


Evolution of a Record-Setting AT-Rich Genome: Indel Mutation, Recombination, and Substitution Bias

Duong T. Nguyen[†], Baojun Wu[†], Shujie Xiao[†], and Weilong Hao ^{*}

Department of Biological Sciences, Wayne State University

*Corresponding author: E-mail: haow@wayne.edu.

[†]These authors contributed equally to this work.

Accepted: 19 September 2020

Data deposition: GenBank accession numbers KU888694–KU888699, MT461932, MT461933, MT534625, and MT534626.

Abstract

Genome-wide nucleotide composition varies widely among species. Despite extensive research, the source of genome-wide nucleotide composition diversity remains elusive. Yeast mitochondrial genomes (mitogenomes) are highly A + T rich, and they provide a unique opportunity to study the evolution of AT-biased landscape. In this study, we sequenced ten complete mitogenomes of the *Saccharomyces ludwigii* yeast with 8% G + C content, the lowest genome-wide %(G + C) in all published genomes to date. The *S. ludwigii* mitogenomes have high densities of short tandem repeats but severely underrepresented mononucleotide repeats. Comparative population genomics of these record-setting A + T-rich genomes shows dynamic indel mutations and strong mutation bias toward A/T. Indel mutations play a greater role in genomic variation among very closely related strains than nucleotide substitutions. Indels have resulted in presence–absence polymorphism of tRNA^{Arg} (ACG) among *S. ludwigii* mitogenomes. Interestingly, these mitogenomes have undergone recombination, a genetic process that can increase G + C content by GC-biased gene conversion. Finally, the expected equilibrium G + C content under mutation pressure alone is higher than observed G + C content, suggesting existence of mechanisms other than AT-biased mutation operating to increase A/T. Together, our findings shed new lights on mechanisms driving extremely AT-rich genomes.

Key words: extreme AT, GC-biased gene conversion, mobile intron, repeat, yeast.

Significance

Genomes with extreme nucleotide composition help us understand mechanisms in genome evolution. This study reports the most AT-rich genome sequenced to date. Comparative genomic analyses show that both mutation and nonmutation mechanisms drive the most extreme genome-wide AT-richness.

Introduction

Genome-wide nucleotide composition has long been recognized to vary across species (Belozersky and Spirin 1958), yet the source of nucleotide composition diversity among genomes remains elusive (Agashe and Shankar 2014; Bohlin and Pettersson 2019). A detailed understanding of base composition evolution requires knowledge of molecular and evolutionary processes including mutation, recombination, natural selection, and random genetic drift (Lynch 2007). Fortunately, comparative genomics of organisms with

extensive nucleotide composition diversity provides a powerful tool for understanding the underlying mechanisms and genomes with extreme base composition offer unique insights (Gardner et al. 2002; McCutcheon and Moran 2010; Smith et al. 2011; Su et al. 2019). Nearly six decades ago, Sueoka (1961) reported that guanine + cytosine (G + C) content of genomic DNA varies approximately from 25% to 75% in bacteria. Recent bacterial genome sequencing has widened the range to 13.5–75.3% (McCutcheon and Moran 2010; Chen et al. 2019) thanks primarily to the

© The Author(s) 2020. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

discovery of extremely A + T-rich genomes in symbiotic bacteria (Moran et al. 2008). Mitochondrial genomes (mitogenomes) evolved from an endosymbiotic alphaproteobacterium (Lang et al. 1999) and are a rich reservoir for unconventional and extreme genomes (Smith and Keeling 2015). Currently sequenced mitogenomes exhibit a wide range of G + C content from 10.9% to 68.1% (Bouchier et al. 2009; Hecht et al. 2011) with most mitogenomes being AT rich (Smith 2012). G + C contents in yeast mitogenomes range from 10.9% in *Nakaseomyces bacillisporus* to 52.7% in *Candida subhashii* (Fricova et al. 2010).

There are several driving forces in the evolution of genome-wide nucleotide composition. First, biases in the mutation process result in diversity of base composition. Despite mutation is mostly biased toward A/T (Hershberg and Petrov 2010), reduced mutation bias is evident in many G + C-rich genomes (Long et al. 2015, 2016; Sun et al. 2017). Second, G + C content can be influenced by G/C-biased gene conversion in recombination (Eyre-Walker 1993). In eukaryotes, G/C-biased gene conversion leads to an increase of GC contents in regions with high recombination activity (Mancera et al. 2008; Pessia et al. 2012). In bacteria and archaea, it has been debated on whether recombination acts as a universal force to increase (Lassalle et al. 2015) or decrease (Bobay and Ochman 2017) G + C content. Third, natural selection often favors increased G + C in protein-coding genes (Hildebrand et al. 2010; Raghavan et al. 2012; Long et al. 2018). In reverse, reduced effective population size and selection efficacy result in decreased G + C. In fact, genomes having the most A + T-rich genes in bacteria (McCutcheon and Moran 2010), mitochondria (James et al. 2013), and plastids (Su et al. 2019) are all in endosymbiotic organisms. Finally, genome-wide nucleotide composition can be shaped by variable length of AT-rich intergenic regions (Xiao et al. 2017). The *Drosophila melanogaster* mitogenome, with the lowest genome-wide %(G + C) in fruit flies, has exceptionally long AT-rich control region (Lewis et al. 1994). *Nakaseomyces bacillisporus* (10.9% G + C), the most AT-rich mitogenome until this study, contains considerable AT-rich intergenic sequences (Bouchier et al. 2009).

In this study, we report the lowest genome-wide %(G + C) in the mitogenome of *Saccharomyces ludwigii*, which is a non-*Saccharomyces* spoilage yeast species in winemaking (Ciani and Maccarelli 1998), and production of low- and non-alcoholic beers (De Francesco et al. 2015). To understand the evolution of this record-setting AT-rich mitogenome, we obtain complete mitogenomes from a total of ten *S. ludwigii* strains. Taking comparative genomic approaches, we investigate genomic variation regarding sequence insertions/deletions (indels) and intron movements. We further detect mitochondrial recombination and examine sequence characteristics of recombinant regions. Finally, we analyze substitution bias and estimate equilibrium G + C content to

investigate mechanisms dictating the genome-wide nucleotide landscape.

Results

Extreme AT-Richness of the *S. ludwigii* Mitogenome

The mitogenome of the *S. ludwigii* type strain Y12793 shows 8.42% overall G + C (more genomic details in [supplementary table S1, Supplementary Material](#) online). Intergenic regions exhibit extremely low %(G + C), introns and genes show a striking elevation in %(G + C) (fig. 1). The intergenic regions of notably elevated %(G + C) are all GC-clusters, which form hairpin structures and are presumably mobile (de Zamaroczy and Bernardi 1986) ([supplementary fig. S1, Supplementary Material](#) online). The ten *S. ludwigii* mitogenomes range from 64,578 to 69,040 nucleotides in size. Their G + C contents range from 7.63% to 8.42% averaging 8.00% ([supplementary table S2, Supplementary Material](#) online).

