Low base-substitution mutation rate in the germline genome of the ciliate *Tetrahymena thermophila*

Hongan Long[1,2,a], David J. Winter[3,a,*], Allan Y.-C. Chang[1], Way Sung[4], Steven H. Wu[3], Mariel Balboa[1], Ricardo B. R. Azevedo[1], Reed A. Cartwright[3,5], Michael Lynch[2], Rebecca A. Zufall[1]

1. Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5001, USA.

2. Department of Biology, Indiana University, Bloomington, Indiana 47405-7005, USA.

3. The Biodesign Institute, Arizona State University, Tempe, Arizona 85287-5301, USA.

4. Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina 28223-0001, USA.

5. School of Life Sciences, Arizona State University, Tempe, Arizona 85287-5301, USA.

a. These authors contributed equally to this work.

* To whom correspondence may be addressed. Email: djwinter@asu.edu

ABSTRACT

Mutation is the ultimate source of all genetic variation and is, therefore, central to evolutionary change. Previous work on *Paramecium tetraurelia* found an unusually low germline base-substitution mutation rate in this ciliate. Here, we tested the generality of this result among ciliates using *Tetrahymena thermophila*. We sequenced the genomes of 10 lines of *T. thermophila* that had each undergone approximately 1,000 generations of mutation accumulation (MA). We applied an existing mutation-calling pipeline and developed a new probabilistic mutation detection approach that directly models the design of an MA experiment and accommodates the noise introduced by mismapped reads. Our probabilistic mutation-calling method provides a straightforward way of estimating the number of sites at which a mutation could have been called if one was present, providing the denominator for our mutation rate calculations. From these methods, we find that *T. thermophila* has a germline base-substitution mutation rate of 7.61×10^{-12} per-site, per cell division, which is consistent with the low base-substitution mutation rate in *P. tetraurelia*. Over the course of the evolution experiment, genomic exclusion lines derived from the MA lines experienced a fitness decline that cannot be accounted for by germline base-substitution mutations alone, suggesting that other genetic or epigenetic factors must be involved. Because selection can only operate to reduce mutation rates based upon the “visible” mutational load, asexual reproduction with a transcriptionally silent germline may allow ciliates to evolve extremely low germline mutation rates.

Key words: mutation accumulation, micronucleus, macronucleus, drift-barrier hypothesis, microbial eukaryote, Oligohymenophorea.

INTRODUCTION

Mutation is the ultimate source of all genetic variation, and the rate, molecular spectrum, and phenotypic consequences of new mutations are all important drivers of biological processes such as adaptation, the evolution of sex, the maintenance of genetic variation, aging, and cancer. However, because mutations are rare, detecting them is difficult, often requiring the comparison of genotypes that have diverged from a common ancestor by at least hundreds or thousands of generations. Further, interpreting the results of such comparisons is complicated by the fact that mutations are frequently eliminated by natural selection before they can be studied.

Mutation accumulation (MA) is a standard method for studying mutations experimentally. In a typical MA experiment, many inbred or clonal lines are isolated and passed repeatedly through bottlenecks. This reduces the effective population size and lessens the efficiency of selection, allowing all but the most deleterious mutations to drift to fixation (Bateman 1959; Mukai 1964). The genome-wide mutation rate and mutational spectrum can then be estimated by comparing the genomes of MA lines with those of their ancestors. Such direct estimates of mutational parameters are now available for a number of model organisms (Denver et al. 2009; Keightley 2009; Keightley et al. 2014; Lee et al. 2012; Lind and Andersson 2008; Lynch et al. 2008; Ness et al. 2012; Ossowski et al. 2010; Sung et al. 2012b; Zhu et al. 2014). However, the narrow phylogenetic sampling of these species still limits our ability to understand how mutation rates and patterns have evolved and, in turn, have influenced evolution across the tree of life.

Microbial eukaryotes are an extraordinarily diverse group, containing many evolutionarily distant lineages, some of which have unusual life histories and genome features (Katz and Bhattacharya 2006). However, microbial eukaryotes are often unsuitable for use in mutational studies because they are difficult to culture in the lab, especially at the small population sizes required to reduce the efficiency of selection in MA experiments. In addition, genomic resources (e.g., completed annotated reference genomes) are lacking for most eukaryotic microbes. These barriers have limited MA experiments to well-annotated model microbial eukaryotes such as *Saccharomyces cerevisiae* (Lynch et al. 2008; Zhu et al. 2014), *Schizosaccharomyces pombe* (Behringer and Hall 2015; Farlow et al. 2015), *Paramecium tetraurelia* (Sung et al. 2012b), *Dictyostelium discoideum* (Saxer et al. 2012), and *Chlamydomonas reinhardtii* (Ness et al. 2012; Ness et al. 2015; Sung et al. 2012a; fig. 1).

The ciliated unicellular eukaryote *Tetrahymena thermophila* is well suited to MA experiments. Like all ciliates, individuals from this species have distinct germline and somatic copies of their nuclear genome. During asexual growth, the contents of the germline genome are duplicated mitotically but are neither expressed nor used to generate a new somatic genome. But unlike most other ciliates (including *P. tetraurelia*, which senesces in the absence of periodic mating or autogamy), *T. thermophila* can be propagated this way indefinitely. Thus, during periods of asexual growth — which can last over thousands of generations (Doerder 2014) — mutations can accumulate in the germline genome without being exposed to natural selection, which is operating on the somatic genome. Long et al. (2013) confirmed that MA lines of *T. thermophila* can be propagated asexually for at least 1,000 generations and inferred that they accumulate mutations in their germline genomes with detectable effects on fitness after the mutations are expressed in the somatic genome. However, Long et al. (2013) did not estimate the mutation rate directly at the molecular level.

The only other existing MA experiment from a ciliate was performed on *Paramecium tetraurelia* (Sung et al. 2012b) and yielded the lowest known base-substitution mutation rate in a eukaryote. Sung et al. (2012b) suggested that this exceptionally low mutation rate is the result of the unusual life history of ciliates, in which a transcriptionally silent germline genome undergoes multiple rounds of cell division between sexual cycles. Measurement of the mutation rate of *T. thermophila* will help reveal whether a low mutation rate is a general feature of ciliates. In addition, natural populations of *T. thermophila* have been the focus of population-genetic studies (Katz et al. 2006; Zufall et al. 2013), so mutational parameters estimated from MA experiments can be related to population and evolutionary processes.

Although the life history of *T. thermophila* is ideal for MA experiments, some features of its genome complicate typical computational approaches to detecting mutations from short-read sequencing. The genome is extremely AT-rich (~78% AT), contains many low complexity and repetitive elements, and has an incomplete reference genome (Eisen et al. 2006). These features make mapping sequencing reads to the reference genome difficult, which may lead to false positives when using naive mutation detection methods. To overcome these difficulties, we have developed a novel probabilistic mutation detection approach that directly models the design of an MA experiment and accommodates the noise introduced by mismapped reads. We used both our new method and an existing mutation-calling pipeline (Sung et al. 2012b) to analyse the MA sequences.

Here we expand the work presented by Long et al. (2013) by directly estimating the base-substitution mutation rate in *T. thermophila*. Our results are consistent with the exceptionally low rate estimated for *P. tetraurelia*, indicating that low germline mutation rates may be a general feature of ciliates. We also use our estimated rate to calculate the effective population size of *T. thermophila* in the wild. Our results establish that it is possible to estimate the mutation rate of *T. thermophila* directly from sequence data, but owing to the extraordinarily low rate, longer and larger MA experiments will be required to confidently estimate the mutational spectrum of a species with such a low mutation rate.

