

Genome analysis

# accuMulate: a mutation caller designed for mutation accumulation experiments

David J. Winter<sup>1,2,\*</sup>, Steven H. Wu<sup>1,3</sup>, Abigail A. Howell<sup>1,4</sup>,  
Ricardo B. R. Azevedo<sup>5</sup>, Rebecca A. Zufall<sup>5</sup> and Reed A. Cartwright<sup>1,4</sup>

<sup>1</sup>The Biodesign Institute, Arizona State University, Tempe, AZ 85287-5301, USA, <sup>2</sup>Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand, <sup>3</sup>Bioconsortia Inc, Davis, CA 95618, USA, <sup>4</sup>School of Life Sciences, Arizona State University, Tempe, AZ 85287-5301, USA and <sup>5</sup>Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA

\*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received and revised on November 10, 2017; editorial decision on March 12, 2018; accepted on March 15, 2018

## Abstract

**Summary:** Mutation accumulation (MA) is the most widely used method for directly studying the effects of mutation. By sequencing whole genomes from MA lines, researchers can directly study the rate and molecular spectra of spontaneous mutations and use these results to understand how mutation contributes to biological processes. At present there is no software designed specifically for identifying mutations from MA lines. Here we describe accuMulate, a probabilistic mutation caller that reflects the design of a typical MA experiment while being flexible enough to accommodate properties unique to any particular experiment.

**Availability and implementation** accuMulate is available from <https://github.com/dwinter/accuMulate>.

**Contact:** david.winter@gmail.com

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Mutation accumulation (MA) is the classic method of directly studying the rates, molecular spectra, and fitness consequences of spontaneous mutations. In a typical MA experiment, replicate inbred or clonal lines are isolated and repeatedly passed through severe bottlenecks. These bottlenecks reduce the effective population size of lines and thus reduce the efficiency of selection, allowing all but the most deleterious mutations to drift to fixation. The development of high-throughput sequencing technologies has led to a renewed interest in MA experiments. By sequencing whole genomes from MA lines, researchers can directly estimate the rate and molecular spectrum of mutations in a given species or strain. Studies combining MA experiments with whole-genome sequencing have provided key insights into the evolution of mutation rates (Lynch *et al.*, 2016), genome evolution (Tenailon *et al.*, 2016), the molecular basis of mutation (Zhu *et al.*, 2014), and the distribution of fitness effects among spontaneous mutations (Dillon and Cooper, 2016).

The majority of the studies described above employ a custom bioinformatic pipeline to identify mutations from MA lines. In the widely used ‘consensus’ approach (Ossowski *et al.*, 2010), a putative mutant is called if the majority of reads mapped to a given site differ from the most common base at that site across all samples. In an alternative approach, putative mutations can be identified by using variant calling software to call the most-likely genotype for each MA line and the ancestral line at every site in the genome. In this approach, samples that are inferred to have a genotype that they could not have inherited from the most-likely ancestral genotype are considered mutants (Zhu *et al.*, 2014). Because these approaches produce many false positive mutations, they are usually coupled with post-analysis filtering (based on sequencing coverage, the frequency of rare bases or quality scores) to produce a final set of putative mutations.

In this article, we describe accuMulate, a mutation caller designed for MA experiments. Our approach can replace the custom pipelines and filtering processes currently used to analyze MA experiments with a unified approach to mutation calling. In addition to saving researchers

time in developing custom pipelines, accuMulate will increase the reproducibility of bioinformatic analyses of MA lines.

## 2 Approach

accuMulate uses the probabilistic approach to mutation detection described by Long *et al.* (2016). The probability that a given site in a genomic alignment contains at least one mutation is calculated from a model that directly reflects the design of MA experiments. In particular, we directly model the transition of alleles from an ancestral strain to descendant MA lines and account for the possibility of heterozygous sites in all lines. Our model also accommodates the noise associated with next generation sequencing data by using a Dirichlet-multinomial model to calculate genotype likelihoods (Wu *et al.*, 2017).

For each putative mutation identified by accuMulate, we calculate a suite of statistics that might be used to identify false-positive mutation calls (Li, 2014). In addition to statistics commonly used in existing approaches to mutation detection, this information includes the results of statistical tests for differences in quality control measures between sequencing reads containing apparently mutant bases and those that contain ancestral bases.

