Estimation of the genome-wide mutation rate and spectrum in the archaeal species *Haloferax volcanii*

Sibel Kucukyildirim*,†, Megan Behringer*,§, Emily M. Williams*,§, Thomas G. Doak*,‡ and Michael Lynch*,§

*: Department of Biology, Indiana University, Bloomington, IN, US 47405
†: Department of Biology, Hacettepe University, Ankara, 06800 TURKEY
§: Biodesign Center for Mechanisms of Evolution, Arizona State University, Tempe, AZ, US 85287
‡: National Center for Genome Analysis Support, Indiana University, Bloomington, IN US 47405
Running Title: Mutational profile of *H. volcanii*

Keywords: *Haloferax volcanii*, mutation accumulation, mutation rate, mutation spectrum, AT bias, genome copy number

Corresponding Author:
Sibel Kucukyildirim
Hacettepe University
Department of Biology
Beytepe 06800 Ankara TURKEY
Phone: 90(535) 2096072

Email: sibelkucukyildirim@gmail.com
ABSTRACT

Organisms adapted to life in extreme habitats (extremophiles) can further our understanding of the mechanisms of genetic stability, particularly replication and repair. Despite the harsh environmental conditions they endure, these extremophiles represent a great deal of the earth’s biodiversity. Here, for the first time in a member of the archaeal domain, we report a genome-wide assay of spontaneous mutations in the halophilic species *Haloferax volcanii* using a direct and unbiased method: mutation-accumulation experiments combined with deep whole-genome sequencing. *H. volcanii* is a key model organism not only for the study of halophilicity, but also for archaeal biology in general. Our methods measure the genome-wide rate, spectrum, and spatial distribution of spontaneous mutations. The estimated base-substitution rate of $3.15 \times 10^{-10}$ per site per generation, or $0.0012$ per genome per generation, is similar to the value found in mesophilic prokaryotes (optimal growth at $\sim 20\text{-}45^\circ\text{C}$). This study contributes to a comprehensive phylogenetic view of how evolutionary forces and molecular mechanisms shape the rate and molecular spectrum of mutations across the tree of life.
INTRODUCTION

Spontaneous mutation is the ultimate source of novel genetic variation in nature and influences the genetic features of all populations. Thus, because mutation affects processes including evolution, inheritance, and genetic disorders, it is important to understand the intrinsic and extrinsic factors influencing the rate and molecular spectrum of mutations. Until recently, most strategies for determining the mutation rate and spectrum were primarily indirect, including interspecies comparison of putatively neutral sites in specific genes (GRAUR AND LI 2000; WIELGOSS et al. 2011) and analyses using reporter-construct genes (DRAKE 1991). However, because selection can affect synonymous sites (LAWRIE et al. 2013), mutation rates can vary significantly across different genomic regions (FOSTER et al. 2013), and only a fraction of reporter-construct mutations are detectable (DRAKE 1991), both of these methods are likely to have significant biases.

In long-term mutation-accumulation (MA) experiments, replicate lines are taken through regular population bottlenecks to minimize the efficiency of selection, allowing cell lines to accumulate mutations in an unbiased fashion (KIBOTA AND LYNCH 1996; LYNCH et al. 2016). By applying high-throughput sequencing to mutation-accumulation experiments (MA-WGS), it has become possible to directly identify these mutations (LYNCH et al. 2008; KEIGHTLEY AND HALLIGAN 2009). As a result, unbiased estimates of the genome-wide rate and molecular
spectrum of spontaneous mutations have been generated for a large number of
However, in order to understand how DNA replication and repair cooperate and
ultimately determine the genome-wide mutation rate across the entire tree of life,
it is necessary to expand these experimental assays to the unexplored archaeal
domain.

The moderately halophilic (optimal NaCl 1.7–2.5 M) and mesothermophilic (30-
49°C) archaeon Haloferax volcanii has a moderately GC-rich genome (65.6%),
containing a primary 2.848 Mb chromosome, three smaller chromosomes, pHV1,
3, 4, and the pHV2 plasmid (HARTMAN et al. 2010). H. volcanii can be cultured
relatively easily—rare for an extremophile, and is a model organism not only for
the study of halophilicity, but also for archaeal biology in general. In the last
decade, H. volcanii has been widely studied in efforts to understand archaeal
DNA replication/repair and transcription mechanisms. For example, it has been
found that the cell machinery involved in replication and repair in eukaryotes is
more closely related to that of archaea than bacteria (BARRY AND BELL 2006).
Previous in vitro studies suggesting the possibility that the fidelity of archaeal
DNA polymerases is higher than that of its bacterial counterparts (KUNKEL 1992;
CLINE et al. 1996; BLOOM et al. 1997; GROGAN et al. 2001) provide further
motivation for the current study.