MATERIALS AND METHODS

Cell lines

The 10 evolved cell lines that were used in this study were generated from 10 parental MA lines (Supplementary Table S1). These lines were established from a single cell of the strain SB210 as described in Long et al. (2013). Briefly, the 10 MA lines were cultured in rich medium (SSP) in test tubes (Gorovsky et al. 1975) and experienced ~50 single-cell bottlenecks and ~1000 cell divisions, except for

M28, which was bottlenecked 10 times and passed ~200 cell divisions. The optical density of cultures was measured prior to each transfer and the number of generations calculated using a standard curve of optical density for the ancestor (Long et al. 2013). Because directly sequencing the *T. thermophila* micronuclear genome is not feasible, we generated autozygous lines with macronuclear genomes derived from haploid copies of our ancestral and descendant micronuclear genomes using genomic exclusion (Allen 1963). Genomic exclusion lines were produced by two rounds of crossing between the MA lines (mating type VI) and a germline-dysfunctional B* strain (mating type VII, Bruns and Cassidy-Hanley, 1999). A mutation in the macronuclear genome of a genomic exclusion line derived from an MA line is assumed to correspond to a germline mutation in that MA line.

We accounted for heterozygosity in the ancestral strain by generating 19 independent genomic exclusion lines from the progenitor line. The DNA from all 19 genomic exclusion lines was pooled before library construction, allowing us to sequence both alleles at any heterozygous sites.

Whole-genome sequencing

DNA libraries with insert size ~350 bp were constructed and Illumina paired-end sequenced by the DNASU core facility at the Biodesign Institute at Arizona State University and the Hubbard Center for Genome Studies, University of New Hampshire. The mean sequencing depth is ~47×, with >90% of the sites in the genome covered in all the sequenced lines (Supplementary Table S1). Sequencing reads are available from the NCBI’s SRA database under a BioProject with accession number PRJNA285268.

Base-substitution analysis

We used two independent approaches to call point-mutations to avoid false negatives that might not be detected by a single approach. First, a widely used consensus approach (Sung et al. 2012b). Second, a probabilistic approach that adapts methods designed for family-based data to the design of MA experiments (Cartwright et al. 2012). Our list of candidates was generated by the union of calls from both methods.

Consensus approach

For the consensus approach we applied the following filters to reduce false positives that may arise from sequencing, read mismapping or library amplification errors. (1) Two mapping programs, BWA 0.7.10 (Li and Durbin 2009) and novoalign (V2.08.01; NOVOCRAFT Inc), were used in two independent pipelines to reduce algorithm-specific read mapping errors. (2) Only uniquely mapped reads were used (BWA option: sampe –n 1; NOVOCRAFT option: novoalign –r None), with mapping/sequencing quality scores > 20 (samtools mpileup –Q 20 –q 20). (3) The line-specific consensus nucleotide at a genomic site needed support from greater than 80% of reads to filter out false positives from mismapping of paralog reads. (4) Three forward and three reverse reads were required to determine the line-specific consensus nucleotide, to reduce false positive calls due to errors in library construction or sequencing. Putative mutations were then called if a single line was different from the consensus of all the remaining lines following the approach of Sung et al. (2012b). This approach has been applied to a wide variety of prokaryotic and eukaryotic organisms and repeatedly verified with Sanger sequencing (Denver et al. 2009; Lee et al. 2012; Long et al. 2015; Ossowski et al. 2010; Sung et al. 2015). The consensus approach also makes predictions consistent with those of the GATK SNP caller (Behringer and Hall 2015; Farlow et al. 2015).

Probabilistic approach using accuMUlate

The challenge of identifying mutations from genomic alignments can also be treated as a hidden-data problem (Cartwright et al. 2012). Fig. 2 illustrates the application of a hidden-data approach to our MA experiment. For a given site in the reference genome, the only data we observe directly is the set of sequencing reads mapped to that site. In order to determine if a mutation has occurred at the site, we have to consider the processes by which the read data was generated. These processes include biological processes (e.g., inheritance, mutation, genomic exclusion) and experimental processes that can introduce errors (e.g., library preparation, sequencing, data processing). Because none of these states are directly observed, we consider them to be hidden data. Each unique combination of hidden states represents a distinct history that could have generated the read data for a given state. See fig. 3 for an example of one such history with hidden and observed data illustrated.

With the above formulation, our challenge is to determine the probability that a site contains at least one *de novo* mutation using our sequencing data (R) as the only observed input

$$P(\text{mutation}|R;\Theta) = \frac{P(\text{mutation},R;\Theta)}{P(R;\Theta)}. \tag{1}$$

Here, $P(\text{mutation},R;\Theta)$ is the joint marginal probability of at least one mutation being present and the sequencing data, and $P(R;\Theta)$ is the marginal probability of the sequencing data. The parameter Θ represents the model parameters and consists of the following:

- θ, the proportion of sites in the ancestor that are heterozygous, approximately (see Equation 6);
- φ_A, the overdispersion parameter for sequencing of the ancestor (described below);
- φ_D, the overdispersion parameter for sequencing of the descendant lines (described below);
- $\boldsymbol{\pi}$, a vector representing the frequency of each nucleotide in the ancestral genome;
- μ, the experiment-long mutation rate per site;
- ε, the rate of sequencing error per site.

The numerator and denominator in Equation 1 are marginal probabilities. To calculate them from the full data we have to sum the probability of mutation over the full set of histories (H). Each of these histories is a unique combination of hidden states that could have generated the read data (an example of one such history is shown in fig. 3),

$$\begin{aligned} P(\text{mutation}|R;\Theta) &= \frac{\sum_{h\in H} P(\text{mutation},R,h;\Theta)}{\sum_{h\in H} P(R,h;\Theta)} \\ &= \frac{\sum_{h\in H} P(\text{mutation}\mid h)\, P(R,h;\Theta)}{\sum_{h\in H} P(R,h;\Theta)}. \end{aligned} \tag{2}$$

Note that the probability of there being at least one mutation in a given history h, $P(\text{mutation}\mid h)$, is known to be either 1 or 0. Therefore, we only need to calculate $P(R,h;\Theta)$, the probability of the full data for the set of model parameters. This amounts to finding the probability that the read data was generated from an ancestral genotype G_A that gave rise to descendants with genotypes specified by the particular history being considered. This can be calculated as the products of the prior probability of genotypes and the likelihoods of those genotypes given the set of all sequencing data (R),

$$P(R,h;\Theta) = \underbrace{P(G_A;\theta,\boldsymbol{\pi})}_{a} \cdot \underbrace{P(\boldsymbol{R_A}|\,G_A;\ \varphi_A,\varepsilon)}_{b} \cdot \prod_{i}^{n} \left[\underbrace{P(G_i\mid G_A;\ \boldsymbol{\pi},\mu\,)}_{c} \cdot \underbrace{P(\boldsymbol{R_i}\mid G_i;\ \varphi_D,\varepsilon)}_{d}\right]. \tag{3}$$

The elements of R are vectors of size four, and each contains the number of A, C, G and T bases mapped to this site for a particular sample (the pileup data). The specific element $\boldsymbol{R_A}$ represents base counts from the ancestral strain. $\boldsymbol{R_i}$ and G_i are the base counts and genotype of the i-th descendant lines, respectively, and n is the total number of descendants.

The terms labeled “b” and “d” in Equation 3 are the probabilities of the observed sequencing data for a given genotype (i.e. genotype likelihoods). We calculate these genotype likelihoods using a Dirichlet-multinomial (DM) distribution. The DM is a compound distribution in which event-probabilities, $\boldsymbol{p}$, of a multinomial distribution is a Dirichlet-distributed random vector. Using a compound distribution provides flexibility to model the complex sources of error in sequencing data. To make this property of our model explicit, we use a parameterization of the DM distribution where $\boldsymbol{p}$ is a vector of length four containing the expected proportion of reads matching each allele and φ is an overdispersion parameter with values in the interval [0,1]. Using this parameterization, the DM distribution is equivalent to a simple multinomial distribution when φ = 0 and becomes increasingly overdispersed (i.e. the variance increases) as φ tends to 1.

