duplication artifacts or derive from a single amplification cluster. Base quality score recalibration was performed to adjust quality score affected by systematic technical errors using BaseRecalibrator in GATK-4.0 [46]. Then base substitutions and small indels were called using HaplotypeCaller implemented in GATK-4.0 [46]. Variants were further filtered with standard parameters described by GATK Best Practices recommendations, except that the Phred-scaled quality score QUAL > 100 and RMS mapping quality MQ > 59 were set, which followed previous studies [46–49]. PCR primers were designed with Primer Premier 5.0 [50] to confirm the presence of mutations identified by the above bioinformatics method. Twenty base substitutions and nine indels were sampled from 11 lines and validated. These lines were chosen because two of these lines showed the highest base-substitution mutation rate and the remaining nine lines showed the longest indel mutations (Table S1). The average number of analyzable sites and the average coverage per site in the *T. eurythermalis* A501 MA lines were 2,123,047 (± 674) and 431 (± 57), respectively.

The base-substitution mutation rate per nucleotide site per cell division (μ) for each line was calculated according to the following equation:

$$\mu = \frac{m}{nG}$$

Where *m* is the number of observed base substitutions, *n* is the number of nucleotide sites analyzed, and *G* is the mean number of cell divisions estimated during the mutation accumulation process. Following a previous study [26], the total standard error of base-substitution mutation rate across all MA lines was calculated by:

$$SE_{pooled} = \frac{s}{\sqrt{N}}$$

where *s* is the standard deviation of the mutation rate across all lines, and *N* is the number of lines analyzed.

The effective population size estimation for Thermococcus eurythermalis

The effective population size (N_e) of a prokaryotic species was calculated following the equation $\pi_S = 2 \times N_e \times \mu$, where π_S is the nucleotide diversity at silent (synonymous) sites among randomly sampled members of a species and μ is the unbiased spontaneous mutation rate. Microbial species commonly harbor genetically structured populations, which has a major influence on π_S and thus N_e estimation. It is therefore important to identify strains allowed for free recombination when calculating N_e for a prokaryotic species [23]. The recently available program PopCOGenT [25] identifies members from a prokaryotic species constituting a panmictic population. The basic idea of PopCOGenT is that the recent homologous recombination erased the single nucleotide polymorphisms (SNPs) and led to identical regions between genomes, and therefore strains subjected with frequent recent gene transfers are expected to show an enrichment of identical genomic regions compared to accumulation of SNPs between genomes lacking recent transfer [25]. In practice, strains were connected via recent gene flow into a network, and a putative population was identified as a cluster, with within-cluster DNA transfer frequency much higher than that of between clusters. Only one strain within each clonal complex was kept, which is also important for π_S estimation because an overuse of strains from a clonal complex is expected to underestimate π_S. Then the cluster containing the largest number of strains was chosen as the panmictic population for a given species. In the case of *T. eurythermalis*, all nine strains together form a panmictic population, but two strains were not used in the calculation because they were repetitive members of clonal complexes.

Next, the single-copy orthologous genes shared by all the seven *T. eurythermalis* genomes were identified by OrthoFinder 2.2.1 [51]. Amino acid sequences of each gene family were aligned with MAFFT v7.464 [52] and then imposed on nucleotide sequences. The number of synonymous substitution per synonymous site (d_S) for each possible gene pair in each gene family was computed with the YN00 program in PAML 4.9e [53]. The π_S of each gene family was obtained by averaging all pairwise d_s values, and then the median π_S across all single-copy gene families together with μ were used to calculate the N_e. We used the median π_S instead of the mean value, because loci showing unusually large d_S as a result of allelic replacement via homologous recombination with divergent lineages are common in bacterial species [54], which are expected to bias the mean value but have a limited effect on the median value across gene loci.