G + C contents in protein genes were compared between *S. ludwigii* and *Hanseniaspora uvarum* in the Saccharomycodaceae family, and against seven species in the Saccharomycetaceae family (fig. 2). *Saccharomyces ludwigii* has significantly lower G + C content in protein genes than *H. uvarum* (Wilcoxon rank sum test, P value = 0.014), but there is no significant difference between *S. ludwigii* versus any Saccharomycetaceae species.

Sequence Evolution in *S. ludwigii*

To compare substitution rates among nine yeast species, seven mitochondrial respiratory protein genes were analyzed. The ribosomal protein gene *rps3* (or *var1*) was excluded due to its extremely high level of sequence variation, which can cause inaccurate sequence alignment among divergent sequences. The branch leading to *S. ludwigii* is 5.9 times or 8.3 times shorter in length than that leading to *H. uvarum* based on synonymous (dS) or nonsynonymous (dN) site divergence, respectively. Similarly, the branch leading to *S. ludwigii* is notably shorter than that leading to *Eremothecium gossypii* on both dS and dN phylogenetic trees. There is no evidence for accelerated rates of substitution in the *S. ludwigii* mitogenome.

Saccharomyces ludwigii does not show elevated nucleotide diversity relative to other yeast species. The average pairwise synonymous nucleotide diversity (π_s) is 0.017 and nonsynonymous nucleotide diversity (π_a) is 0.0002. Compared against species in the Saccharomycetaceae family with >5 mitogenomes from our previous study (Xiao et al. 2017), *S. ludwigii* has π_s and π_a values lower than *Saccharomyces cerevisiae* ($\pi_s = 0.033$, $\pi_a = 0.0029$), *Saccharomyces paradoxus* ($\pi_s = 0.032$, $\pi_a = 0.0019$), *Lachancea kluyveri* ($\pi_s = 0.032$, $\pi_a = 0.0013$), but higher than *Lachancea thermotolerans* ($\pi_s = 0.005$, $\pi_a = 0.0001$). The π_a/π_s ratio, as a proxy for the level of genetic drift, in

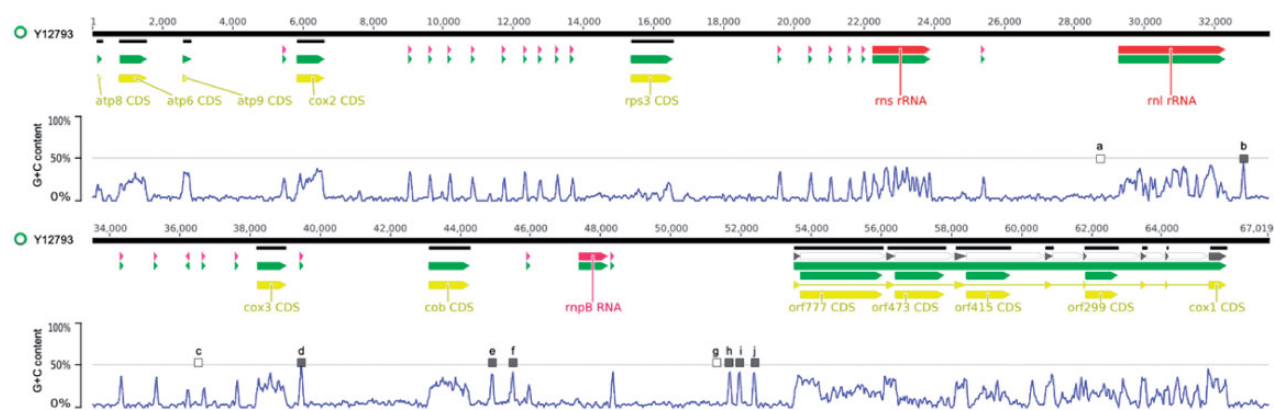


FIG. 1.—Mitogenome map of *Saccharomyces ludwigii* Y12793 (type strain). The genome is circular but presented in a linear view to better illustrate G + C content variation (measured by 50-nucleotide sliding windows) along the genome. GC-rich nongenic regions (%G + C > 35) are often polymorphic and shown in squares (a–j) with filled squares signifying presence in the 12793^T mitogenome, whereas open squares indicating absence in the 12793^T mitogenome. tRNA^{Arg} (ACG) (labeled as d) is included because it is polymorphic among the ten sampled strains (fig. 4).

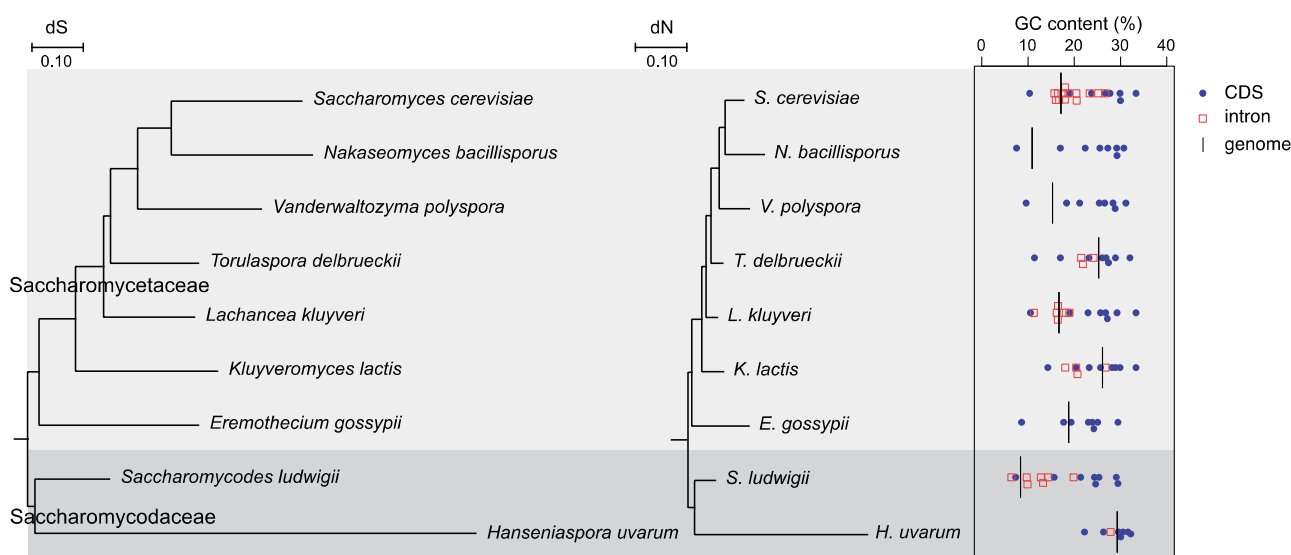


FIG. 2.—Phylograms of synonymous (dS) and nonsynonymous (dN) site divergence for the concatenated alignment of seven respiratory protein genes. G + C content of protein-coding genes (CDSs), introns, and the whole mitogenome is shown on the right for each species.

S. ludwigii (0.012) is lower than the ratio in *Saccharomyces cerevisiae* (0.089), *Saccharomyces paradoxus* (0.060), or *L. kluveri* (0.041). These findings suggest that mitochondrial genes in *S. ludwigii* are not under more relaxed selection than in other yeast species.