MA experiments are frequently undertaken in order to estimate the rate of mutation in a particular strain or species. Accurate estimates of mutation rates require both a numerator (the number mutations detected) and a denominator (the number of sites at which a mutation could have been detected if one were present). Our probabilistic approach to mutation calling provides a straight forward means to estimating this denominator. Mutations can be simulated for a given sample at a given site by altering bases in sequencing reads from that sample. Only those sites that generate a mutation probability greater than the threshold used for mutation detection and pass all additional filtering criteria used in a particular analysis should count towards the denominator for mutation rate calculations (Long *et al.*, 2016).

## 3 Implementation

The accuMulate package is written in C++ and contains two executable files. The main program, accuMulate is a command line program that takes a single genomic alignment (in BAM format) with sequencing reads from the ancestral line (if they are available) and all MA lines as input. A number of additional arguments can be passed to accuMulate to customize an analysis to a particular experiment. These arguments can be passed via the command line or through a simple text file. accuMulate writes information for each site found to have a mutation probability higher than a user-set threshold. A second executable, denominate can be used to calculate the number of sites at which mutation could have been detected if one was present using the same parameters and filtering criteria used in mutation calling.

## 4 Demonstration

We demonstrate the use of accuMulate by reanalysing data generated from a previously published MA experiment. Shaw *et al.* (2000) allowed several lines of *Arabidopsis thaliana* to accumulate mutations. These lines have been the subject of two sequencing efforts. Ossowski *et al.* (2010) sequenced individuals from five lines to identify putative mutations, which they then validated by Sanger sequencing. Subsequently, Becker *et al.* (2011) generated longer sequencing reads from individuals representing 12 MA lines for an analysis of DNA methylation. We downloaded sequencing data from five of the lines Becker *et al.* (2011) sequenced, including two

lines analysed by Ossowski *et al.* (2010). By comparing mutation calls generated by accuMulate with the location of validated mutations (Wei *et al.*, 2014), we are able to demonstrate both the sensitivity of accuMulate and the degree to which the various summary statistics reported for each site differ among validated mutations and other putative mutants.

A summary of this demonstration is described in a document provided as a [Supplementary Material](#). We show accuMulate was able to recover all validated mutations along with a number of putative mutations that were not reported by Ossowski *et al.* (2010). The mapping quality and insert-size statistics reported by accuMulate differ substantially between validated and non-validated mutations. We were able to use these statistics to filter likely false positives from all lines and, using denominate, to estimate the number of callable sites under these filtering criteria. Analysing this data further produces results similar to those reported by Ossowski *et al.* (2010). Our point-estimate of the mutation rate is slightly higher ( $8.3 \times 10^{-9}$  base substitutions per site per generation compared  $7 \times 10^{-9}$  in the published work), but both studies show a mutational spectrum that is strongly biased toward G:C > A:T transitions. These results show that accuMulate accurately identifies mutations from MA experiments. The statistics reported for each putative mutation provide researchers with a straightforward way to detect potential false positives and denominate can generate a direct estimate of the number of callable sites in a given experiment. The reproducible data analysis in [Supplementary Material](#) shows how the files produced by these programs can be used to produce the sorts of results usually reported from MA experiments. The accuMulate distribution thus allows all of the key steps in the analysis of sequencing data from an MA experiment to be undertaken in a single framework.

## Funding

This work was supported by the National Institute of General Medical Sciences [grant number R01GM101352 to RAZ, RBRA, and RAC] and an Arizona State University SOLUR scholarship to A.H.

*Conflict of Interest:* none declared.

## References

- Becker, C. *et al.* (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, **480**, 245–249.
- Dillon, M.M. and Cooper, V.S. (2016) The fitness effects of spontaneous mutations nearly unseen by selection in a bacterium with multiple chromosomes. *Genetics*, **204**, 1225–1238.
- Li, H. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843.
- Long, H. *et al.* (2016) Low base-substitution mutation rate in the germline genome of the ciliate *Tetrahymena thermophila*. *Genome Biol. Evol.*, **8**, 3629–3639.
- Lynch, M. *et al.* (2016) Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.*, **17**, 704–714.
- Ossowski, S. *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.
- Shaw, R.G. *et al.* (2000) Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics*, **155**, 369–378.
- Tenaillon, O. *et al.* (2016) Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, **536**, 165–170.
- Wei, W. *et al.* (2014) Smal: a resource of spontaneous mutation accumulation lines. *Mol. Biol. Evol.*, **31**, 1302.
- Wu, S.H. *et al.* (2017) Estimating error models for whole genome sequencing using mixtures of dirichlet-multinomial distributions. *Bioinformatics*, **33**, 2322–2329.
- Zhu, Y.O. *et al.* (2014) Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci. USA*, **111**, E2310–E2318.