Data synthesis

To enable a comparative analysis of *T. eurythermalis* relative to other prokaryotic species, the available μ values of other 29 prokaryotic species determined with the MA/WGS technique were collected from the literature (Table S4). Among these, 20 species each had multiple isolates' genomes available from the NCBI Refseq database [55], and thus were used for N_e calculation. The calculation of N_e for these species followed the abovementioned procedure detailed for *T. eurythermalis*, which started with the identification of members constituting a panmictic population by PopCOGenT, followed by the calculation of π_S. A few species have thousands of isolates' genomes available in Refseq (Table S4), which are not amenable for the PopCOGenT analysis. For these species, we started from the populations previously identified by ConSpeciFix [51] and used these genomes as the input of PopCOGenT. The ConSpeciFix delineates populations based on homoplasious SNPs, which retains historical recombination

signal and blurs the boundary of the ecological populations enriched with recent gene transfers [25]. In the case of the species *Ruegeria pomeroyi* DSS-3, a model heterotrophic marine bacterium with its mutation rate available [26], since closely related isolates has not been available, we turned to its closely related species *Epibacterium mobile* (previously known as *Ruegeria mobile*) with multiple isolates' genomes available.

Next, the pairwise linear relationship between μ, N_e, and genome size across the prokaryotic species was initially assessed with the generalized linear model (GLM) implemented in *stats* package in R v4.0.2 [56]. The Bonferonni adjusted outlier test was performed with *outlierTest* function in *car* package [57]. A data point with Bonferroni *p*-value smaller than 0.05 would be identified as the outlier. For μ versus genome size, all 30 species were used. In the case of N_e versus μ and N_e versus genome size, only the 21 species each containing multiple strains' genomes were used. To test whether there was a phylogenetic signal of these traits, the Pagel's λ [52] was estimated using the *pgls* function of the *caper* package [58] which took the phylogeny of 30 species or the phylogeny of 21 species as an input. The species phylogeny was approximated by the 16S rRNA gene tree constructed using IQ-TREE 2.0 [59] with ModelFinder [60] which assigns the best substitution model and with 1,000 ultrafast bootstrap replicates. The value of λ ranges from 0 to 1, with 0 indicating no phylogenetic signal and 1 indicating a strong phylogenetic signal due to Brownian motion. The *p* values for the lower and upper bounds represent whether the λ is significantly different from 0 and 1, respectively. The results of this test indicate that there was an intermediate phylogenetic signal for the relationship of N_e versus μ ($\lambda = 0.81$, lower bound $p = 0.29$, upper bound $p = 0.06$), but not for that of N_e versus genome size and μ versus genome size (in both cases, $\lambda = 0$, lower bound $p = 1$, upper bound $p < 0.001$). To control for the phylogenetic effect on the correlations of the traits, the pairwise linear relationship between μ,

400 N_e, and genome size was further assessed with the phylogenetic generalized least square (PGLS)
401 regression implemented in the *caper* package [61] in R v4.0.2 [56]. The PGLS and GLM regression
402 lines were largely overlapped for N_e versus genome size and μ versus genome size (Fig. 2BC).
403 This is because no phylogenetic signal was detected in these relationships. A data point was
404 identified as an outlier in the PGLS result if the associated absolute value of studentized residual
405 is greater than three [62,63].
406
407
408 **Acknowledgement**
409 This research is supported by the National Key R&D Program of China (2018YFC0309800),
410 National Nature of Science China (NSFC 41530967), China Ocean Mineral Resources R & D
411 Association DY125-22-04. HL is also supported by the Hong Kong Research Grants Council
412 Area of Excellence Scheme (AoE/M-403/16).
413
414 **Competing interests**
415 The authors declare no competing commercial interests in relation to the submitted work.
416
417