Densities of Short Tandem Repeats

Saccharomyces ludwigii has the lowest density of mononucleotide repeats (smaller than 1-bp repeats per kb) among yeast mitogenomes, which contradicts the expected density (fig. 3). The densities of mononucleotide repeats are also low in *H. uvarum* (smaller than 1-bp repeats per kb), *E. gossypii* (2.8-bp repeats per kb), *Vanderwaltozyma polyspora* (3.9-bp

repeats per kb), and *N. bacillisporus* (6.4-bp repeats per kb). All except *E. gossypii* and *H. uvarum* have higher than expected densities of dinucleotide repeats. All mitogenomes have higher than expected densities of trinucleotide repeats. All but *H. uvarum* have higher than expected densities of tetra-nucleotide repeats. *Saccharomyces ludwigii* and *N. bacillisporus* are consistently the two species with highest densities of dinucleotide, trinucleotide, and tetra-nucleotide repeats.

Movements of Mitochondrial Introns

Among the eight mitochondrial intron sites in *S. ludwigii*, five are group I introns and three are group II introns

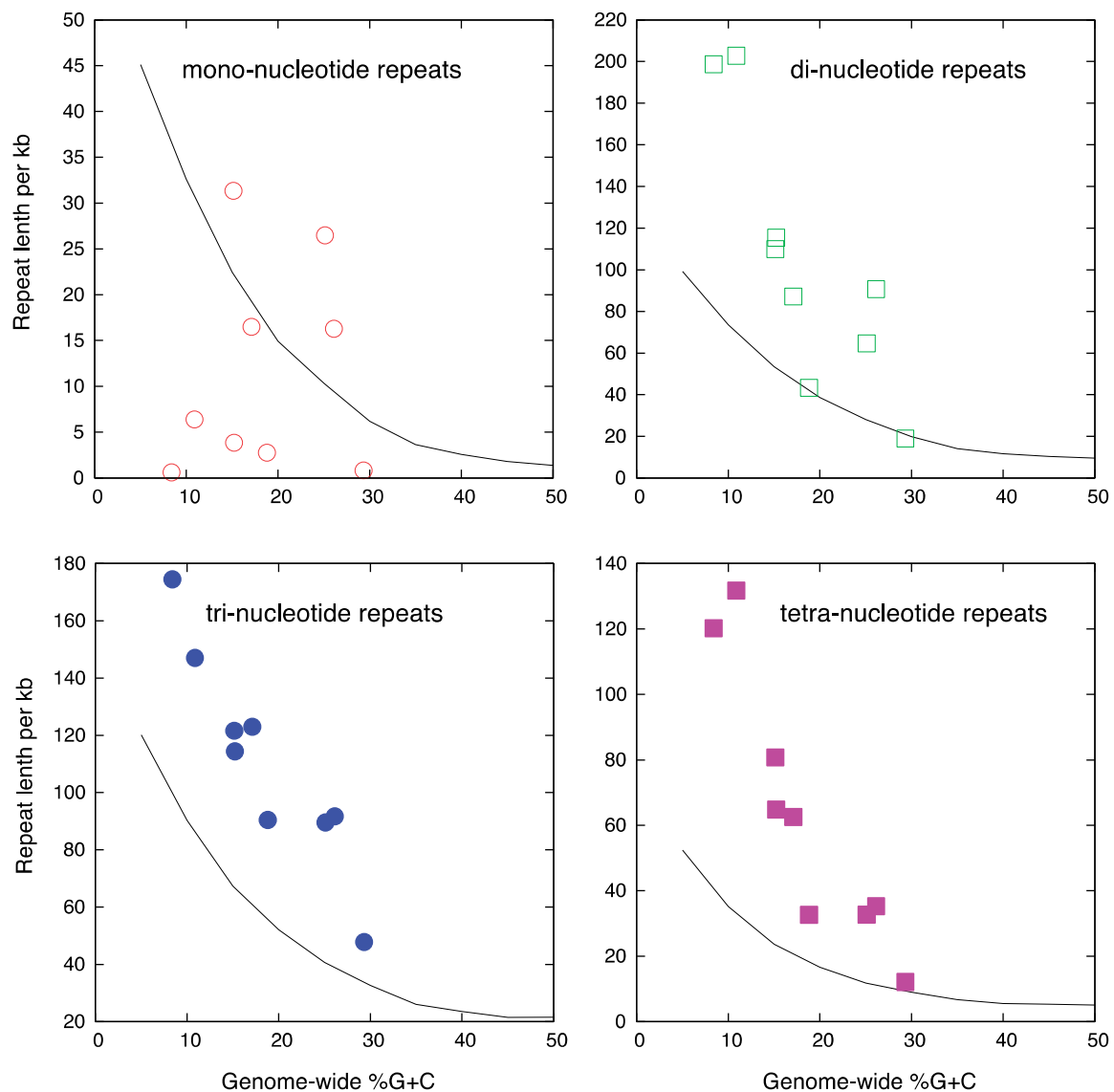


FIG. 3.—Correlation between short repeat densities (repeats per 1 kb) and G + C content. The line in each panel shows the expected repeat densities based on G + C content. The species in each panel from left to right are *Saccharomyces ludwigii*, *Nakaseomyces bacillisporus*, *Lachancea kluyveri*, *Vanderwaltozyma polyspora*, *Saccharomyces cerevisiae*, *Eremothecium gossypii*, *Torulaspora delbrueckii*, *Kluyveromyces lactis*, and *Hanseniaspora uvarum*.

(supplementary table S3, Supplementary Material online). The *cox1* intron 2, intron 3, and intron 5 contain open reading frames (ORFs) that encode endonucleases, the *cox1* intron 1 contains a maturase-encoding ORF. The omega intron in *ml* rRNA and *cox1* intron 1 is presence–absence polymorphic (fig. 4). The omega intron is present in four strains (NCYC732, Y12861, Y12860, and NCYC3345), which do not form a monophyletic group but are identical at the sequence level. The *cox1* intron 1 is sporadically distributed in five strains. Among the five sequences, four of them (Y12793, Y94, Y8871, and NCYC3345) are identical with Y12860 differing by just a single nucleotide. The identical or

nearly identical intron sequences suggest events of recent intron invasion.