We demonstrate the calculation of genotype likelihoods using the term for the ancestral genotype in Equation 3 (“b” term) as an example. To calculate $P(\boldsymbol{R_A} |\ G_A;\ \varphi_A, \varepsilon)$, we use the probability mass function of the DM distribution

$$P(\boldsymbol{R_A} = \boldsymbol{r} \mid G_A = g;\ \varphi_A, \varepsilon) = \binom{N}{\boldsymbol{r}} \frac{\Gamma(\omega_\mathrm{A})}{\Gamma(\omega_A + N)} \prod_b \frac{\Gamma(\boldsymbol{p}_b \omega_A + \boldsymbol{r}_b)}{\Gamma(\boldsymbol{p}_b \omega_A)}. \tag{4}$$

Here N is the total number of reads, $\binom{N}{\boldsymbol{r}}$ is the multinomial coefficient, Γ is the gamma function, and $\omega_a = (1 - \varphi_A)/\varphi_A$. The parameter vector $\boldsymbol{p}$ contains the expected frequency of bases in {A, C, G, T} for the site under consideration and is indexed by b. Values in $\boldsymbol{p}$ are determined by both the probability of sequencing error (ε) and the ancestral diploid genotype $g = \{g_1, g_2\}$ following Equation 5.

$$\boldsymbol{p_b} = \begin{cases} 1 - \varepsilon, & \text{if } b = \ g_1 = g_2 & \text{(homozygous match)} \\ \dfrac{1}{2} - \dfrac{\varepsilon}{3}, & \text{if } b = g_1 \neq g_2 \ \text{ or } b = g_2 \neq g_1 & \text{(heterozygous match)} \\ \dfrac{\varepsilon}{3}, & \text{otherwise} & \text{(error / mismatch)} \end{cases} \tag{5}$$

We now consider the remaining terms in Equation 3. The term labeled “a” represents the prior probability that the ancestor had a particular genotype (g_A below) at the site under consideration given the nucleotide composition of the *T. thermophila* genome and average heterozygosity of the ancestral strain. We calculate this via a finite-sites model with parent-independent mutation,

$$P(G_A = g_A; \theta, \boldsymbol{\pi}) = \begin{cases} \pi_{b_1} \dfrac{1}{1+\theta} + \pi_{b1}\pi_{b_2} \dfrac{\theta}{1+\theta} & \text{if } g_1 = g_2 = b_1 & \text{(homozygote)} \\ 2\,\pi_{b_1}\pi_{b_2} \dfrac{\theta}{1+\theta} & \text{if } g_1 = b_1 \text{ and } g_2 = b_2 & \text{(heterozygote)} \end{cases}. \tag{6}$$

Here $\frac{1}{1+\theta}$ is the probability that the ancestor is autozygous at a site, and $\boldsymbol{\pi}$ is the vector of stationary nucleotide/allele frequencies in ancestral genome and b_1 and b_2 refer to the indices of the g_1 and g_2 alleles. See Wright (1949) for more details on this model and its biological assumptions.

To complete Equation 3 we need to consider the term labeled “c”, which represents the probability that the *i*-th MA line inherited a particular genotype, given the ancestral genotype and the probability of mutation. We calculate this via the Felsenstein (1981) model of nucleotide substitution. This model incorporates equilibrium nucleotide frequencies, allowing us to include the extreme AT-bias present in the *T. thermophila* genome.

Using the approach described above, we used Equation 2 to calculate both the probability of at least one point-mutation and the probability of exactly one point-mutation at every site along the *T. thermophila* reference genome. In MA experiments, multiple mutations at the same site are unlikely; therefore, sites that contain a strong signal of more than one mutation are likely false positives due to systematic errors in sequencing and mapping of reads.

This model is implemented in a C++ program called accuMUlate, which makes use of the Bamtools (Barnett et al. 2011) library. The source code used to perform the calculations described above is available under an MIT license from https://github.com/dwinter/accumulate; the specific version of the code used in these analyses is archived at http://dx.doi.org/10.5281/zenodo.19942. We ran our model on a genomic alignment produced by using Bowtie version 2.1.0 (Langmead and Salzberg 2012) to map reads to the December 2011 release of the *T. thermophila* macronuclear genome from the *Tetrahymena* Genome Database (Stover et al. 2006). One site in the reference contained a gap character, which we removed since our reads indicated that it was an artifact. We processed the resulting alignments to remove sequencing and mapping artifacts that could lead to false-positive mutation calls. In particular, we identified and marked duplicate reads using the MarkDuplicates tool from Picard 1.106 (http://pricard.sourceforge.net) and performed local realignment around potential indels using GATK 3.2

(DePristo et al. 2011; McKenna et al. 2010). We adjusted raw base quality scores by running GATK's BaseRecalibrator tool, using a set of putative single nucleotide variants detected with SAMtools mpileup as input (Li et al. 2009).

The putative mutations from this approach were preliminarily identified by running accuMUlate to identify sites with a mutation probability > 0.1 with parameter-values: $\varphi_A = \varphi_D = 0.001, \varepsilon = 0.01$, $\mu = 10^{-8}$, $\theta = 0.0001$ and only considering reads with Phred-scaled mapping and base quality scores ≥ 13 (corresponding to an estimated 5% probability of error). The validation phase showed that false-positive mutations were frequently associated with poorly-mapped reads, low coverage regions surrounding deletions with respect to the reference genome, or the presence of rare bases in all samples. Thus, we re-ran the accuMUlate model, excluding all reads with a mapping quality < 25 (corresponding to an estimated 0.3% chance of error), and using the overdispersion parameters $\varphi_A = 0.03$ and $\varphi_D = 0.01$. Setting $\varphi_A > \varphi_D$ allowed us to accommodate the increased variance generated by sequencing pooled genomic exclusion lines to infer the ancestral genotype. In addition, we filtered out putative mutations that were not supported by at least 3 reads in both forward and reverse orientation. This final filtering step removed sites with unusually low coverage and those displaying strand bias, both characteristics associated with mismapped reads. We investigated the influence of our over-dispersion parameters by calculating the overall likelihood of the data using the initial and final parameter values. In order to make these results directly comparable, these calculations were performed on a data set consisting of all bases with quality score ≥ 13 from reads with a mapping quality ≥ 25.

Validation of putative mutations

The validity of a subset of putative mutations was tested by Sanger sequencing. All mutations identified by either the consensus or the probabilistic approach were tested with suitable primers up to 500 bp away from the mutation site. Primers were designed using the default parameters of Primer3 (Koressaar and

Remm 2007; Untergrasser et al. 2012) as implemented in Geneious (Kearse et al. 2012). Successful PCR products were purified and directly sequenced at Lone Star Labs (Houston, TX).

Mutation rate calculations

Our probabilistic approach to mutation detection also provides a way to calculate the number of sites at which we could have detected a mutation if one was present, and thus the correct denominator to use for mutation rate calculations. Using our final model parameters, we shuffled the vector of read-counts generated from a given sample in order to simulate mutations in our data. This procedure was repeated for every site in the reference genome, shuffling the read counts from each descendent separately then recalculating the probability of a mutation. A site was treated as missing from a sample if the mutation probability calculated from shuffled read-counts was < 0.1 or if the most probable mutant allele was not supported by at least 3 reads in both the forward and the reverse orientation. To investigate the impact of our final parameter values and filtering criteria on the number of callable sites we repeated this procedure using the initial parameter set (i.e. with $\varphi_A = \varphi_D = 0.001$ and removing reads with mapping quality < 13). The number of callable sites detected using this approach for each line is given in Table 1.