References

1. Martin, W., Baross, J., Kelley, D. & Russell, M. J. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **6**, 805–814 (2008).
2. Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA* **114**, E4602–E4611 (2017).
3. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
4. Schut, G. J. *et al.* The Order Thermococcales and the Family Thermococcaceae. in *The Prokaryotes: Other Major Lineages of Bacteria and The Archaea* (eds. Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) 363–383 (Springer, 2014). doi:10.1007/978-3-642-38954-2_324.
5. Zhao, W., Zeng, X. & Xiao, X. Thermococcus eurythermalis sp. nov., a conditional piezophilic, hyperthermophilic archaeon with a wide temperature range for growth, isolated from an oil-immersed chimney in the Guaymas Basin. *Int. J. Syst. Evol. Microbiol.* **65**, 30–35 (2015).
6. Drake, J. W. Avoiding Dangerous Missense: Thermophiles Display Especially Low Mutation Rates. *PLoS Genet.* **5**, e1000520 (2009).
7. Grogan, D. W., Carver, G. T. & Drake, J. W. Genetic fidelity under harsh conditions: Analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius. *Proc. Natl. Acad. Sci. USA* **98**, 7928–7933 (2001).
8. Mackwan, R. R., Carver, G. T., Kissling, G. E., Drake, J. W. & Grogan, D. W. The rate and character of spontaneous mutation in Thermus thermophilus. *Genetics* **180**, 17–25 (2008).
9. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. USA* **109**, 18488–18492 (2012).
10. Borges, N., Matsumi, R., Imanaka, T., Atomi, H. & Santos, H. Thermococcus kodakarensis Mutants Deficient in Di-myo-Inositol Phosphate Use Aspartate To Cope with Heat Stress. *J. Bacteriol. Res.* **192**, 191–197 (2010).
11. Friedman, R., Drake, J. W. & Hughes, A. L. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. *Genetics* **167**, 1507–1512 (2004).
12. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. USA* **109**, E2774–E2783 (2012).
13. Long, H. *et al.* Antibiotic treatment enhances the genome-wide mutation rate of target cells. *Proc. Natl. Acad. Sci. USA* E2498–E2505 (2016) doi:10.1073/pnas.1601208113.
14. Williams, A. B. Spontaneous mutation rates come into focus in Escherichia coli. *DNA Repair (Amst.)* **24**, 73–79 (2014).
15. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**, 704–714 (2016).
16. He, Y., Xiao, X. & Wang, F. Metagenome reveals potential microbial degradation of hydrocarbon coupled with sulfate reduction in an oil-immersed chimney from Guaymas Basin. *Front. Microbiol.* **4**, 148 (2013).
17. Roussel, E. G. *et al.* Extending the sub-sea-floor biosphere. *Science* **320**, 1046 (2008).
18. Takai, K. & Nakamura, K. Archaeal diversity and community development in deep-sea hydrothermal vents. *Curr. Opin. Microbiol.* **14**, 282–291 (2011).
19. Zhao, W. *et al.* Cross-stress adaptation in a piezophilic and hyperthermophilic archaeon from deep sea hydrothermal vent. *Front. Microbiol.* **11**, (2020).

20. Frenoy, A. & Bonhoeffer, S. Death and population dynamics affect mutation rate estimates and evolvability under stress in bacteria. *PLoS Biol.* **16**, e2005056 (2018).
21. Farlow, A. *et al.* The Spontaneous Mutation Rate in the Fission Yeast Schizosaccharomyces pombe. *Genetics* **201**, 737–744 (2015).
22. Long, H. *et al.* Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* **2**, 237–240 (2018).
23. Rocha, E. P. C. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. *Mol. Biol. Evol.* **35**, 1338–1347 (2018).
24. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High-throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *Nat. Commun.* **9**, 5114 (2018).
25. Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J. & Polz, M. F. A reverse ecology approach based on a biological definition of microbial populations. *Cell* **178**, 820–834 (2019).
26. Sun, Y. *et al.* Spontaneous mutations of a model heterotrophic marine bacterium. *ISME J* **11**, 1713–1718 (2017).
27. Senra, M. V. X. *et al.* An Unbiased Genome-Wide View of the Mutation Rate and Spectrum of the Endosymbiotic Bacterium Teredinibacter turnerae. *Genome Biol. Evol.* **10**, 723–730 (2018).
28. Dillon, M. M., Sung, W., Sebra, R., Lynch, M. & Cooper, V. S. Genome-Wide Biases in the Rate and Molecular Spectrum of Spontaneous Mutations in Vibrio cholerae and Vibrio fischeri. *Mol. Biol. Evol.* **34**, 93–109 (2017).
29. Hughes, A. L. Near neutrality: leading edge of the neutral theory of molecular evolution. *Ann. N. Y. Acad. Sci.* **1133**, 162–179 (2008).
30. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
31. Song, Q. *et al.* Induction of a Toxin-Antitoxin Gene Cassette under High Hydrostatic Pressure Enables Markerless Gene Disruption in the Hyperthermophilic Archaeon Pyrococcus yayanosii. *Appl. Environ. Microbiol.* **85**, (2019).
32. Sato, T., Fukui, T., Atomi, H. & Imanaka, T. Improved and versatile transformation system allowing multiple genetic manipulations of the hyperthermophilic archaeon Thermococcus kodakaraensis. *Appl. Environ. Microbiol.* **71**, 3889–3899 (2005).
33. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol. Evol.* **5**, 966–977 (2013).
34. Marais, G. A. B., Calteau, A. & Tenaillon, O. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica* **134**, 205–210 (2008).
35. Luo, H., Huang, Y., Stepanauskas, R. & Tang, J. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. *Nat. Microbiol.* **2**, 17091 (2017).
36. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13–26 (2011).
37. Itoh, T., Martin, W. & Nei, M. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12944–12948 (2002).
38. Kuo, C.-H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).