As is characteristic in many clades of Euryarchaeota (e.g., Halobacteriales,
Methanosarcinales, Thermococcales, and Methanococcales), Haloferax volcanii
– while genetically haploid – has been shown to carry a variable number of genome copies, with 2-45 genome copies per cell depending on the growth phase (BREUERT et al. 2006; ZERULLA et al. 2014). It has been proposed that high genome copy numbers may have selective advantages for prokaryotes, especially those living in extreme environments. The possible advantages of many genome copies in prokaryotes include efficient DNA repair by homologous recombination, higher rates of transcription/translation in nutrient limiting conditions, restrained phenotypic expression of deleterious recessive mutations (BREUERT et al. 2006; ZAHRADKA et al. 2006), and the use of genomic DNA as a phosphate storage polymer (ZERULLA et al. 2014). Some of these effects might either increase or decrease indirect mutation-rate estimations.

Previous mutation-rate estimates in a few Archaeal species have been derived using indirect methods, such as fluctuation tests, which rely on the production of a phenotype to indicate the random mutation of a known locus or loci and are sometimes coupled with Sanger sequencing: Haloferax volcanii (MACKWAN et al. 2007), Halobacterium salinarum (BUSCH AND DI RUGGIERO 2010), Sulfolobus acidocaldarius (JACOBS AND GROGAN 1997; GROGAN et al. 2001), and Thermus thermophilus (MACKWAN et al. 2008). In this study—for the first time to our knowledge—we directly determine the rate, spectrum, and distribution of genome-wide mutation in an archaeon, thereby enhancing understanding of how different phylogenetic backgrounds influence the mutation process.
MATERIALS AND METHODS

Mutation Accumulation

One hundred independent *H. volcanii* DS2 (ATCC 29605) MA lines were initiated from a single colony. *Halobacterium* medium 974, as recommended by the ATCC, was used for the mutation-accumulation line transfers. Every week a single colony from each MA line was transferred by streaking to a new plate, ensuring that each line regularly passed through a single-cell bottleneck. This bottlenecking procedure ensures that mutations accumulate in an effectively neutral fashion (Muller 1928; Bateman 1959; Mukai 1964; Kibota and Lynch 1996). MA lines were incubated at 30°C under aerobic conditions and each line passed through ~3000 cell divisions (Supplementary Table S1). Frozen stocks of all lineages were prepared by growing a final colony per isolate in 1 ml *Halobacterium* broth incubated at 30°C, and freezing in 20% glycerol at −80°C.

The number of generations undergone by each MA line per transfer was determined based on the average number of cells in a colony. Specifically, each month, one single colony randomly chosen from at least ten MA lines was transferred to a sterile tube with basal salt solution (medium without carbon source), vortexed, serially diluted and plated. After 6 days of incubation at 30°C, colony forming units (CFU; denoted as *N* here) of the diluted cultures were counted and averaged. The number of generations (*n*) was calculated by $n = \log_2 N$. 


**DNA Extraction and Sequencing**

Lines surviving to the end of the MA experiment were prepared for whole-genome sequencing: DNA was extracted with the Wizard® Genomic DNA Purification kit (Promega, Madison, Wisconsin, USA) and DNA libraries for Illumina HiSeq 2500 sequencing (insert size 300 bp) were constructed using the Nextera DNA Sample Preparation kit (Illumina, San Diego, CA). Paired-end 150 nt read sequencing of MA lines was done by the Hubbard Center for Genome Studies, University of New Hampshire, with an average sequencing depth of 126× across all lines (Supplementary Table S1).