We calculated the mutation rate by summing the number of validated mutations (n_i) across MA lines, and then dividing it by the product of the number of analyzed sites (L) and the number of generations (T) in each MA line (i): $\hat{\mu} = \sum_i n_i/(L\,T)$. Assuming that the number of mutations in each line follows a Poisson distribution (but not necessarily the same distribution) and ignoring uncertainty in our estimates for L and T, the standard error for our estimate of mutation rate was estimated as $\mathrm{SE}(\hat{\mu}) = \sqrt{\hat{\mu}/(L\,T)}$, and a 95% confidence interval was constructed as $\hat{\mu} \pm 1.96\,SE(\hat{\mu})$. We also performed the same calculation on log-transformed values of $\hat{\mu}$, L, and, T to produce a “log-space” confidence interval. To calculate genomic mutation rates we assumed a haploid genome size of 104 Mb (Eisen et al. 2006).

Annotation of mutations

We annotated the functional context of identified mutations using snpEff (Cingolani et al. 2012) and the December 2011 release of the *T. thermophila* macronuclear genome annotation file from the *Tetrahymena* Genome Database.

RESULTS

Mutation detection and validation

To estimate the micronuclear mutation rate, we sequenced the whole macronuclear genomes of 10 homozygous genomic-exclusion lines, each derived from a separate *T. thermophila* line that had undergone MA for approximately 1000 generations. Using two different mutation-detection approaches (a widely used "consensus" method and a new probabilistic approach described in the Materials and Methods), we identified 92 sites for which there was some evidence of a mutation in at least one lineage. On closer inspection we found an unusual pattern — more than half of the apparent mutations were from lines M47 and M51, and in many cases reads containing the apparent mutant allele from one of these lines were also sequenced from the other line (but absent or very rare in all other lines).

To investigate this anomaly further we analyzed the frequency of non-reference bases in all samples across the whole genome (Supplementary Data). These analyses demonstrated that M47 and M51 differ from all other lines in the frequency of non-reference bases and in patterns of sequencing coverage. We do not know what caused the anomaly. It is possible that some cellular process occurred in these lines but not others (e.g., the incorporation of sequences usually restricted to the micronucleus, or the inclusion of DNA from the B* strain during genomic exclusion). It is extremely unlikely that M47 and M51 independently accrued more shared mutations than independent mutations during our MA experiment. For this reason, we have excluded these lines from all subsequent analyses.

Forty putative mutations remained after lines M47 and M51 were removed (Supplementary Table S2). We attempted to validate each of these mutations using Sanger sequencing. Only 4 of these mutations were validated. The remaining sites were either shown to be false positives (11 sites) or failed to generate either PCR amplicons or clean sequence traces (25 sites). Closer inspection of the data underpinning both the false positive and inconclusive mutations showed these sites to have unusually low sequencing coverage and low mapping quality, and to be subject to strand bias. All of these properties are associated with mapping error, and are known to generate false positive variant calls (Li 2014). For this reason, we re-ran our probabilistic mutation caller using stricter filters for mapping quality and excluding putative mutations that did not have at least 3 sequencing reads supporting a mutation in both the forward and reverse orientation. None of the inconclusive or false positive sites were called as mutations in this analysis, which also detected an additional mutation that was confirmed by Sanger sequencing. Thus, we detected a total of 5 mutations across 8 MA lines, with no line having more than one confirmed mutation (Table 1). Our probabilistic method produced more false positives than the consensus approach but generated no false negatives (Supplementary Table S2). Of the 5 mutations detected, 2 are non-synonymous, 2 are synonymous, and one is in an intergenic region.

Number of callable sites

We estimated the denominator for our mutation rate estimates by calculating the number of sites at which a mutation could be called if one was present. An average of 86.1% of the reference genome was callable per line (Table 1). Sites for which we lacked power to detect mutations in at least one line are in relatively gene-poor regions; 30% of such sites are in exons compared to 49% of always-included sites. We also considered the impact of our final filtering steps and model parameters on our analyses. The more stringent filtering steps we used to generate our final mutation set reduced the proportion of callable sites per line, with the median proportion of callable sites declining from 93% to 88% (Table 1). The final overdispersion values used in our probabilistic mutation caller produced a better fit to our data than the initial values, with the overall log likelihood improving by 8×10^4.

Mutation rate

Given the number of callable sites, the 5 mutations that we detected yield a base-substitution mutation rate estimate of 7.61×10^{-12} per base pair per asexual generation (95% confidence interval, CI = [0.691×10^{-12}, 14.53×10^{-12}] using the standard method or [4.68×10^{-12}, 12.38×10^{-12}] using log-transformed values). This point estimate is approximately one third of the rate reported for *P. tetraurelia*, although the 95% CIs of both estimates overlap (fig. 1), and equates to a genome-wide rate of 0.8 base-substitution mutations per haploid genome per thousand asexual generations (95% CI = [0.07, 1.50]).

If our estimate of the base-substitution mutation rate holds for the portions of the genome from which we did not have sufficient power to detect mutations, then we estimate that we have failed to detect an additional 0.87 mutations across all of the macronuclear genomes sequenced.

DISCUSSION

We have used whole-genome sequencing and a novel mutation-detection approach to estimate the base-substitution mutation rate of *T. thermophila* from 8 MA lines (Long et al. 2013) and obtained an estimate of 7.61×10^{-12} mutations per-site per-generation. This is the lowest estimate of base-substitution mutation rate from a eukaryote (for surveys of mutation-rate estimates see fig. 1 and Sung et al 2012b), and indeed lower than that observed in any prokaryote. However, it is not significantly different from the rate in either the social amoeba *Dictyostelium discoideum* (Saxer et al. 2012) or the ciliate *P. tetraurelia* (Sung et al. 2012b). The fact that the two lowest mutation rates have been recorded in ciliates supports the hypothesis that ciliates in general have low germline mutation rates (Sung et al. 2012b).

Direct estimates of the mutation rate from MA lines can only be as accurate as the methods used to detect mutations. Our estimate of a low mutation rate in *T. thermophila* could conceivably result from a high rate of false-negative results. However, we believe that this is unlikely. Our approach to mutation detection was designed to maximize the sensitivity of our analyses. We initially applied lenient filters to our data and attempted to validate all putative mutations detected by two separate methods. Most of the putative mutations suggested by this initial analysis could not be validated by Sanger sequencing. For this reason, we developed filters and model-parameters that improved the specificity of our mutation-calling method (producing negligible mutation probabilities for all of our unconfirmed mutations, while still supporting our confirmed mutations). It is possible that this increased stringency also led us to miss mutations present in our descendant lines. To account for the possibility of such false negatives in our mutation rate estimates, we simulated mutations in our data. This allowed us to identify sites at which we would not be able to detect a mutation in a given line even if one was present. Sites for which we could not call a simulated mutation were not included in the denominator of our mutation rate calculation. Thus, we are confident that our mutation rate estimate is accurate, at least for the regions of the genome from which we could call mutations.

Our mutation rate estimate allows us to estimate the effective population size of *T. thermophila*. If we assume that silent sites in protein-coding genes are effectively neutral and under drift-mutation equilibrium, the population-level heterozygosity at silent sites (π_s) has expected value $4N_e\mu$, where N_e is the effective population size, and μ is mutation rate per site per generation. Using the estimate $N_e \times \mu = 8 \times 10^{-4}$ reported by Katz et al. (2006), if we assume that mutation rates in the germline and somatic genomes are equal, our N_e estimate for *T. thermophila* is 1.12×10^8, which is almost identical to that of *P. tetraurelia* (N_e=1.24×10^8; Sung et al. 2012b). These estimates may seem surprising given the observations that *P. tetraurelia* is cosmopolitan and regularly isolated from different continents (Catania et al. 2009), while *T. thermophila* has a distribution limited to the eastern United States (Zufall et al. 2013). However, the relationship between census population size and genetic diversity (and therefore estimated N_e) is not a simple one (Leffler et al. 2012; Lewontin 1974). In very large populations stochastic processes, including demographic events that prevent populations from reaching mutation-drift equlibrium (Haigh and Maynard Smith 1972; Leffler et al. 2012) and the effects of selection on sites linked to neutral variants (Gillespie 2001; Lynch 2007; Neher et al. 2013), limit genetic diversity across the whole genome. Regardless, the large effective population size estimated here suggests that selection will have considerable power in the evolution of *T. thermophila*.