509 39. Liu, L., Wang, F., Xu, J. & Xiao, X. Molecular diversity of Thermococcales isolated
510 from Guaymas Basin hydrothermal vents. *Acta Oceanol. Sin.* **32**, 75–81 (2013).
511 40. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank.
512 *Nucleic Acids Res.* **35**, D21–D25 (2007).
513 41. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to
514 single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
515 42. Matsumi, R., Manabe, K., Fukui, T., Atomi, H. & Imanaka, T. Disruption of a sugar
516 transporter gene cluster in a hyperthermophilic archaeon using a host-marker system based on
517 antibiotic resistance. *J. Bacteriol.* **189**, 2683–2691 (2007).
518 43. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
519 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
520 44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
521 transform. *Bioinformatics* **25**, 1754–1760 (2009).
522 45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
523 2078–2079 (2009).
524 46. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
525 analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
526 47. Long, H., Behringer, M. G., Williams, E., Te, R. & Lynch, M. Similar Mutation Rates
527 but Highly Diverse Mutation Spectra in Ascomycete and Basidiomycete Yeasts. *Genome Biol.*
528 *Evol.* **8**, 3815–3821 (2016).
529 48. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-
530 generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
531 49. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the
532 Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1-33
533 (2013).
534 50. Singh, V. K., Mangalam, A. K., Dwivedi, S. & Naik, S. Primer premier: program for
535 design of degenerate primers from a protein sequence. *BioTechniques* **24**, 318–319 (1998).
536 51. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
537 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157
538 (2015).
539 52. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
540 Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
541 53. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**,
542 1586–1591 (2007).
543 54. Sun, Y. & Luo, H. Homologous recombination in core genomes facilitates marine
544 bacterial adaptation. *Appl. Environ. Microbiol.* **84**, e02545-17 (2018).
545 55. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
546 taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
547 56. R Development Core Team. *R: A language and environment for statistical computing.* (R
548 Foundation for Statistical Computing, Vienna).
549 57. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (Sage, 2019).
550 58. Orme, D. *et al. The Caper Package: Comparative Analysis of Phylogenetics and*
551 *Evolution in R R Package Version 1.0.1*. (2018).
552 59. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and
553 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*
554 **32**, 268–274 (2015).

555 60. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
556 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–
557 589 (2017).
558 61. Orme, D. The caper package: comparative analysis of phylogenetics and evolution in R.
559 36 (2013).
560 62. Powell, L. E., Isler, K. & Barton, R. A. Re-evaluating the link between brain size and
561 behavioural ecology in primates. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **284**, 20171765
562 (2017).
563 63. Harrison, T. L., Simonsen, A. K., Stinchcombe, J. R. & Frederickson, M. E. More
564 partners, more ranges: generalist legumes spread more easily around the globe. *Biol. Lett.* **14**,
565 20180616 (2018).
566
567