**Mutation Identification and Analyses**

Adaptors of paired-end reads were removed with Trimmomatic 0.32 (BOLGER et al. 2014); trimmed reads were mapped to the reference genome (National Center for Biotechnology Information (NCBI) accession numbers: NC_013964-NC_013968), using BWA, version 0.7.12 (LI AND DURBIN 2009). The output was parsed with SAMTOOLS (LI et al. 2009); we also applied duplicate-read removal using picardtools-1.141, and read realignment around indels using GATK 3.6, before performing SNP and insertion-deletion discovery with standard hard filtering parameters described in GATK Best Practices recommendations (Phred-scaled quality score QUAL > 100 and RMS mapping quality MQ > 59 for both variant and non-variant sites; ploidy setting: diploid higher ploidy did not change mutation detection) (MCKENNA et al. 2010; DEPRISTO et al. 2011). Base substitutions and small indels were called using the Haplotype Caller in GATK. In
order to call a variant, a minimum of ten reads was needed and 99% of reads in
a line were required to call the line-specific consensus nucleotide at a site; a ~1%
cut off was set to allow for aberrant reads originating from sequencing errors,
contamination of pure indexes during library construction, or barcode degeneracy
during sequence de-multiplexing. All novel homozygous variants and
heterozygous variants (mutations that are segregating within an individual) that
were supported by at least three reads for both alleles were kept. Presence of
structural variants (transpositions, deletions, inversions and duplications) were
called using GRASPER (LEE et al. 2016) and RetroSeq (KEANE et al. 2013), while
relative ploidy level within an MA line was determined by creating sequencing
coverage maps of 10,000 bp sliding windows with BEDTools (QUINLAN AND HALL
2010). All structural variants were confirmed by visual examination using IGV
(ROBINSON et al. 2011).

Statistics and Calculations

The per site per cell division mutation rate ($\mu$) was calculated for each line as $\mu = \frac{m}{(nT)}$, where $m$ is the number of observed mutations, $n$ is the number of sites
analyzed in the MA line, and $T$ is the number of generations for the line. The
standard error of the mean mutation rate was calculated with the equation $\text{SEM} = \frac{\text{SD}}{\sqrt{n}}$, where SD is the standard deviation of mutation rates of each line.

The expected GC content at mutation equilibrium was calculated as (LYNCH
2007): $\frac{\mu_{A/T \rightarrow G/C}}{\mu_{G/C \rightarrow A/T} + \mu_{A/T \rightarrow G/C}}$, where $\mu_{A/T \rightarrow G/C}$ is the rate of mutations at A:T sites
resulting in an A/T → G/C change (including A/T → G/C transitions and A/T → C/G transversions), and \( \mu_{G/C\rightarrow A/T} \) is the rate of mutations at G:C sites resulting in a G/C → A/T change (including G/C → A/T transitions and G/C → T/A transversions) (Supplementary Table S1). Mutation bias in the G/C direction was calculated by \( \frac{\mu_{A/T\rightarrow G/C}}{\mu_{G/C\rightarrow A/T}} \). We used R v3.1.0 (R DEVELOPMENT CORE TEAM 2014) for all statistical tests. 95% Poisson confidence intervals were calculated using the Poisson test in R. While testing if the observed conditional base-substitution mutation spectra differed between chromosomes, the cutoff \( p \) value was determined using Bonferroni correction, i.e. \( p = 0.05/n \), where \( n \) equals 6 – the number of conditional base substitution types.

A Consideration of Genome Copy Number Effect on MA Experiment

Outcomes

Because *H. volcanii* is reported to maintain a large number of genome copies (2-45 genome copies/cell) (ZERULLA et al. 2014), a newly arising neutral mutation needs time to reach fixation within a cell. In a single-cell lineage, the average time to fixation is approximately two times the ploidy for a newly arisen selectively neutral mutation (KIMURA AND OHTA 1969). Thus, if genomes are randomly transmitted to new cells, and each generation starts from a single cell, conditional on drifting to fixation, a new mutation in *H. volcanii* should reach fixation in 4-90 generations. Considering that each *H. volcanii* MA line passed through an average of 3000 cell divisions in this study, heterozygous mutations should be a very small fraction of the total (even if genome copy number were
as high as 45, heterozygous mutations should still be a small fraction (~0.05 (45 genome copies x 3866751 nucleotide sites per genome copy x 3.15 * 10^-10 base-pair substitutions per site) and of negligible significance if they were undetected. Furthermore, even if the genome-copy number increases at some stage of colony development, the 2-genome stage acts as a bottleneck to accelerate the fixation of a mutation within a lineage. Therefore, the number of fixed mutations per generation is very close to the mutation rate per generation. However, as a conservative approach, we conducted a secondary estimation of mutation rate taking heterozygous variants into account. Here, heterozygous variants were scaled by their proportion of supporting reads where \( m_{scaled} = \frac{\text{# of reads identifying the mutation}}{\text{# of reads mapping to the site}} \) and mutation rate was then recalculated for each sample as \( \mu_{ploidy} = \frac{\sum m_{scaled}}{nT} \).