The unusual genome structure and life history of ciliates may explain their low mutation rates. Sung et al. (2012a) argued that mutation rates are minimized to the extent made possible by the power of natural selection — the "drift barrier" hypothesis. Selection operates to reduce the mutation rate based on the "visible" mutational load, and mutations that accumulate in the germline genome in ciliates during asexual generations are not expressed and exposed to selection until they are incorporated in a new somatic genome following sexual reproduction. Thus, the mutation rate per selective event is equal to the mutation rate per asexual generation multiplied by the number of asexual generations between rounds of sexual reproduction. The low mutation rates reported for ciliates may have evolved naturally as a consequence of the many asexual generations in between bouts of sexual reproduction, combined with large effective population sizes that promote strong selection for low mutation rates.

Unlike *P. tetraurelia*, *T. thermophila* does not undergo senescence in the absence of sex, and we lack a good estimate for the frequency of sexual reproduction in natural populations (Doerder et al. 1995). Therefore, we cannot put an upper bound on the number of asexual generations between conjugation events. However, we can estimate a lower bound because cells arising from sexual reproduction enter a period of immaturity lasting ~50–100 divisions (Lynn and Doerder 2012). We know that the germline genome divides at least this many times without opportunity for selection on any newly acquired mutations. Using the immaturity period as a proxy for the frequency of sex gives an estimate of the base-substitution mutation rate of ~0.1 mutations per haploid genome per conjugation event — much closer to that of other eukaryotes per round of DNA replication (Sung et al. 2012b).

Most mutations with effects on fitness are deleterious, so the accumulation of mutations in the absence of selection is expected to lead to a reduction in organismal fitness (Bateman 1959; Halligan and Keightley 2009; Mukai 1964; Muller 1928). The fitness of a genomic exclusion line derived from an MA line of *T. thermophila* should, in part, reflect the germline mutations in that MA line. If most germline mutations are base-substitutions, the low germline base-substitution rate would lead us to predict modest effects on the fitness of the genomic exclusion lines we studied. Surprisingly, some of these lines experienced substantial fitness losses relative to the ancestor (Long et al. 2013). For example, we did not detect any base-substitution mutations in the line with largest observed loss in fitness (M50, w=0.38) (Table 1). It is unlikely that the fitness losses observed in these MA lines can be explained by other undetected single-base substitutions, as our mutation calling method had power to detect mutations in an average of 86.1% of the genome (Table 1) and the excluded portion of the genome is relatively gene poor. Rather, it seems likely the fitness of these lines is determined in part by indels and other structural variants that we did not include in this study. Furthermore, non-Mendelian patterns of inheritance could obscure the relationship between mutations and fitness. For example, the fitness of an individual line may be influenced by epigenetic processes, such as cortical inheritance (Sonneborn 1963) or small RNA guided genome rearrangement (Mochizuki and Gorovsky 2004).

Our probabilistic mutation calling method can directly model the overdispersion produced by modern sequencing techniques. Our initial run of this method used very low values for the overdispersion parameters and produced many false positive mutation calls. We believe that these false positives arose from sites with data that did not fit the expectation of relatively clean data that these low overdispersion values represent. Using data form the validation phase, we were able to show that increasing the overdispersion parameter values (and being more stringent about which sequencing reads were included in our analysis) improved the accuracy of our method such that it produced fewer false positives while still correctly identifying all validated mutations. In addition to improving the accuracy of our mutation calling, increasing the values of the overdispersion parameters substantially increased the fit of our model to the sequencing data. Although the small number of true mutations in this experiment prevents us from performing a more complete analysis, this study demonstrates that modeling overdispersion in sequencing data can improve mutation calling methods.

In this study we have established that it is possible to detect mutations in *T. thermophila* MA lines through short-read sequencing, and thus to directly study the nature of mutation in this model organism. Although we were able to show that *T. thermophila* shares a low mutation rate with *P. tetraurelia* (the only other ciliate for which a mutation rate has been directly estimated), there is still much to learn about mutation in this species. For instance, the unusual genome structure of ciliates presents a novel test of the drift-barrier hypothesis of mutation rate evolution (Sung et al. 2012a). If the mutation rates of the germline and somatic nuclei can evolve independently then we would expect the somatic mutation rate to be higher (i.e. more similar to the mutation rates of other eukaryotes) because somatic mutations are exposed to selection after each cell division. Furthermore, the small number of mutations accumulated over this experiment has prevented us from analyzing the spectrum of mutations arising in *T. thermophila* and determining the influence of mutational biases on genome evolution. Similarly, the few mutations that we detect seem inadequate to explain the observed losses of fitness during MA. Future studies using more MA lines evolving over longer periods and detecting indels and other structural variants accrued during MA will be needed to fully understand the effects of mutation and selection in *T. thermophila*.

ACKNOWLEDGEMENTS

We thank Kristen Dimond, Robert Coyne, Tom Doak, Kale Dai, Adam Orr, and Rachel Schwartz for technical help and two anonymous reviewers for helpful comments. This study is funded by NIH R01GM101352 (RAZ, RBRA, RAC) and Multidisciplinary University Research Initiative award W911NF-09-1-0444 (ML) from the US Army Research Office, NIH grant R01GM036827 (ML) and National Science Foundation Grant MCB-1050161 (ML).

REFERENCES

Allen SL 1963. Genomic exclusion in *Tetrahymena*: genetic basis. J Protozool. 10: 413–420.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 27: 1691–1692.

Bateman AJ 1959. The viability of near-normal irradiated chromosomes. Int J Radiat Biol 1: 170–180.

Behringer MG, Hall DW 2015. Genome-wide estimates of mutation rates and spectrum in *Schizosaccharomyces pombe* indicate CpG sites are highly mutagenic despite the absence of DNA methylation. G3 6: 149–160.

Bruns P, Cassidy-Hanley D 1999. Methods for genetic analysis. Methods Cell Biol 62: 229–240.

Cartwright RA, Hussin J, Keebler JEM, Stone EA, Awadalla P 2012. A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. Stat Appl Genet Molec 11: Article 6.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6: 80–92.

Denver DR, et al. 2009. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc Natl Acad Sci USA 106: 16310–16314.

DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.

Doerder FP 2014. Abandoning sex: multiple origins of asexuality in the ciliate *Tetrahymena*. BMC Evol Biol 14: 112.

Doerder FP, Gates MA, Eberhardt FP, Arslanyolu M 1995. High frequency of sex and equal frequencies of mating types in natural populations of the ciliate *Tetrahymena thermophila*. Proc Natl Acad Sci USA 92: 8715–8718.

Eisen JA, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol 4: e286.

Farlow A, et al. 2015. The spontaneous mutation rate in the fission yeast *Schizosaccharomyces pombe*. Genetics 201: 737–744.

Felsenstein J 1981. Evolutionary trees from DNA-sequences - a maximum-likelihood approach. J Mol Evol 17: 368–376.

Gillespie JH 2001. Is the population size of a species relevant to its evolution? Evolution 55: 216–219.

Gorovsky M, Yao M, Keevert J, Pleger G 1975. Isolation of micro- and macronuclei of *Tetrahymena pyriformis*. Methods Cell Biol 9: 311–327.

Haigh J, Maynard Smith J 1972. Population size and protein variation in man. Genet Res 19: 73–89.

Halligan DL, Keightley PD 2009. Spontaneous mutation accumulation studies in evolutionary genetics. Annu Rev Ecol Evol Syst 40: 151–172.

Katz LA, Bhattacharya D. 2006. Genome evolution in eukaryotic microbes. Oxford, UK: Oxford University Press.