Figure Legends

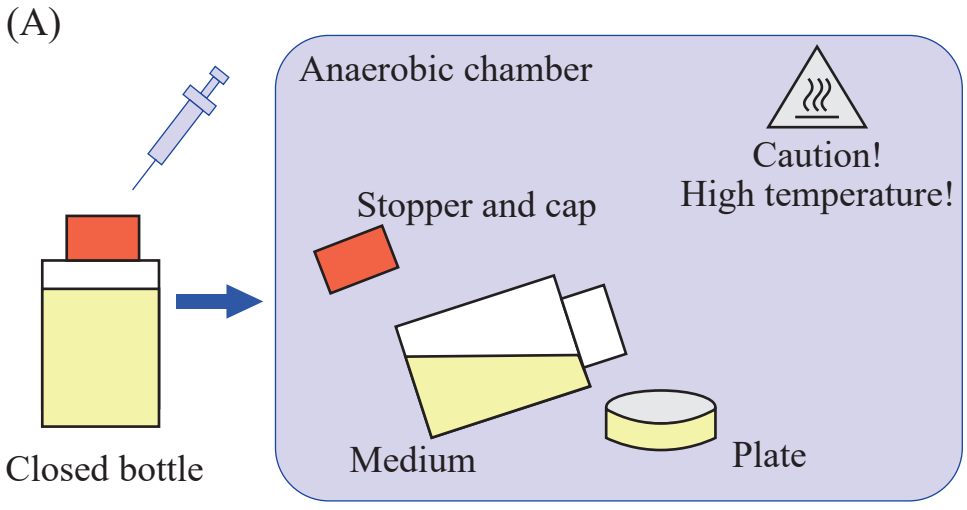
Figure 1. Experimental determination of the unbiased mutation rate of the *Thermococcus eurythermalis* A501 is challenging because this archaeon has unusual physiology (i.e., obligate anaerobic and obligate hyperthermophilic). (A) The preparation of anaerobic high temperature tolerant gelrite plate. After sterilization and polysulfide addition via syringe, the plates are made in an anaerobic chamber. (B) The incubation of the strain *T. eurythermalis* A501 at 85℃ in liquid medium. (C) The initiation of mutation accumulation (MA) by spreading cells from a single founding colony to 100 lines. Plates are placed in an anaerobic jar for incubation in strictly anaerobic condition at 85℃. (D) The MA process followed by whole genome sequencing and data analysis. Single colony of each line is transferred to a new plate for N times (here N=20). (E) Base-substitution mutations and insertion/deletion mutations across the whole genome of *T. eurythermalis*. The dashed vertical line separates the chromosome and plasmid. The height of each bar represents the number of base-substitution mutations across all MA lines within 10 kbp window. Green and red triangles denote insertion and deletion, respectively. The locus tags of the 14 genes with statistical enrichment of mutations are shown.

Figure 2. The scaling relationship involving the base-substitution mutation rate per cell division per site (μ), the estimated effective population size (N_e), and genome size across 28 bacterial and two archaeal species. All three traits' values were logarithmically transformed. The mutation rates of these species are all determined with the mutation accumulation experiment followed by whole genome sequencing of the mutant lines. The mutation rate of species numbered 1-29 (blue) is collected from literature and that of the species 30 (red) is determined in the present study. The numbered species: 1, *Agrobacterium tumefaciens*; 2, *Bacillus subtilis*; 3, *Burkhoderia*