The presence–absence polymorphism of mitochondrial introns has been associated with mitochondrial recombination (Wu and Hao 2014). To detect recombination, phylogenetic analyses were performed on the six introns present in all ten *S. ludwigii* strains. Among the *cox1* intron 2 sequences, the most basal strain Y5447 is identical with NCYC3345, and both are more closely related to the clade containing the type strain Y12793, than to another basal strain Y12860 (fig. 5). The genome-wide nucleotide distance between Y5447 and NCYC3345 is 0.008 (supplementary fig. S2, Supplementary

Material online). If we assume that the occurrence of mutations in the mitogenome follows a Poisson distribution, the probability to have identical 1731 nucleotides (the length of *cox1* intron 2) between Y5447 and NCYC3345 is 1.01×10^{-6} . Among the *cox1* intron 7 sequences, the basal strain Y12860 is identical with three strains (Y12793, Y974, and Y8871). Based on the genome-wide nucleotide distance (0.018) between Y12860 and Y12793, the probability to have identical 1,202 nucleotides (the length of *cox1* intron 7) between Y12860 and Y12793 is 4.16×10^{-10} . The *cox1* intron 2 is an ORF-containing group I intron, whereas *cox1* intron 7 is an ORF-lacking group II intron. This is consistent with our previous findings that mitochondrial recombination is common in both group I and group II introns, and

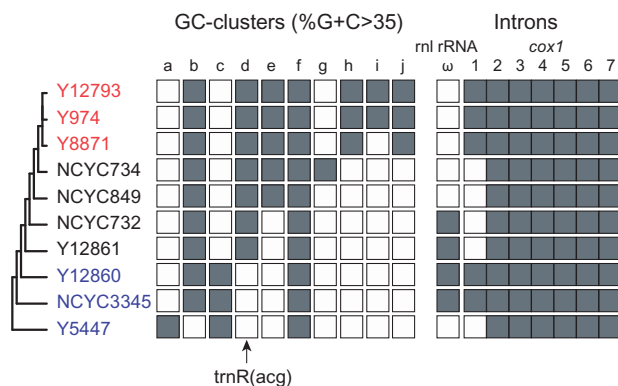


FIG. 4.—Distribution of GC-cluster regions and mitochondrial introns in *Saccharomyces ludwigii*. Filled squares signify presence, whereas open squares indicate absence. The d region is the only GC-rich region with an annotated function (tRNA^{Arg} with an ACG anticodon). A cladogram on the left is based on the phylogeny constructed from the intron-free mitogenome alignment after the removal of indels and recombinant regions. For illustration purpose, the three strains lacking tRNA^{Arg} (ACG) are in blue, the Y12793 type strain and its two closely related strains are in red.

recombination is not dependent on the presence of intron-encoded endonuclease or maturase (Wu and Hao 2014; Wu et al. 2015). Unexpectedly high sequence similarity and phylogenetic incongruence are common in the remaining introns (supplementary fig. S2, Supplementary Material online), suggesting widespread recombination in *S. ludwigii* introns.

Presence–Absence Polymorphism of tRNA^{Arg} (ACG)

Eight out of nine GC-clusters exhibit presence–absence polymorphism among *S. ludwigii* mitogenomes (fig. 4). The arginine tRNA gene tRNA^{Arg} (ACG) was analyzed along with GC-clusters due to its presence–absence polymorphism and relatively high G + C content (fig. 1). Unlike the dominant arginine tRNA gene tRNA^{Arg} (TCT) that is universally present, tRNA^{Arg} (ACG) is sporadically distributed in yeast mitogenomes (table 1). The presence of tRNA^{Arg} (ACG) seems to be associated with the usage of CGN codons (recognized by tRNA^{Arg} [ACG]) in the *rps3* gene in *Saccharomyces cerevisiae*, *Torulaspora delbrueckii*, and *L. kluyveri*. In *S. ludwigii*, there is no usage of CGN codons for arginine in any strains regardless of the presence or absence of tRNA^{Arg} (ACG) (table 1). The three strains (Y5447, NCYC3345, and Y12860) that lack tRNA^{Arg} (ACG) are basal lineages on the rooted *S. ludwigii* tree (supplementary fig. S3, Supplementary Material online). The tRNA^{Arg} (ACG) region does not show evidence of sequence degeneration in any of the three tRNA^{Arg} (ACG)-lacking strains (supplementary fig. S4, Supplementary Material online), instead, it suggests rapid clean gain/loss of tRNA^{Arg} (ACG). The most parsimonious explanation would be that tRNA^{Arg} (ACG) was acquired once after the *S. ludwigii* species evolved. However, the possibility of multiple tRNA^{Arg} (ACG) losses cannot be ruled out if tRNA^{Arg} (ACG) loss takes place at a much higher rate than tRNA^{Arg} (ACG) gain.

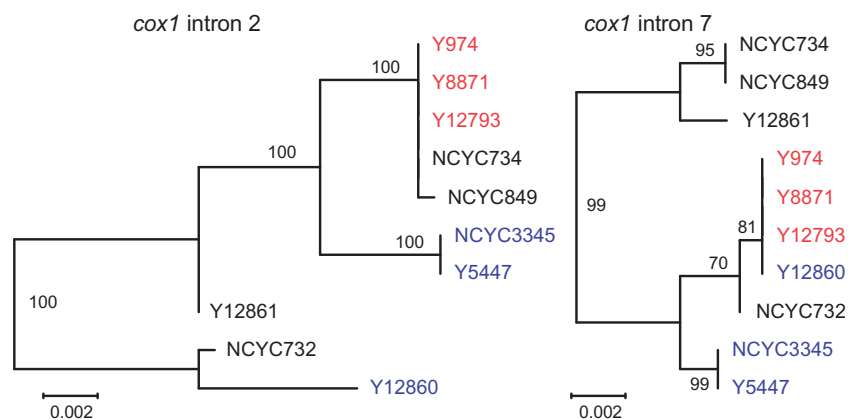


FIG. 5.—Phylogenetic evidence for recombination of *cox1* intron 2 and *cox1* intron 7. Bootstrap values >70% are shown. The strain colors are per figure 3.

Table 1

Usage of CGN and AGR Codons for Arginine in Yeast Mitogenomes

Species/Strains ^a	Presence of tRNA ^{Arg} (ACG)	Seven Respiratory Protein Genes		Ribosomal Protein Gene (<i>rps3</i>)	
		AGR Codons	CGN Codons	AGR Codons	CGN Codons
<i>Saccharomyces cerevisiae</i>	Yes	39	0	3	1
<i>Nakaseomyces bacillisporus</i>	No	35	0	3	0
<i>Vanderwaltozyma polyspora</i>	No	36	0	4	0
<i>Torulaspota delbrueckii</i>	Yes	39	0	3	2
<i>Lachancea kluyveri</i>	Yes	40	0	4	2
<i>Kluyveromyces lactis</i>	No	40	0	7	0
<i>Eremothecium gossypii</i>	No	34	0	6	0
<i>Hanseniaspora uvarum</i> ^b	No	40	0	–	–
<i>Saccharomyces ludwigii</i>					
Y974	Yes	35	0	3	0
Y12793	Yes	35	0	3	0
Y8871	Yes	35	0	3	0
NCYC734	Yes	35	0	3	0
NCYC849	Yes	35	0	3	0
NCYC732	Yes	35	0	3	0
Y12861	Yes	35	0	3	0
Y12860	No	35	0	3	0
NCYC3345	No	35	0	3	0
Y5447	No	35	0	3	0

^aSpecies/strains are shown following the phylogenetic order in figures 1 and 6.

^bThe *rps3* gene is absent in *Hanseniaspora uvarum*.