Katz LA, Snoeyenbos-West O, Doerder FP 2006. Patterns of protein evolution in *Tetrahymena thermophila*: implications for estimates of effective population size. Mol Biol Evol 23: 608–614.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649.

Keightley P, Trivedi U, Thomson M, Oliver F, Kumar S et al. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res 19: 1195–1201.

Keightley PD, Ness RW, Halligan DL, Haddrill PR 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics 196: 313–320.

Koressaar T, Remm M 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23: 1289–1291.

Langmead B, Salzberg SL 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

Lee H, Popodi EM, Tang H, Foster PL 2012. 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.

Leffler EM, et al. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol 10: e1001388.

Lewontin RC. 1974. The genetic basis of evolutionary change. New York, USA: Columbia University Press.

Li H 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30: 2843–2851.

Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754–1760.

Li H, et al. 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078–2079.

Lind PA, Andersson DI 2008. Whole-genome mutational biases in bacteria. Proc Natl Acad Sci USA 105: 17878–17883.

Long H, Paixao T, Azevedo RBR, Zufall RA 2013. Accumulation of spontaneous mutations in the ciliate *Tetrahymena thermophila*. Genetics 195: 527–540.

Long H, et al. 2015. Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair (MMR) deficient *Pseudomonas fluorescens* ATCC948. Genome Biol Evol 7: 262–271.

Lynch M. 2007. The origins of genome architecture. Sunderland, MA: Sinauer Associates, Inc.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci USA 105: 9272–9277.

Lynn DH, Doerder FP 2012. The life and times of *Tetrahymena*. Methods Cell Biol 109: 9–27.

McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Mochizuki K, Gorovsky MA 2004. Small RNAs in genome rearrangement in *Tetrahymena*. Curr Opin Genet Dev 14: 181–187.

Mukai T 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics 50: 1–19.

Muller HJ 1928. The measurement of gene mutation rate in *Drosophila*, its high variability, and its dependence upon temperature. Genetics 13: 279–357.

Neher RA, Kessinger TA, Shraiman BI 2013. Coalescence and genetic diversity in sexual populations under selection. Proc Natl Acad Sci U S A 110: 15836–15841.

Ness RW, Morgan AD, Colegrave N, Keightley PD 2012. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. Genetics 192: 1447–1454.

Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD 2015. Extensive *de novo* mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. Genome Res 25: 1739–1749.

Ossowski S, et al. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327: 92–94.

Saxer G, et al. 2012. Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba *Dictyostelium discoideum*. PLoS One 7: e46759.

Sonneborn TM. 1963. Does preformed cell structure play an essential role in cell heredity? In: Allen JM, editor. The nature of biological diversity. New York: McGraw-Hill.

Stover NA, et al. 2006. *Tetrahymena* Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. Nucleic Acids Res 34: D500–D503.

Sung W, et al. 2015. Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. Mol Biol Evol 32: 1672–1683.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M 2012a. Drift-barrier hypothesis and mutation rate evolution. Proc Natl Acad Sci USA 109: 18488–18492.

Sung W, et al. 2012b. Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. Proc Natl Acad Sci USA 109: 19339–19344.

Untergrasser A, et al. 2012. Primer3 - new capabilities and interfaces. Nucleic Acids Res 40: e115.

Wright S 1949. The genetical structure of populations. Ann Hum Genet 15: 323–354.

Zhu YO, Siegal ML, Hall DW, Petrov DA 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA 111: E2310–E2318.

Zufall RA, Dimond KL, Doerder FP 2013. Restricted distribution and limited gene flow in the model ciliate *Tetrahymena thermophila*. Mol Ecol 22: 1081–1091

FIGURE LEGENDS

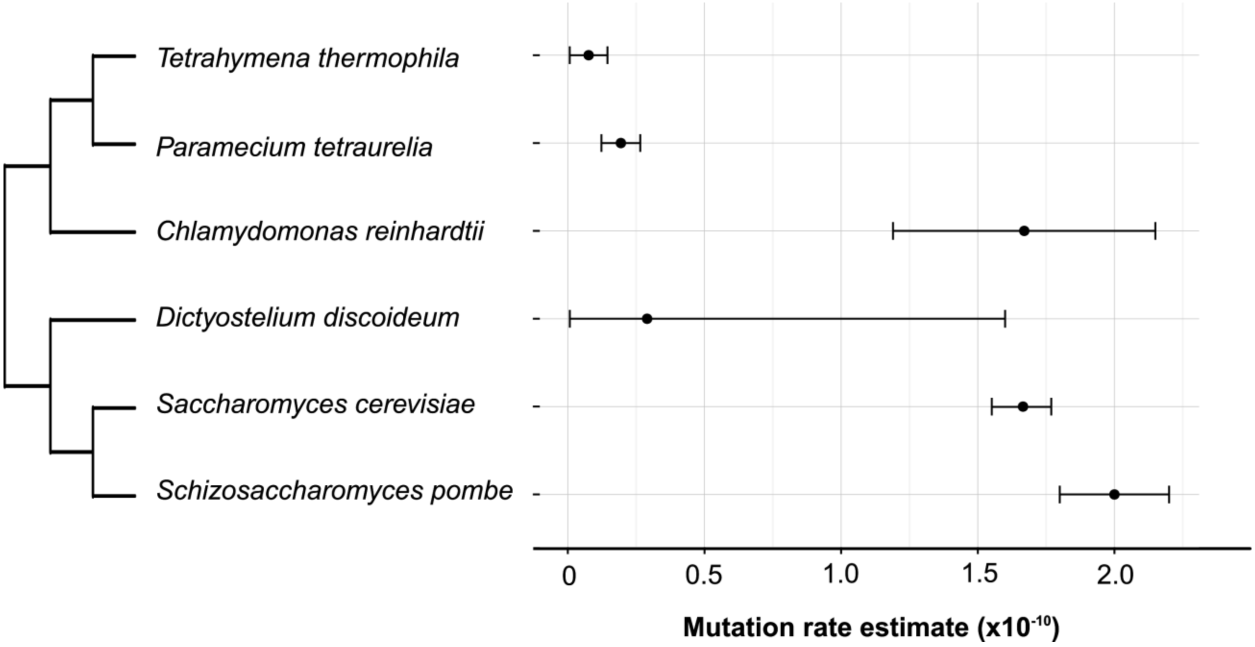


FIG. 1.—Mutation rate estimates for unicellular eukaryotes. Base-substitution mutation rates per nucleotide per generation estimated for different unicellular eukaryotes: *T. thermophila* (this paper), *P. tetraurelia* (Sung et al. 2012b), *C. reinhardtii* (Ness et al. 2015), *D. discoideum* (Saxer et al. 2012), *Sa. cerevisiae* (Zhu et al. 2014), and *Sc. pombe* (Farlow et al. 2015). The phylogenetic tree was retrieved from the Open Tree of Life (Hinchliff et al. 2015); branch lengths are arbitrary. Error bars are 95% confidence intervals calculated by ignoring uncertainty in the number of sites at which a mutation could be detected and the number of generations that each MA ran experiment ran for.