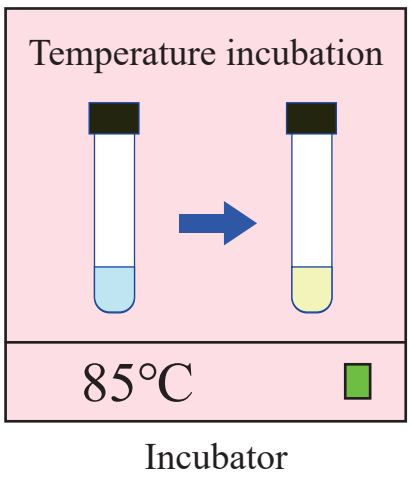
591 *cenocepacia*; 4, *Caulobacter crescentus*; 5, *Escherichia coli*; 6, *Haloferax volcanii*; 7,
592 *Janthinobacterium lividum*; 8, *Lactococcus lactis*; 9, *Mesoplasma florum*; 10, *Mycobacterium*
593 *smegmatis;* 11, *Pseudomonas aeruginosa*; 12, *Rhodobacter sphaeroides*; 13, *Ruegeria pomeroyi*;
594 14, *Salmonella enterica*; 15, *Staphylococcus aureus*; 16, *Staphylococcus epidermidis;* 17,
595 *Teredinibacter turnerae*; 18, *Vibrio cholerae*; 19, *Vibrio fischeri*; 20, *Vibrio shilonii*; 21,
596 *Arthrobacter* sp.; 22, *Colwellia psychrerythraea*; 23, *Deinococcus radiodurans*; 24,
597 *Flavobacterium* sp.; 25, *Gemmata obscuriglobus*; 26, *Kineococcus radiotolerans*; 27,
598 *Leeuwenhoekiella* sp.; 28, *Micrococcus* sp.; 29, *Nonlabens* sp.; 30, *Thermococcus eurythermalis*.
599 Among these, the species #6 *Haloferax volcanii* is facultative anaerobic halophilic archaeon, and
600 the species #30 is an obligate anaerobic hyperthermophilic archaeon. (A) The scaling
601 relationship between μ and N_e. (B) The scaling relationship between μ and genome size. (C) The
602 scaling relationship between genome size and N_e. Numbered data points 21-29 are not shown in
603 (A) and (C) because of the lack of population dataset for estimation of N_e. The dashed gray lines
604 and blue lines represent the generalized linear model (GLM) regression and the phylogenetic
605 generalized least square (PGLS) regression, respectively. The Bonferroni adjusted outlier test for
606 the GLM regression show that #7 *Janthinobacterium lividum* is an outlier in the scaling
607 relationship between μ and N_e, and #9 *Mesoplasma florum* is an outlier in the scaling relationship
608 between genome size and N_e. No outlier was identified in the PGLS regression results.
609

Figure 1

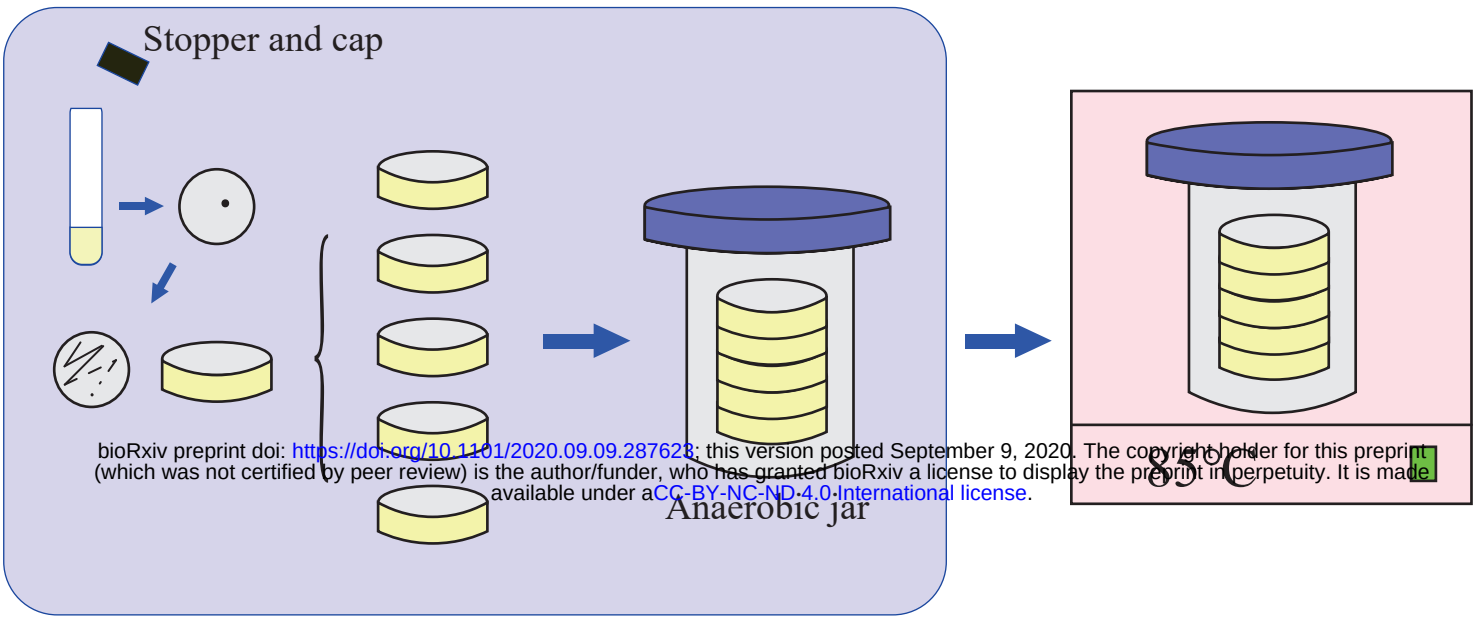
(A)