**Data Availability**

Workflow deposited at the GitHub repository (https://github.com/sibelkucukyildirim) and raw sequences are available at the Sequence Read Archive at NCBI (Bioproject No.: PRJNA386190).
RESULTS

To estimate the mutation rate in *H. volcanii* DS2, a mutation-accumulation experiment was carried out for 37 months (~3000 generations) with 100 independent lineages, all derived from a common progenitor colony of *H. volcanii*. Every week, a single colony from each line was re-streaked onto a fresh plate with transfers producing colonies from probably just a single cell. Such treatment minimizes the effective population size, such that all but lethal mutations should accumulate in an effectively neutral manner (Kibota and Lynch 1996). The number of generations undergone by each MA line per transfer was determined based on the number of cells in the colonies at the time of transfer (23.6 ± 6.4 generations per transfer). There was no significant change in the number of cells per colony throughout the experiment, suggesting that the number of generations per transfer stayed constant. The 57 lines that survived over 122 transfers—to the end of the MA experiment—were prepared for whole-genome sequencing. Of these lines, 54 had sufficient depth of coverage (>156x) to calculate the mutation rate and spectrum. After sequencing and detecting mutations in those 54 MA lines, we noted that three mutations were shared by all the MA lines, indicating that those mutations were present in the ancestor line and were removed from further downstream analysis (Supplementary Table S1).

**Mutation Rate**

Across 54 *H. volcanii* MA lines, we identified 167 single base-substitution changes (Supplementary Tables S2), yielding an overall base-substitution
mutation rate of $3.15 \times 10^{-10}$ (SE = $0.27 \times 10^{-10}$) per site per generation, or 0.0012 per genome per generation. In addition to single base-substitutions we also identified 14 multinucleotide mutations (MNM; defined as multiple mutations that occur within 50 nucleotides in the same MA line) (SCHRIDER et al. 2011) (Supplementary Table S3). Including these 14 MNMs increases the overall base-substitution mutation rate to $3.24 \times 10^{-10}$ (SE = $0.28 \times 10^{-10}$) per site per generation. As noted in the Material and Methods section, because the *H. volcanii* genome is known to be transiently maintained in a polyploid state of up to 45 genome copies we repeated our analysis to include heterozygous mutations (only the heterozygous variants that were supported by at least three reads for both alleles). This analysis revealed 12 heterozygous mutations, all of which were multinucleotide mutations (Supplementary Table S3). These heterozygous mutations represent a very small fraction of our data (~0.02) and account for a negligible difference in the estimated mutation rate. Thus, the mutation rate in *H. volcanii* (per haploid genome) is similar to mutation rates commonly observed in mesothermophilic bacteria (LONG et al. 2018).

We also found 19 short insertions and deletions 1-30 bps in length (2 insertions, and 17 deletions), yielding an insertion/deletion rate of $3.58 \times 10^{-11}$ (SE = $0.82 \times 10^{-11}$) per site per generation (Supplementary Tables S2 and S4). These small indel events comprise 10.2% of all accumulated mutations in *H. volcanii*, consistent with results from many other microbial MA experiments (SUNG et al. 2016). The deletion rate for *H. volcanii* is ~8.5× greater than the insertion rate
resulting in a small net rate of loss of genomic sites, consistent with a suggested universal prokaryotic deletion-bias hypothesis (MIRA et al. 2001) and 21% of the small indels occur in simple sequence repeats (SSRs), e.g. homopolymer runs (Supplementary Table S4).

One of the goals of this study was to examine whether mutation rates and spectra differ between the main chromosome and the mini chromosomes (pHV1, pHV3 and pHV4), which differ in size and content. The heterogeneity of overall base-substitution mutation rates of each chromosome was evaluated with a χ² test. We were unable to reject the null expectation that the number of substitutions is proportional to the number of sites covered on each chromosome (χ²=7.34, p=0.062) (Supplementary Tables S5-S8 and Figure S1). This may not appear to be surprising, given that all chromosomes are expected to be replicated and repaired by the same mechanisms.