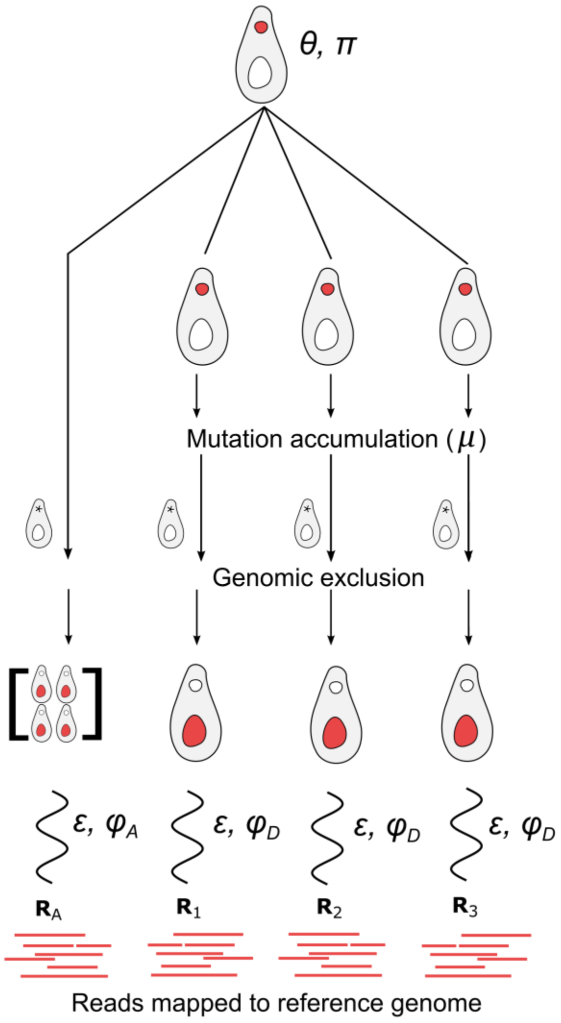


FIG. 2.—Experimental design in relation to parameters of probabilistic mutation-detection model. A complete description of the experiment is presented in Long et al. (2013). Here, we describe how the experiment relates to the parameters used in our probabilistic mutation-calling model. Specifically, the ancestral line with average heterozygosity θ and genome-wide nucleotide frequencies $\boldsymbol{\pi}$ is used to generate a set of MA lines. Each of these lines accumulates mutations at a rate μ per nucleotide-site per generation for 1000 generations. Genomic exclusion, an auto-diploidization process, is used to generate lines with macronuclei representing one haploid-copy of each MA line (and multiple copies of the ancestral line, in order to detect ancestral heterozygosity). The macronuclear genomes of these genomic exclusion lines are then sequenced with a sequencing error rate of ε and overdispersion caused by library

preparation and other correlated errors modeled as φ_A and φ_D for ancestral and descendant lines respectively. A full description of this model and its parameters is given in the subsection of the Materials and Methods labeled “Probabilistic approach using accuMUlate”

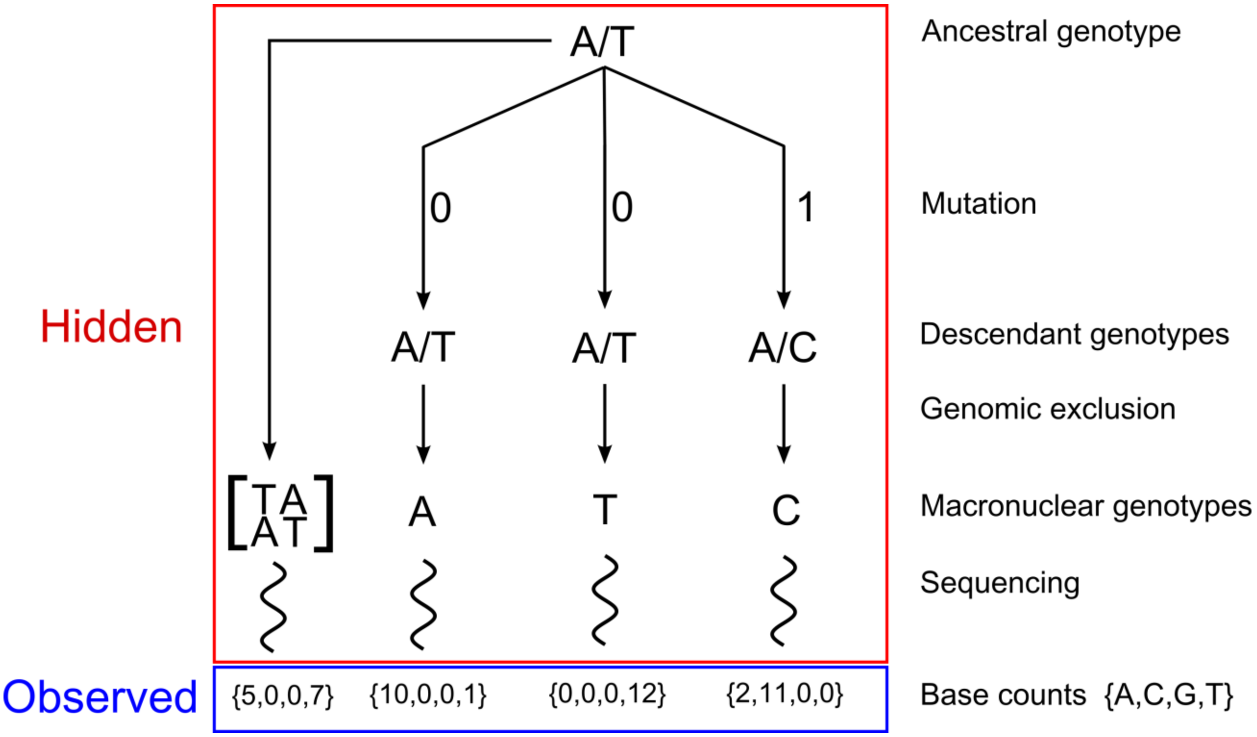

A/T
Ancestral genotype
0
0
1
Mutation
Hidden
A/T
A/T
A/C
Descendant genotypes
Genomic exclusion
TA
AT
A
T
C
Macronuclear genotypes
Sequencing
Observed
{5,0,0,7}
{10,0,0,1}
{0,0,0,12}
{2,11,0,0}
Base counts {A,C,G,T}

FIG. 3.—Illustration of a single history in the accuMUlate method. In our model, a history is a unique combination of states (i.e. the genotypes of ancestral and MA lines, results of genomic exclusions and errors introduced during sequencing) generated during an MA experiment. Here we illustrate one such history by giving values to the different states in a model reflecting the same experimental design as fig. 2 and show how we calculate the probability that this history occurred and generated the observed sequencing data. Because we treat sites in the reference genome independently, we describe the process for a single site. Specifically, we consider a history in which an ancestor that is heterozygous with genotype A/T is used to establish three MA lines. One of those lines experiences a mutation from A/T to A/C, and the C allele of this mutant is passed on to a new macronuclear genome through genomic exclusion. The only data we observe for this locus is the set of bases mapped to this site that pass our filtering steps. We represent this data as vectors containing the number of A, C, G and T bases mapped to a given site (the base counts). We can use Equation 3 to calculate the probability that this sequencing data was generated by the specific history shown here. To do this, we first calculate the probability that the

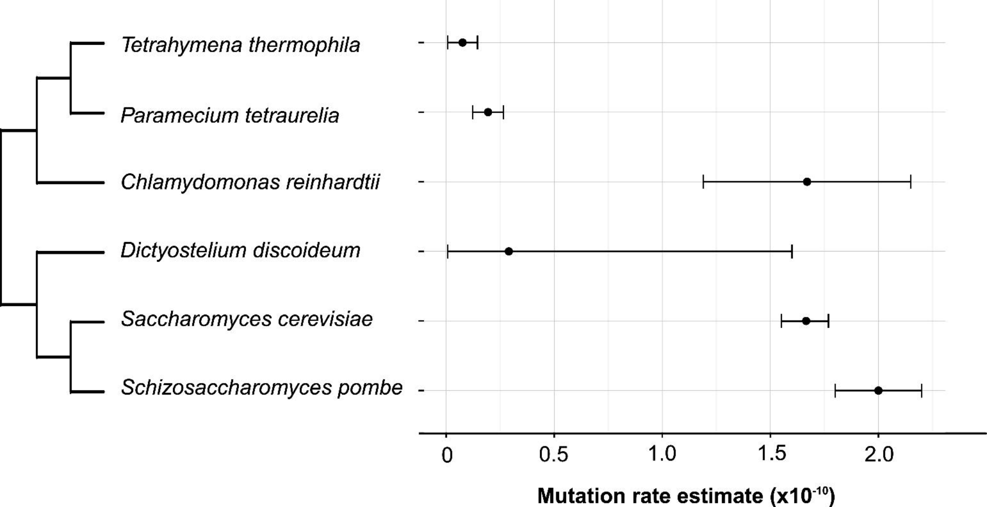
ancestor would have genotype A/T and that the observed sequencing data from the ancestor could be generated from this genotype (using Equations 6 and 4, respectively). Next, we consider the MA (descendant) lines, calculating the probability that the three descendant lines would have genotypes A, T and C and that the observed sequencing data could be generated from these genotypes. In this case we use the Felsenstein (1981) model of nucleotide substitution to calculate the probabilities that genomic exclusions generated from the MA lines would have these genotypes. We use the same genotype likelihood model (Equation 4) to calculate the probability that the sequencing data was generated from MA lines with these genotypes. Because each of the descendant lines is independent of each other, the overall probability of the history is simply the product of the probabilities for the ancestral and all descendant lines (Equation 3). We calculate the probability of a site containing at least one mutation by repeating this procedure for all possible histories at a given site (i.e. all possible combinations of genotypes) and keeping track of those histories that contain one or more mutations (Equation 2)