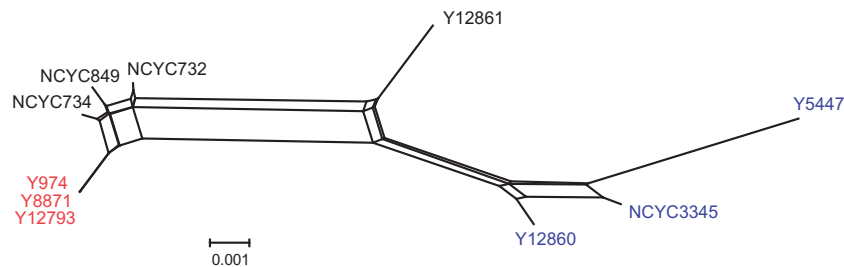


Fig. 6.—Phylogenetic network of *Saccharomyces ludwigii* strains. The Neighbor-Net dendrogram was calculated on the intron-free mitogenome alignment using SplitsTree. The strain colors are per figure 4.

Genome-Wide Analysis of Mitochondrial Recombination

We removed introns and indels from the alignment before recombination analyses. There is significant evidence for recombination in the *S. ludwigii* mitogenome (PHI-test [Bruen et al. 2006], P value = 0). A phylogenetic network is shown in figure 6 to illustrate reticulate evolution. Recombinant events were inferred by the four-gamete test (Hudson and Kaplan 1985). Along the 51,469-nucleotide intron-free genome alignment, the regions that violate four-gamete test are 14,981 nucleotides (29.1% of the total alignment, [supplementary table S4, Supplementary Material](#) online). Overall, there are a significant amount of sequences in the *S. ludwigii* mitogenome associated with recombination. Stand-alone endonuclease genes in yeast mitogenomes can

cause recombination (Wu and Hao 2019), but *S. ludwigii* lacks mitochondrial-encoded stand-alone endonuclease genes. Therefore, other mechanisms are involved in mitochondrial recombination in *S. ludwigii*.

Variation of %G + C among Sequence Regions

Intron sequences in all ten strains have an average G + C content 11.9%. After the removal of introns, the G + C content of genome alignment drops to 7.35% ([supplementary table S4, Supplementary Material](#) online). Insertion/deletion (indel) regions in the genome alignment have a significantly lower G + C content (5.28%) than the entire alignment (χ^2 test, P value < 0.001). After the removal of indels, the G + C

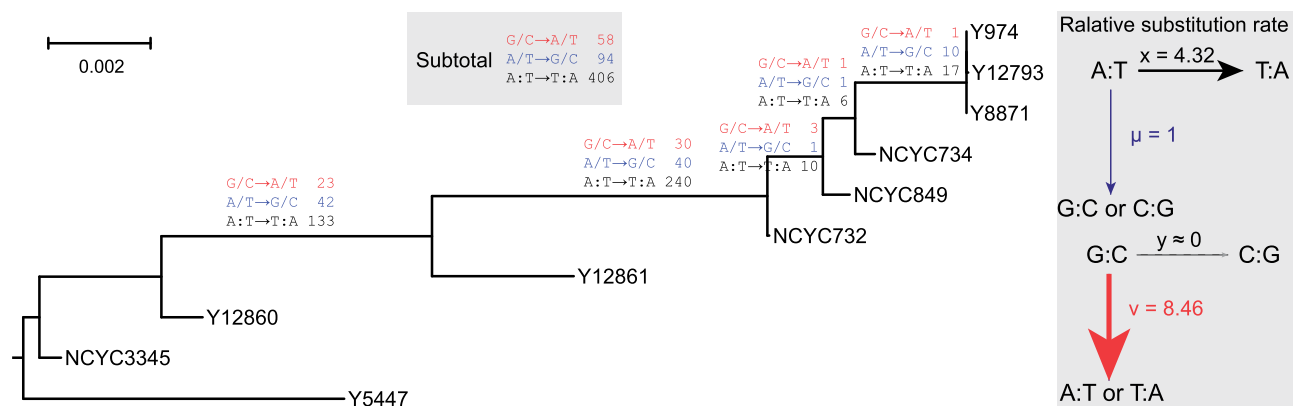


FIG. 7.—Maximum-likelihood phylogeny constructed from the intron-free mitogenome alignment after the removal of indels and recombinant regions. The tree is rooted by placing Y5447 as the most basal strain following the concatenated mitochondrial protein-gene tree in [supplementary figure S3, Supplementary Material](#) online. Substitutions are counted and shown on corresponding branches.

content of genome alignment becomes 7.56% ([supplementary table S4, Supplementary Material](#) online). Regions that violate four-gamete test have a significantly higher G + C content (9.51%) than the entire alignment (χ^2 test, P value < 0.001). This is consistent with previous reports on increased GC content in regions with high recombination activity (Mancera et al. 2008; Pessia et al. 2012). We propose that GC-biased gene conversion is also prevalent in mitochondrial recombination.

Indel Sequences among the Three Most Closely Related Mitogenomes

Among Y12793, Y974, and Y8871 mitogenomes, there are two indel regions, but no substitutions in the whole mitogenome alignment. A two-nucleotide indel is located in an (AT) dinucleotide repeat starting at alignment position 12274. Y12793 has (AT)₁₀ (i.e., ten AT units), whereas Y974 and Y8871 have (AT)₉. The other indel has 292 nucleotides in alignment region 51918–52209. The indel sequence is present in Y12793 and Y974, but absent from Y8871. Its first 74 nucleotides have 52.70% G + C and correspond to GC-cluster *i* in figure 4, whereas the remaining 218 nucleotides have only 3.67% G + C. These results show the importance of short tandem repeats and GC-clusters in indel mutations.

Mutation Bias and Expected Equilibrium %G + C under Mutation Bias

Nucleotide changes were mapped on the phylogeny (fig. 7) to infer mutation bias using the intron-free genome alignment excluding indels and recombinant regions. There are 58 changes from G/C to A/T, 94 from A/T to G/C, 406 from A/T to T/A, and no changes from C/G to G/C. Based on the empirical nucleotide composition (6.77% G + C in the analyzed sequence alignment), the substitution rate of A/T → T/A is 4.32 time higher than that of A/T → G/C, the substitution rate of G/C → A/T is 8.46 time higher than that of A/T → G/C.

There is a very strong mutation bias toward A/T. The expected equilibrium G + C content under mutation pressure alone would be 10.54% with 95% confidence interval (7.86–14.20%), which is still higher than the observed G + C content (6.77%) of the analyzed sequence alignment. The phenomenon of observed G + C content being lower than expected equilibrium G + C content suggests the existence of evolutionary mechanism(s) operating to increase A/T composition.

Discussion

No Extreme Acceleration of Sequence Evolution in *S. ludwigii*

Unlike AT-rich endosymbiotic bacteria whose genes are under accelerated evolution and increased fixation of deleterious mutations (Moran 1996), the *S. ludwigii* mitogenome does not show evidence of accelerated evolution at either synonymous or nonsynonymous sites (fig. 2). *Saccharomyces ludwigii* has much lower substitution rates than *H. uvarum*, yet its protein genes have lower G + C contents than those in *H. uvarum*. This suggests that accelerated evolution under relaxed selection is not a major force driving the extreme AT-richness of the *S. ludwigii* mitogenome.