Using the annotated *H. volcanii* DS2 genome, we identified the functional context of each base substitution (Supplementary Table S9). Across the 54 sequenced MA lines, 120 of the 167 (71.9%) substitutions occurred in coding regions (~86% of the genome consists of coding sequence), while the remaining 47 were found at non-coding sites (Supplementary Table S9). Given the codon usage and transition/transversion ratio in *H. volcanii*, the expected ratio of nonsynonymous to synonymous mutations is 2.77 in the absence of selection, which is not significantly different from the observed ratio of 2.24 (83/37)
\[\chi^2 = 1.06, \ p = 0.3\]. Thus, selection does not appear to have had a significant influence on the distribution of mutations in this experiment, as has been shown repeatedly for the MA/WGS method (LEE et al. 2012; DILLON et al. 2015; LONG et al. 2015a; KUCUKYILDIRIM et al. 2016; SENRA et al. 2018).

**Structural Variants**

We called medium/large deletions, transpositions, inversions, and tandem duplications using a combination of GRASPER (LEE et al. 2016), RetroSeq (KEANE et al. 2013) and IGV (ROBINSON et al. 2011). We detected 24 total structural variants consisting of 3 large deletions on the main chromosome, and 3 inversions and 18 large-deletions on the plasmid pHV4. Large deletions were the most common form of structural variation (21 events) accounting for a genome-wide large-scale deletion rate of 3.77 x 10^-11 per site per generation (Supplementary Table S10). Of these large-scale deletions, 19 were mediated by recombination between insertion sequence (IS) elements.

IS elements are short mobile elements in prokaryotic (bacterial and archaeal) genomes, initially identified by their capacity to generate mutations. Previous studies have identified IS-element related mutations as mostly deleterious or selectively neutral (KIDWELL AND LISCH 2001; CHANDLER AND MAHILLON 2002). The *H. volcanii* genome has two major IS element types (ISH51 and the newly described IS4-type elements), both of which are members of the IS4 family (HOFMAN et al. 1986; HARTMAN et al. 2010). IS-element mediated genomic
rearrangement has previously been reported in *Halobacterium salinarum* (SAPIENZA *et al.* 1982).

In *H. volcanii*, IS-elements are unevenly distributed amongst the chromosomes with pHV4 harboring more than the other chromosomes (HARTMAN *et al.* 2010). Thus, we looked to see if a chromosomal bias existed for IS-mediated deletions. If IS-mediated deletions were distributed evenly between the main chromosome and pHV4 with respect to size, we would expect 85% of the deletions to occur on the main chromosome and 15% of the deletions on pHV4. However, we see the opposite relationship, i.e. a strong bias for IS-mediated deletion towards pHV4 ($\chi^2=45.296$, $p=1.23 \times 10^{-11}$) Our results are consistent with the highly dynamic nature of archaeal genomes (REDDER AND GARRETT 2006; BRIDGER *et al.* 2012), and indicate that a high deletion rate related to IS elements may be affected by genome composition and by the fact that various IS elements have different activities.

**Mutation Spectrum**

Across the 54 MA lines, there were 98 transitions and 69 transversions, resulting in a transition/transversion ratio of 1.42. There are 76 G:C$\rightarrow$A:T transitions and 33 G:C$\rightarrow$T:A transversions at GC sites, yielding a mutation rate in the AT direction of $\mu_{G/C\rightarrow A/T} = 3.16 \times 10^{-10}$ per G:C site per generation. In contrast, 22 A:T$\rightarrow$G:C transitions and 4 A:T$\rightarrow$C:G transversions yielded a mutation rate in the GC direction of $\mu_{A/T\rightarrow G/C} = 1.40 \times 10^{-10}$ per A:T site per generation.
(Supplementary Table S2 and Figure 1), which is significantly lower than the
$\mu_{G/C\rightarrow A/T}$ rate (95% Poisson confidence intervals for $\mu_{G/C\rightarrow A/T}$ are $2.6\text{--}3.8 \times 10^{-10}$
and, for $\mu_{A/T\rightarrow G/C}$ are $0.9\text{--}2.0 \times 10^{-10}$). As such, the expected GC-content under
mutation-drift equilibrium is $0.31\pm0.04$ (SEM), lower than the genome-wide GC-
content (0.66).

We performed a $\chi^2$ test to investigate whether the observed conditional base-
substitution mutation spectra differ between chromosomes (main chr/pHV1: $\chi^2$=
4.73, df=5, $p=0.45$; main chr/pHV3: $\chi^2$=5.72, df=5, $p=0.33$; main chr/pHV4:
$\chi^2$=12.6, df=5, $p=0.03$, after the Bonferroni correction, the cutoff $p$ value =
0.05/6= 0.0083). These comparisons show that there is no significant
chromosomal variation in the overall base-substitution spectra of *H. volcanii*
(Supplementary Tables S5-S8 and Figure S1).