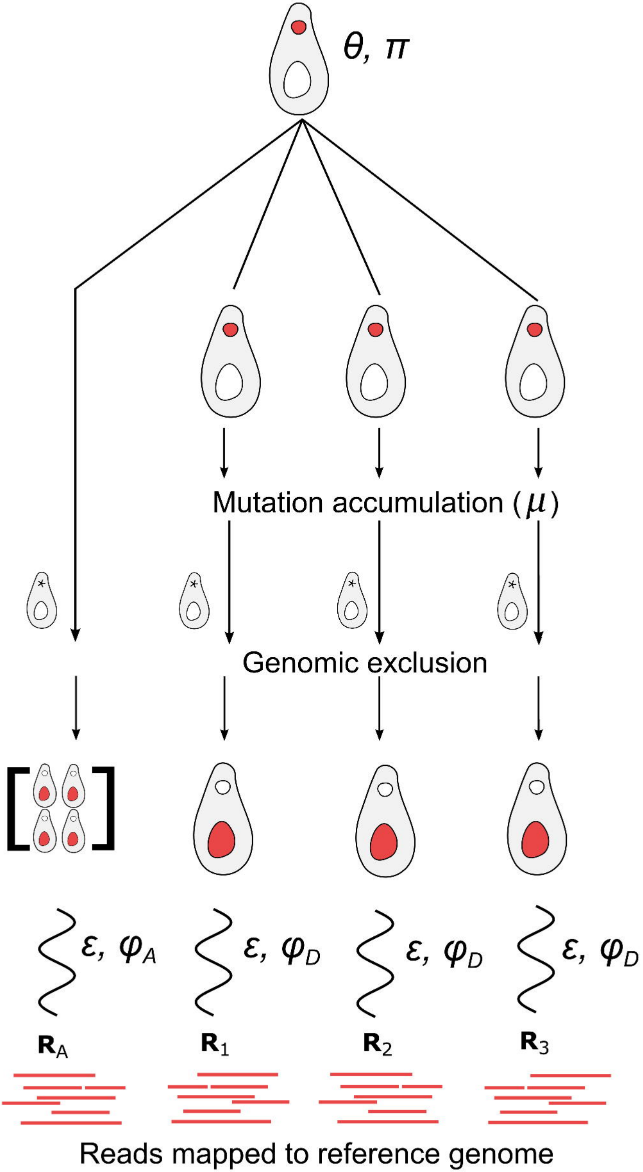
				Callable^b		Mutation				
Line	Coverage	Generations	Fitnes^s	Initial	Final	Scaffold	Substitution	Feature	Locus	Effect
M5	64.17	1000	0.56	0.92	0.88	scf_8254658	g.334881C>T	Exon	TTHERM_00675900A	Synonymous (gaC>gaT, D>D)
M19	53.05	1000	0.64	0.93	0.88	-	-	-	-	-
M20	34.42	1000	0.44	0.93	0.88	scf_8254594	g.239327A>T	Intron	TTHERM_00286840A	-
M25	30.88	1000	0.57	0.93	0.88	scf_8254607	g.179325C>T	Exon	TTHERM_00439220A	Non Synonymous (Gtt>Att, V>I)
M28	50.65	200	0.65	0.92	0.87	-	-	-	-	-
M29	31.36	1000	0.49	0.93	0.88	scf_8254365	g.304140T>G	Intergenic	-	-
M40	16.84	1000	0.57	0.83	0.63	scf_8254002	g.27830G>A	Exon	TTHERM_01128590A	Non Synonymous (tGT>tAt, C>Y)
M50	106.65	1000	0.38	0.93	0.88	-	-	-	-	-

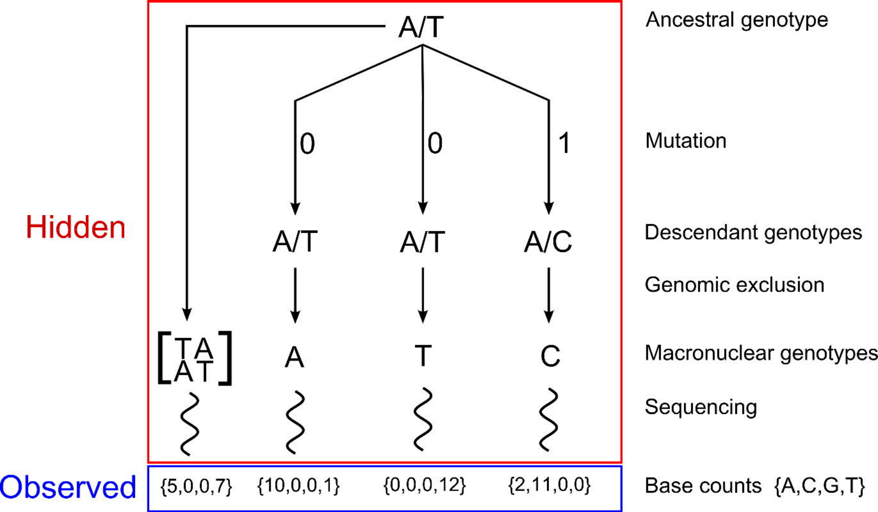
a. Fitness data from Long et al. (2013), using exponential growth rate as fitness metric and normalized by dividing the ancestral growth rate.

b. The proportion of all sites in the MAC genome (1.04Mb) from which a mutation could have been called if one was present. "Initial" refers to the first analysis (with reads with mapping quality < 13 removed and parameter values $\varphi_A = 0.001, \varphi_D = 0.001$), "final" refers to the subsequent analysis (with reads with mapping quality < 30 removed and putative mutations supported by < 3 read in forward and reverse orientation removed, and with parameter values $\varphi_A = 0.03, \varphi_D = 0.01$).

Table 1: Summary of sequencing data and detected mutations. Note, no mutations were detected from lines M50, M28 or M19.


Tetrahymena thermophila
Paramecium tetraurelia
Chlamydomonas reinhardtii
Dictyostelium discoideum
Saccharomyces cerevisiae
Schizosaccharomyces pombe
0
0.5
1.0
1.5
2.0
Mutation rate estimate (x10^{-10})


θ, π
Mutation accumulation (μ)
Genomic exclusion
ε, φ_A
ε, φ_D
ε, φ_D
ε, φ_D
R_A
R_1
R_2
R_3
Reads mapped to reference genome


A/T
Ancestral genotype
0
0
1
Mutation
Hidden
A/T
A/T
A/C
Descendant genotypes
Genomic exclusion
TA
AT
A
T
C
Macronuclear genotypes
Sequencing
Observed
{5,0,0,7}
{10,0,0,1}
{0,0,0,12}
{2,11,0,0}
Base counts {A,C,G,T}