Limited Role of tRNA^{Arg} (ACG) in the *S. ludwigii* Mitogenome

Mitochondrial-encoded tRNA^{Arg} (ACG) has been recently acquired in *S. ludwigii*, but there is no evidence for usage of CGN codons in mitochondrial-encoded protein-coding genes in any of the *S. ludwigii* strains. tRNA^{Arg} (ACG) has undergone gain and loss events in the mitogenomes of the Saccharomycetaceae family, where the usage of CGN codons in the *rps3* gene is evident in tRNA^{Arg} (ACG)-containing species. The absence of CGN codon usage in the *S. ludwigii* mitogenomes suggests either that tRNA^{Arg} (ACG) is not

functional or that the usage of CGN codons has not been established after the recent acquisition of tRNA^{Arg} (ACG).

GC-Biased Gene Conversion in Mitochondrial Recombination

A significant amount of sequences in the *S. ludwigii* mitogenome are associated with recombination. In yeast nuclear genomes, recombination events increase G + C content in the converted sequences because of GC-biased gene conversion (Mancera et al. 2008). This study shows that recombinant regions have higher G + C content than nonrecombinant regions sequences in the *S. ludwigii* mitogenome. To the best of our knowledge, this is the first study to show evidence of GC-biased gene conversion in mitogenomes. Importantly, GC-biased gene conversion occurs during recombination even in the most AT-rich mitogenome.

Dynamic Indel Mutations

Indel mutations are major drivers in yeast mitogenome diversity, even among mitogenomes with no substitutions. There are two major types of indel mutations in yeast mitogenomes, presence–absence polymorphism, and length variation of short tandem repeats. Presence–absence polymorphism of introns and GC-clusters can be, in part, explained by their nature as mobile genetic elements (Goddard and Burt 1999; Wu and Hao 2015). Length variation of short tandem repeats can be caused by slippage and unequal recombination (Toth et al. 2000; Dieringer and Schlotterer 2003). The *S. ludwigii* mitogenome has substantially underrepresented mononucleotide repeats but high densities of other short tandem repeats. This is in sharp contrast to repeat densities in nuclear genomes, whose mononucleotide repeats are abundant (Toth et al. 2000). It is possible that mononucleotide repeats are strongly selected against in AT-rich yeast mitogenomes, and/or that repeats with >1-bp-repeated units undergo recent and very rapid expansions in extremely AT-rich and gene-poor mitogenomes such as in *S. ludwigii*.

AT-Biased Mutation in Mitogenomes

There is a very strong mutation bias toward A/T in the *S. ludwigii* mitogenome. Strong mutation bias toward A/T has been shown in many AT-rich mitogenomes such as *Drosophila melanogaster* (Haag-Liautard et al. 2008), *Caenorhabditis briggsae* (Howe et al. 2010), *Daphnia pulex* (Xu et al. 2012), and *Caenorhabditis elegans* (Konrad et al. 2017). The degree of mutation bias varies among *Paramecium* species, with a significant AT bias in the AT-rich *Paramecium caudatum* mitogenome (22% G + C), and little bias in mildly AT-rich *Paramecium* mitogenomes (39–42% G + C) (Johri et al. 2019). Interestingly, the mitogenome of budding yeast *Saccharomyces cerevisiae* is AT rich (with 17% G + C) but has been shown to have mutations biased

toward G/C (Lynch et al. 2008). Taken together, AT-rich mitogenomes often exhibit a strong AT mutation bias and there also exist species-specific factors dictating mutation bias.

Mutations from A/T to T/A

Mechanisms causing A/T to T/A substitutions in the *S. ludwigii* mitogenome are unclear. Mutation rates and patterns can be influenced by sequence context, especially in alternating purine–pyrimidine sequences (Hess et al. 1994; Sung et al. 2015). The frequent A → T mutations in this study can be associated with the abundant alternating A-T sequences (i.e., dinucleotide repeats) in the *S. ludwigii* mitogenome. Hypermutation tends to occur in A nucleotide sites in mammal immunoglobulin genes (Spencer and Dunn-Walters 2005; Zivojnovic et al. 2014). Though mutations in immunoglobulin genes and mitogenomes involve different DNA polymerase (η , PolH vs. γ , PolG) (Delbos et al. 2007), they all have frequent A → T mutations. Interestingly, A → T mutations in immunoglobulin genes are significantly favored in the alternating A-T sequence context TAT (with the focal base underlined) (Spencer and Dunn-Walters 2005; Zivojnovic et al. 2014).

Can AT-Richness Be Favored?

Natural selection often favors increased G + C in protein-coding genes (Hildebrand et al. 2010; Raghavan et al. 2012; Long et al. 2018). Interestingly, the observed G + C content in this study is lower than the expected equilibrium G + C content under mutation bias. This suggests existence of mechanisms promoting AT-richness at the mitogenome level in *S. ludwigii*. GC-biased gene conversion is evident among *S. ludwigii* mitogenomes. Our results, however, could not exclude AT-biased gene conversion (Bobay and Ochman 2017) especially if AT-biased gene conversion took place before the *S. ludwigii* species evolved. Similarly, AT-rich intracellular elements in bacteria are selectively favored because of limited availability and high costs of G + C nucleotides (Dietel et al. 2019). Base composition is known to be influenced by environmental factors (Foerstner et al. 2005; Reichenberger et al. 2015; Hellweger et al. 2018). Because yeast mitogenomes have different G + C contents from their corresponding nuclear genomes, any environmental factors influencing mitochondrial base composition must be at the subcellular level. The subcellular conditions between mitochondria and nucleus are indeed different in many ways. For instance, mitochondria have a temperature about 10 °C higher than the rest of the cell (Chretien et al. 2018), dNTP pools within mitochondria are not in equilibrium with cytosolic dNTP pools (Rampazzo et al. 2004). Finally, the life history of *S. ludwigii* is not well understood, further studies on its life history could shed new lights on the evolution of this extreme AT-rich mitogenome. With the remarkable mitogenomic diversity and readily available genetic tools to manipulate subcellular conditions in yeasts,

yeast mitogenomes become an excellent model system to investigate genetic mechanisms driving the evolution of nucleotide composition.

Materials and Methods

Mitogenome Sequencing, Assembly, and Annotation

Six strains of *S. ludwigii* (Y974, Y5447, Y8871, Y12793, Y12860, and Y12861) were kindly provided by the National Center of Agricultural Utilization Research (IL, USA). Four strains of *S. ludwigii* (NCYC732, NCYC734, NCYC849, and NCYC3345) were purchased from National Collection of Yeast Cultures (NCYC, UK). Genomic DNA of each strain was extracted from a 2-day culture of a single colony inoculation. The ten mitogenomes were sequenced at read depths ranging from 214× to 1431× by the Illumina MiSeq platform (Paired-End 250 bp) and assembled using SPAdes v3.7.1 (Bankevich et al. 2012) with k-mers of 55, 75, 89, 97, and 127. Each assembly was validated by mapping raw reads back to the assembly using BWA-MEMv0.7.12 (Li and Durbin 2009), following by visual inspection using Integrative Genomics Viewer (IGV v2.3.60) (Robinson et al. 2011). All ten mitogenomes were assembled into single, circularized genomes of length 64.6–69.0 kb with G + C content ranging from 7.63% to 8.42% (supplementary table S2, Supplementary Material online). The assembled genomes were annotated using MFannot (<http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl>), followed by manual correction of intron boundaries.