Spontaneous DNA damage, such as deamination of cytosine to uracil and C5
methylcytosine (m5C) to thymine, or oxidation of guanine to 7,8-dihydro-8-
oxoguanine (8-oxoG) can lead to mutational biases. Methylated bases are well-
known mutational hotspots in bacteria (COULONDRE *et al.* 1978; DUNCAN AND
(2015) found that *H. volcanii* DS26 contains methyltransferases that modify two
motifs in the genome: GCA$_{m6}$BN$_{8}$VTGC (B stands for C or G or T, V stands for
A or C or G and N stands for any base) and C$_{m4}$TAG. Although the percentage
of methylation of these sites is high in the *H. volcanii* genome (77% and 28%
respectively) (OUELLETTE *et al.* 2015), we found no elevation of the mutation rate
at the above noted methylated sites. We also examined bacterial canonical and
non-canonical sites (e.g., GATC, CTAG, GACC and CACC)—previously
suggested as methylation sites (CLARK *et al.* 2012; LONG *et al.* 2015a)—but found
no indication of mutation-rate elevation.

DISCUSSION

The biological processes related to mutation (prevention, production, and repair)
vary among genetic contexts. This variation leads to biases in the types,
locations, and numbers of genetic changes that occur. As a consequence,
organisms with different genome compositions may experience different types of
genetic variation. As such, determining variation in the mutation process is
important to understanding how genome composition affects neutral and
adaptive sequence evolution.

Analyses of spontaneous mutation have shown that most mesothermophilic
bacteria share two mutational characteristics: (i) total genome-wide mutation
rates in the general range of 0.0025 to 0.0046 per genome per generation (DRAKE
1991; DRAKE *et al.* 1998; LYNCH *et al.* 2016), with the variation potentially
explained by the influence of the drift barrier (KUO AND OCHMAN 2009; SUNG *et al.*
2012; LYNCH *et al.* 2016); and (ii) base substitutions comprising ~83% of the total
mutations (SUNG *et al.* 2016). However, previous analyses of Archaea based on
reporter constructs estimated a *H. volcanii* genomic mutation rate of 0.00045 per
genome per generation, ~10-fold lower than the consensus value of 0.004
observed in other mesothermophilic prokaryotes as well as a predominance of small insertion-deletion events (GROGAN et al. 2001; MACKWAN et al. 2007; MACKWAN et al. 2008).

It has been observed that mutation-rate estimates can differ between MA experiments and specific-locus methods, even with carefully chosen reporter genes (Table 1) and these differences are often significant (paired t-test, two-tailed $p = 0.029$) (DRAKE 2012). These discrepancies likely arise in part because mutation rates vary among individual nucleotide types (LONG et al. 2015a; KUCUKYILDIRIM et al. 2016; LONG et al. 2018), depending on their neighbor content (SUNG et al. 2012), and vary among chromosomal regions (DILLON et al. 2015; LONG et al. 2015b; SUNG et al. 2015; NICCUM et al. 2019). In addition, only a fraction of reporter-construct mutations are molecularly detectable (DRAKE 1991).

Thus, for any organism, a single-locus will hardly ever provide an unbiased estimate of the genomic mutation rate or molecular spectrum. As next-generation sequencing becomes increasingly less expensive, and higher-quality reference genomes are now available for a large number of microbial species, MA combined with WGS provides a more comprehensive and nearly unbiased description of spontaneous mutations (for the given environmental setting) across the entire genome as supported by our crosschecks on synonymous versus nonsynonymous base substitutions. This reduces the risk of spurious conclusions based on a single gene and avoids other methodological issues associated with reporter-gene approaches (e.g., mutation detectability) for mutation-rate measurement. We report a base-substitution mutation rate for $H.$
*volcanii* of $3.15 \times 10^{-10}$ per base pair per generation, nearly three-fold higher than the previously reported rate of $1.08 \times 10^{-10}$ based on the *pyrE2* reporter locus (Mackwan et al. 2007), but still lower than the average mesophilic bacterial base-substitution rate (Figure 2). Moreover, we do not observe an excess of insertion-deletion mutations, contrary to a previous estimate for *Haloferax volcanii* (Mackwan et al. 2007), which is the only study to date exhibiting this pattern in any organism. The reason for the observed difference between the two mutation rate estimates may be that not all mutations could be molecularly detected at the reporter locus (~75%).