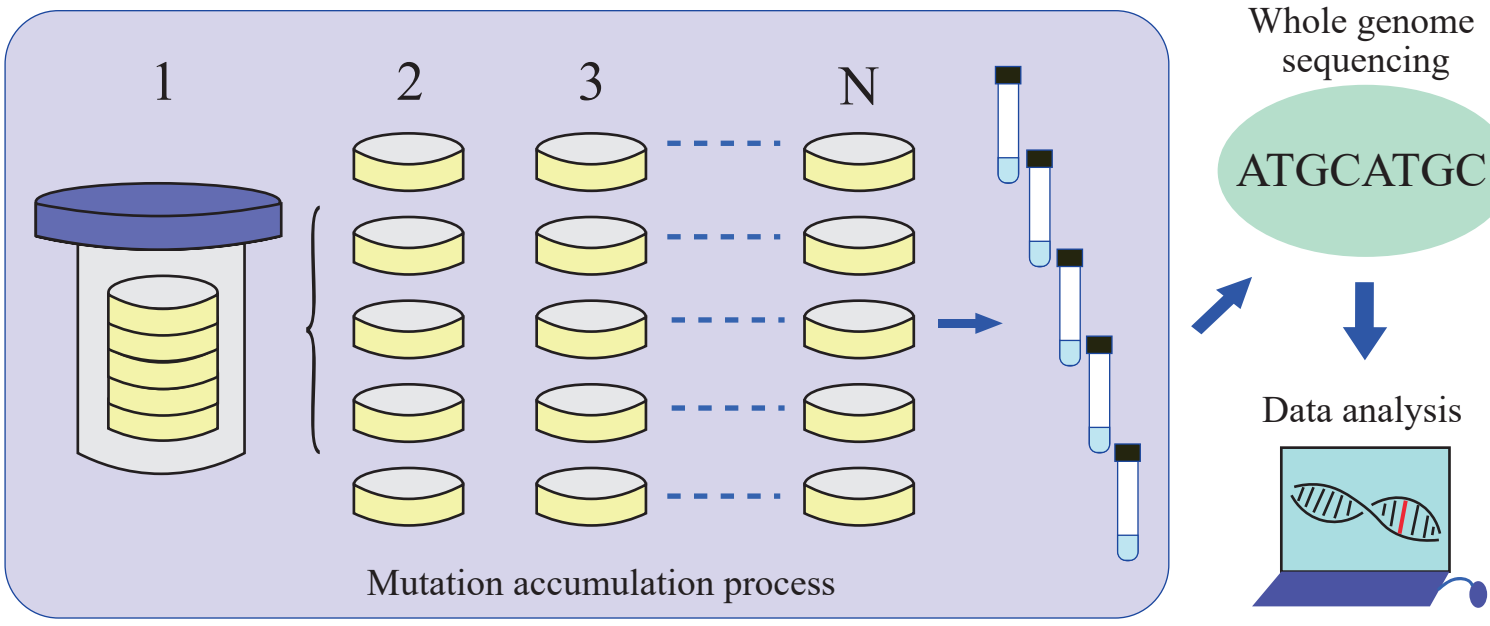
(B)