Sequence Alignment and Phylogenetic Analyses

The seven respiratory protein genes (*atp6*, *atp8*, *atp9*, *cob*, *cox1*, *cox2*, and *cox3*) were used for phylogenetic analyses. The ribosomal protein gene *rps3* (or *var1*) was excluded due to its extremely high level of sequence variation. Homologous sequences were aligned using MUSCLE (Edgar 2004). Phylogenetic trees were constructed using PhyML (Guindon et al. 2010) under a GTR + Γ nucleotide substitution model. The synonymous and nonsynonymous tree was constructed using SEAVIEW (Gouy et al. 2010). The concatenated mitochondrial protein-gene tree of *S. ludwigii* was rooted using *Kluyveromyces lactis* as the outgroup (supplementary fig. S2, Supplementary Material online). *Hanseniaspora uvarum* and *E. gossypii* were not chosen for the outgroup taxa as they have undergone accelerated evolution in overall substitutions (*H. uvarum*) or nonsynonymous substitutions (*E. gossypii*) (fig. 2). Whole mitogenomes were aligned using the Mauve program (Darling et al. 2004). Because mitochondrial introns are of presence–absence polymorphism (Wu and Hao 2014) and subject to homologous recombination (Wu et al. 2015), all introns were removed before genome alignment and analyzed separately.

Detection of Recombination

Existence of recombination was detected by the PHI-test using the software PhiPack (Bruen et al. 2006). The Neighbor-Net dendrogram was calculated using SplitsTree4 (Huson and Bryant 2006). The four-gamete test (Hudson and Kaplan 1985) is a method for detecting recombination events. Regions violating four-gamete test were removed using RminCutter.pl v1.05 (https://github.com/RILAB/rmin_cutter).

Analysis of Short Tandem Repeats

Short tandem repeats were counted when repeat unit >2 and repeat length >6. The minimum numbers of repeat units are 7 for mononucleotide, 4 for dinucleotide, 3 for trinucleotide, and 3 for tetra-nucleotide repeats. The expected repeat densities in a given %G + C were calculated on randomly generated 1,000-kb sequences by assuming a random distribution of nucleotides using custom PERL scripts.

Inference of Substitution Patterns on Genome Alignment

Substitution patterns were inferred only using nonrecombinant regions (i.e., not violating four-gamete test). A phylogeny was constructed and rooted by placing Y5447 as the most basal strain following the concatenated mitochondrial protein-gene tree in supplementary figure S3, Supplementary Material online. To avoid ambiguity and reliably determine substitution patterns on an internal branch, we applied strict monophyly (no homoplasy allowed) to infer derived characters and also excluded derived substitutions associated with any of the three basal strains. For instance, a substitution X → Y is inferred on the branch leading to the common ancestor of Y974, Y12793, and Y8871, only when Y is in Y974, Y12793, and Y8871, X is in all the remaining seven strains.

Mutation Bias and Expected Equilibrium G + C Content

The extent of mutation bias is estimated by comparing rates of different substitution types. Substitution rate of A/T → G/C was calculated as (number of A/T → G/C substitutions)/(total number of utilizable A/T sites), substitution rate of A/T → T/A was calculated as (number of A/T → T/A substitutions)/(total number of utilizable A/T sites), and substitution rate of G/C → A/T was calculated as (number of G/C → A/T substitutions)/(total number of utilizable G/C sites). The relative substitution rate of A/T → G/C (μ) was set to be 1 ($\mu = 1$), so that substitution rates of A/T → T/A (χ) and G/C → A/T (ν) can be easily compared relative with A/T → G/C (fig. 7). Mutation bias toward A/T (m) was calculated as, $m = \nu/\mu$, and the expected equilibrium G + C content was calculated as $1/(1 + m)$ following Lynch (2007). The 95% confidence interval was computed from sampling from the Poisson distribution with a mean of observed mutation count following Hershberg and Petrov (2010). In brief, 100,000 values were sampled from the Poisson distribution once with a mean of

observed 58 G/C → A/T mutations and once with a mean of observed 94 A/T → G/C mutations using the R program, rpois. They were used to recalculate 100,000 equilibrium %G + C values and the resulting values were sorted and used to estimate the 95% confidence interval for expected equilibrium %G + C.

Data Availability

The ten mitogenomes have been deposited to GenBank with accession numbers of KU888694–KU888699, MT461932, MT461933, MT534625, and MT534626.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank Prof. Ken Wolfe and two anonymous reviewers for very helpful comments on a previous version of this manuscript. The authors are grateful for the grid computing service from Computing & Information Technology of Wayne State University. This work was supported by funds from Wayne State University to W.H.

Author Contributions

W.H. designed the project; D.T.N. and S.X. performed research; D.T.N., B.W., S.X., and W.H. analyzed data; and W.H. wrote the article.

Literature Cited

- Agashe D, Shankar N. 2014. The evolution of bacterial DNA base composition. *J Exp Zool B Mol Dev Evol.* 322(7):517–528.
- Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19(5):455–477.
- Belozersky AN, Spirin AS. 1958. A correlation between the compositions of deoxyribonucleic and ribonucleic acids. *Nature* 182(4628):111–112.
- Bobay LM, Ochman H. 2017. Impact of recombination on the base composition of bacteria and archaea. *Mol Biol Evol.* 34(10):2627–2636.
- Bohlin J, Pettersson JH. 2019. Evolution of genomic base composition: from single cell microbes to multicellular animals. *Comput Struct Biotechnol J.* 17:362–370.
- Bouchier C, Ma L, Créno S, Dujon B, Fairhead C. 2009. Complete mitochondrial genome sequences of three *Nakaseomyces* species reveal invasion by palindromic GC clusters and considerable size expansion. *FEMS Yeast Res.* 9(8):1283–1292.
- Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665–2681.
- Chen J, Xing C, Zheng X, Li X. 2019. Complete genome sequence of *Cellulomonas* sp. strain Y8, a high-GC-content plasmid-free heavy metal-resistant bacterium isolated from farmland soil. *Microbiol Resour Announc.* 8(46):e01066.
- Chretien D, et al. 2018. Mitochondria are physiologically maintained at close to 50 °C. *PLoS Biol.* 16(1):e2003992.
- Ciani M, Maccarelli F. 1998. Oenological properties of non-*Saccharomyces* yeasts associated with wine-making. *World J Microbiol Biotechnol.* 14(2):199–203.
- Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14(7):1394–1403.
- DeFrancesco G, Turchetti B, Sileoni V, Marconi O, Perretti G. 2015. Screening of new strains of *Saccharomyces ludwigii* and *Zygosaccharomyces rouxii* to produce low-alcohol beer. *J Inst Brew.* 121(1):113–121.
- de Zamaroczy M, Bernardi G. 1986. The GC clusters of the mitochondrial genome of yeast and their evolutionary origin. *Gene* 41(1):1–22.
- Delbos F, Aoufouchi S, Faili A, Weill JC, Reynaud CA. 2007. DNA polymerase eta is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *J Exp Med.* 204(1):17–23.
- Dieringer D, Schlotterer C. 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.* 13(10):2242–2251.
- Dietel AK, Merker H, Kaltenpoth M, Kost C. 2019. Selective advantages favour high genomic AT-contents in intracellular elements. *PLoS Genet.* 15(4):e1007778.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5):1792–1797.
- Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc Biol Sci.* 252(1335):237–243.
- Foerstner KU, von Mering C, Hooper SD, Bork P. 2005. Environments shape the nucleotide composition of genomes. *EMBO Rep.* 6(12):1208–1213.
- Fricova D, et al. 2010. The mitochondrial genome of the pathogenic yeast *Candida subhashii*: GC-rich linear DNA with a protein covalently attached to the 5' termini. *Microbiology* 156(7):2153–2163.
- Gardner MJ, et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906):498–511.
- Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc Natl Acad Sci U S A.* 96(24):13880–13885.
- Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27(2):221–224.
- Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59(3):307–321.
- Haag-Liautard C, et al. 2008. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. *PLoS Biol.* 6(8):e204.
- Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biol Evol.* 3:344–358.
- Hellweger FL, Huang Y, Luo H. 2018. Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. *ISME J.* 12(5):1180–1187.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9):e1001115.
- Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. *J Mol Biol.* 236(4):1022–1033.
- Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.* 6(9):e1001107.
- Howe DK, Baer CF, Denver DR. 2010. High rate of large deletions in *Caenorhabditis briggsae* mitochondrial genome mutation processes. *Genome Biol Evol.* 2:29–38.
- Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111(1):147–164.

- Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23(2):254–267.
- James TY, et al. 2013. Shared signatures of parasitism and phylogenomics unite Cryptomycota and microsporidia. *Curr Biol.* 23(16):1548–1553.
- Johri P, Marinov GK, Doak TG, Lynch M. 2019. Population genetics of *Paramecium* mitochondrial genomes: recombination, mutation spectrum, and efficacy of selection. *Genome Biol Evol.* 11(5):1398–1416.
- Konrad A, et al. 2017. Mitochondrial mutation rate, spectrum and heteroplasmy in *Caenorhabditis elegans* spontaneous mutation accumulation lines of differing population size. *Mol Biol Evol.* 34(6):1319–1334.
- Lang BF, Gray MW, Burger G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. *Annu Rev Genet.* 33(1):351–397.
- Lassalle F, et al. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet.* 11(2):e1004941.
- Lewis DL, Farr CL, Farquhar AL, Kaguni LS. 1994. Sequence, organization, and evolution of the A+T region of *Drosophila melanogaster* mitochondrial DNA. *Mol Biol Evol.* 11:523–538.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.
- Long H, Behringer MG, Williams E, Te R, Lynch M. 2016. Similar mutation rates but highly diverse mutation spectra in Ascomycete and Basidiomycete yeasts. *Genome Biol Evol.* 8(12):3815–3821.
- Long H, et al. 2015. Background mutational features of the radiation-resistant bacterium *Deinococcus radiodurans*. *Mol Biol Evol.* 32(9):2383–2392.
- Long H, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol.* 2(2):237–240.
- Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
- Lynch M, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A.* 105(27):9272–9277.
- Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454(7203):479–485.
- McCutcheon JP, Moran NA. 2010. Functional convergence in reduced genomes of bacterial symbionts spanning 200 My of evolution. *Genome Biol Evol.* 2:708–718.
- Moran NA. 1996. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A.* 93(7):2873–2878.
- Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet.* 42(1):165–190.
- Pessia E, et al. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol.* 4(7):675–682.
- Raghavan R, Kelkar YD, Ochman H. 2012. A selective force favoring increased G + C content in bacterial genes. *Proc Natl Acad Sci U S A.* 109(36):14504–14507.
- Rampazzo C, et al. 2004. Mitochondrial deoxyribonucleotides, pool sizes, synthesis, and regulation. *J Biol Chem.* 279(17):17019–17026.
- Reichenberger ER, Rosen G, Hershberg U, Hershberg R. 2015. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biol Evol.* 7(5):1380–1389.
- Robinson JT, et al. 2011. Integrative genomics viewer. *Nat Biotechnol.* 29(1):24–26.
- Smith DR. 2012. Updating our view of organelle genome nucleotide landscape. *Front Genet.* 3:175.
- Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci U S A.* 112(33):10177–10184.
- Smith DR, et al. 2011. The GC-rich mitochondrial and plastid genomes of the green alga *Coccomyxa* give insight into the evolution of organelle DNA nucleotide landscape. *PLoS One* 6(8):e23624.
- Spencer J, Dunn-Walters DK. 2005. Hypermutation at A-T base pairs: the A nucleotide replacement spectrum is affected by adjacent nucleotides and there is no reverse complementarity of sequences flanking mutated A and T nucleotides. *J Immunol.* 175(8):5170–5177.
- Su HJ, et al. 2019. Novel genetic code and record-setting AT-richness in the highly reduced plastid genome of the holoparasitic plant *Balanophora*. *Proc Natl Acad Sci U S A.* 116(3):934–943.
- Sueoka N. 1961. Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old and new data. *J Mol Biol.* 3(1):31–40.
- Sun Y, et al. 2017. Spontaneous mutations of a model heterotrophic marine bacterium. *ISME J.* 11(7):1713–1718.
- Sung W, et al. 2015. Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. *Mol Biol Evol.* 32(7):1672–1683.
- Toth G, Gaspari Z, Jurka J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10(7):967–981.
- Wu B, Buljic A, Hao W. 2015. Extensive horizontal transfer and homologous recombination generate highly chimeric mitochondrial genomes in yeast. *Mol Biol Evol.* 32(10):2559–2570.
- Wu B, Hao W. 2014. Horizontal transfer and gene conversion as an important driving force in shaping the landscape of mitochondrial introns. *G3 (Bethesda)* 4:605–612.
- Wu B, Hao W. 2015. A dynamic mobile DNA family in the yeast mitochondrial genome. *G3 (Bethesda)* 5:1273–1282.
- Wu B, Hao W. 2019. Mitochondrial-encoded endonucleases drive recombination of protein-coding genes in yeast. *Environ Microbiol.* 21(11):4233–4240.
- Xiao S, Nguyen DT, Wu B, Hao W. 2017. Genetic drift and indel mutation in the evolution of yeast mitochondrial genome size. *Genome Biol Evol.* 9(11):3088–3099.
- Xu S, et al. 2012. High mutation rates in the mitochondrial genomes of *Daphnia pulex*. *Mol Biol Evol.* 29(2):763–769.
- Zivojnovic M, et al. 2014. Somatic hypermutation at A/T-rich oligonucleotide substrates shows different strand polarities in Ung-deficient or -proficient backgrounds. *Mol Cell Biol.* 34(12):2176–2187.

Associate editor: Kenneth Wolfe