**Possible Effects of Temperature**

In principle, mutation rates can be influenced by a number of factors including growth conditions, growth medium, number of generations, temperature, and genome copy number (ploidy level). Accordingly, growth temperature may contribute to the differences observed in the results for our study, conducted at 30°C, and the previous study of the *Haloferax volcanii* mutation rate, conducted at 37°C and 41°C (Mackwan et al. 2007). An extensive literature with a wide range of organisms including viruses, bacteria, microbial eukaryotes, insects and plants has shown that extremely high temperatures are stressful and may result in very low growth/survival rates and have mutagenic effects (Zamenhof and Greer 1958; Drake 1966; Lindgren 1972). Recent work by Chu et al. (2018) has shown that for *E. coli*, within a range of unstressful temperatures, the genome-wide spontaneous base-pair substitution rate is 1.5-fold higher at 37°C.
than at the cooler, suboptimal growth temperatures of 25 and 28°C. If this relationship between temperature and base-pair substitution rate extends to *H. volcanii* then we would expect to observe a lower base-pair substitution rate at 30°C than at 37°C/41°C, not higher. Therefore, it seems unlikely that temperature is the main factor contributing to the differences in mutation-rate estimates in *H. volcanii* reported here and in the previous study (MACKWAN *et al.* 2007).

**Possible Effects of Genome Copy Number**

Another feature with a potential to affect the mutation process is genome copy number, which can vary among prokaryotes as a function of phyla and growth rate. As in other haloarchaeal species (such as *Halobacterium salinarum* and *Haloferax mediterranei*), *Haloferax volcanii* can maintain multiple copies (2-45) of its genome, sometimes at high copy-numbers (ZERULLA *et al.* 2014), which may result in detection of fewer mutations in short-term surveys that depend on phenotype and Sanger sequencing to estimate base-pair substitution rates (MACKWAN *et al.* 2007). All else being equal, the number of mutations accumulated in diploid cells is expected to be 2-fold higher than haploid cells, as they have two sets of homologous chromosomes, and more targets for mutation (CROW AND KIMURA 1965). However, unlike haploid cells, diploids have continual access to homologous DNA. This access may reduce the rate of DNA damage, but the presence of homologous DNA may also increase the probability of spontaneous structural variants and rearrangements. As recently shown in yeast, diploid cells have greater replication fidelity (for single nucleotide changes) but are more prone to large-scale mutations with
deleterious fitness effects (SHARP et al. 2018). Thus, ploidy level may affect the mutational profile.

In an MA experiment, newly arising mutations in organisms with many genome copies will be present at low frequency and may be incorrectly filtered by the hard-mapping criteria. However, we found 12 heterozygous mutations (Supplementary Table S10) which represent a very small fraction of our data (~0.02 even if we consider them independent events) and account for a negligible difference in the estimated mutation rate. This negligible contribution of heterozygotes is expected based on population-genetic theory: if \( n \) equals the copy number of a chromosome then \( \frac{1}{n} \) is both the frequency of a mutation on the chromosome when it initially occurs as well as the probability of the random fixation of that mutation within the genome. When these values are combined with the per site mutation rate, \( \mu \), then the product is the total fixation probability of a mutation \( \mu n \cdot \frac{1}{n} = \mu \), or the mutation rate. Even if the genome-copy number increases is as high as 45 during colony growth, the single-cell bottleneck that occurs during streaking will effectively fix mutations on average within 45 generations assuming random segregation of chromosomes to progeny, so the number of heterozygous base pair substitutions per MA line at anytime is ~0.05. Taken together, the time needs for intracellular fixation of any novel base-pair mutation is short relative to total length of the MA experiment.
Our study demonstrates for the first time the spectrum of spontaneous mutations in an archaeon. A fundamental question that arises is how intrinsic forces normally shape the GC content of genomes. Here, in the moderately GC-rich *H. volcanii*, we observed that mutations are biased in the AT direction (\( \mu_{G/C \rightarrow A/T} = 3.16 \times 10^{-10} \) vs. \( \mu_{A/T \rightarrow G/C} = 1.39 \times 10^{-10} \)) (Supplementary Table S2), consistent with near-universal AT mutation bias (HERSHBERG AND PETROV 2010). As recently discussed in LONG et al. (2018), genome-wide nucleotide composition variation among species is a consequence of both interspecific differences in mutation bias and the efficiency of selection or biased gene conversion for different nucleotides, and a GC-rich genome in species with AT mutational bias is not particularly surprising.