(C)



(D)



(E)

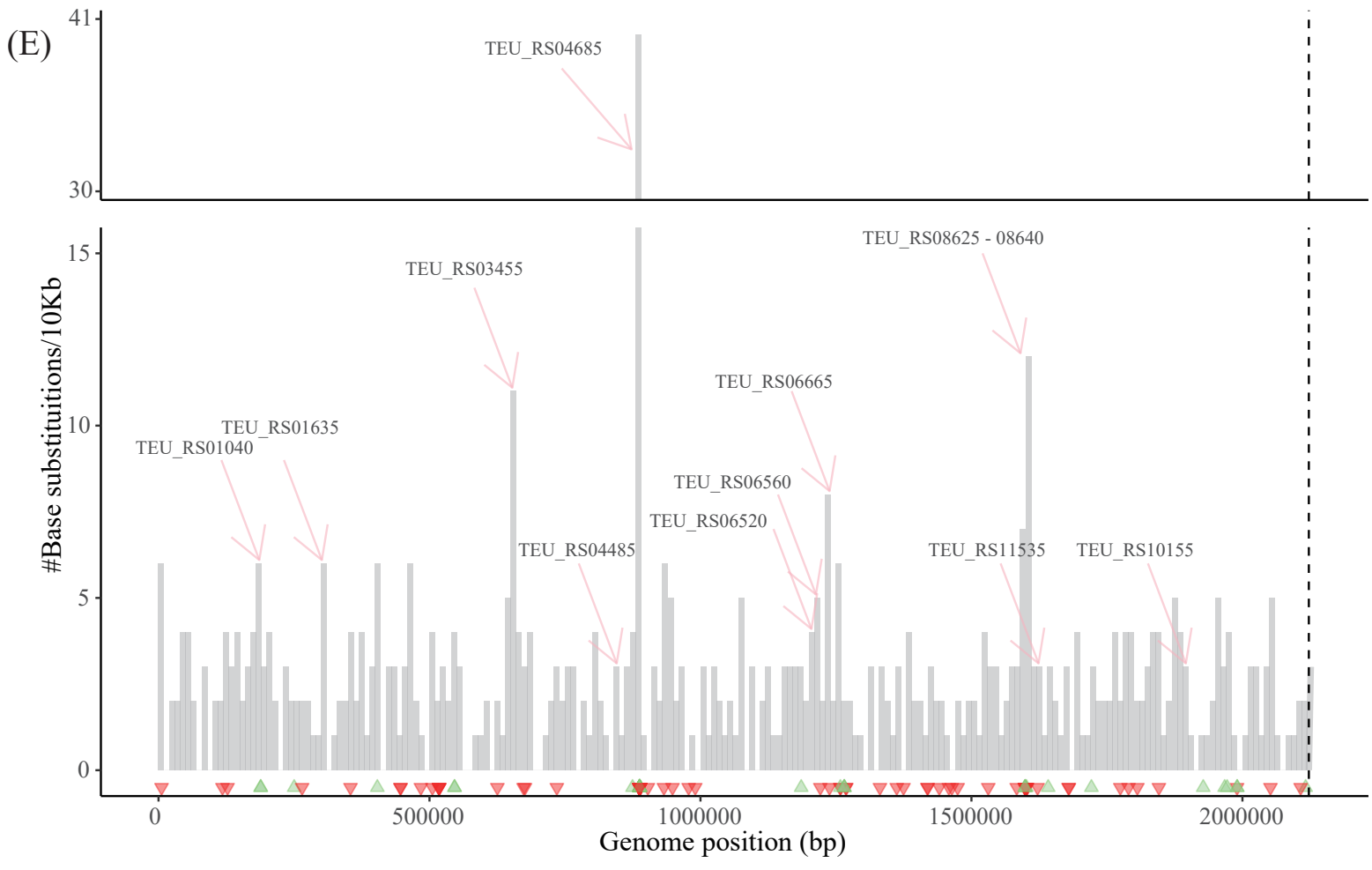


Figure 2