Interestingly, we did not observe increased mutation rates for nucleotide sequences known to be methylation targets in *H. volcanii*. Mutation rates at these sites are expected to be higher, especially if there is mutational bias towards AT, as G/C nucleotides are more prone to DNA damage due to oxidative stress and spontaneous deamination. However, uracil DNA glycosylases have been noted to protect the genome against deamination in GC-rich bacteria (VENKATESH et al. 2003) and may explain why we found no elevation of the mutation rate at methylation sites in this organism.

In conclusion, by using the MA approach combined with WGS, we have shown
that *H. volcanii* has a mutation rate similar to the consensus value of other prokaryotic organisms, as well as a commonly observed A/T biased mutational spectrum. More precise genetic and biochemical assays are necessary to clarify the mechanisms maintaining genomic stability in Archaea.

**ACKNOWLEDGEMENTS**

This study is funded by a Multidisciplinary University Research Initiative award (W911NF-09-1-0444) from the US Army Research Office and National Institutes of Health awards (R01GM036827, R35-GM122566, and F32-GM123703). We thank Samuel F. Miller for technical support and Hongan Long for helpful discussions.
Table 1. Examples for base-substitution mutation rate ($\mu$) estimates by using whole genome sequencing of MA experiments and reporter-construct estimates.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Base-substitution mutation rate (x10$^{-10}$ per site per generation)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MA</td>
<td>Reporter-construct</td>
</tr>
<tr>
<td>Archaea</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Halofax volcanii</em></td>
<td>3.15</td>
<td>1.1</td>
</tr>
<tr>
<td>Eubacteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>3.28</td>
<td>7.4</td>
</tr>
<tr>
<td><em>Deinococcus radiodurans</em></td>
<td>4.99</td>
<td>16.9</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>2.23</td>
<td>2.6</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>0.79</td>
<td>8.4</td>
</tr>
<tr>
<td>Unicellular eukaryotes</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Plasmodium falciparum</em></td>
<td>2.45</td>
<td>20.8</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>1.67</td>
<td>7.2</td>
</tr>
<tr>
<td><em>Schizosaccharomyces pombe</em></td>
<td>2.00</td>
<td>8.2</td>
</tr>
</tbody>
</table>
Figure 1. Conditional base-substitution mutation rates for *H. volcanii* MA lines.

Rates for each base-substitution type normalized by genome-base composition; error bars indicate the standard error.
Figure 2. The relationship of the base-substitution mutation rate per site per generation with total haploid genome size.

Data points correspond to the following species: At: Agrobacterium tumefaciens (SUNG et al. 2016), Bs: Bacillus subtilis (SUNG et al. 2015), Bc: Burkholderia cenocepacia (DILLON et al. 2015); Dr: Deinococcus radiodurans (LONG et al. 2015a), Ec: Escherichia coli (LEE et al. 2012), Go: Gemmata obscuriglobus (LONG et al. 2018), Hv: Haloferax volcanii (this study), Ms: Mycobacterium smegmatis (KUCUKYILDIRIM et al. 2016), Pa: Pseudomonas aeruginosa (DETTMAN et al. 2016), Sa: Staphylococcus aureus (LONG et al. 2018), Se: Staphylococcus epidermidis (SUNG et al. 2016) Vf: Vibrio fischeri and Vc: Vibrio cholera (DILLON et al. 2017). All data are derived from MA-WGS projects.
REFERENCES

Behringer, M. G., and D. W. Hall, 2016 Genome-wide estimates of mutation rates and spectrum in *Schizosaccharomyces pombe* indicate CpG sites are highly mutagenic despite the absence of DNA methylation. G3 (Bethesda) 6: 149-160.


Hofman, J. D., L. C. Schalkwyk and W. F. Doolittle, 1986 ISH51: a large, degenerate family of insertion sequence-like elements in the genome of the


Sung, W., M. S. Ackerman, M. M. Dillon, T. G. Platt, C. Fuqua et al., 2016 Evolution of the insertion-deletion mutation rate across the tree of life. G3 (Bethesda) 6: 2583-2591.
Sung, W., M. S. Ackerman, J.-F. Gout, S. F. Miller, E. Williams et al., 2015 Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. Mol Biol Evol 32: 1672-1683.
Wielgoss, S., J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming et al., 2011 Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with Escherichia coli. G3 (Bethesda) 1: 183-186.