# accuMUlate: A mutation caller designed for mutation accumulation experiments

David J. Winter [1,2,*], Steven H. Wu [1,3], Abigail A. Howell [1,4], Ricardo B. R. Azevedo [5], Rebecca A. Zufall [5] & Reed A. Cartwright [1,4]

[1] The Biodesign Institute, Arizona State University, Tempe, AZ 85281 USA,

[2] Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand,

[3] Bioconsortia, Davis, CA 95618 USA

[4] School of Life Sciences, Arizona State University, Tempe, AZ 85281 USA,

[5] Department of Biology and Biochemistry, University of Houston, Houston, TX 77004 USA.

[*] To whom correspondence may be addressed: david.winter@gmail.com

## Abstract

**Motivation**: Mutation accumulation (MA) experiments are the most widely used method for directly studying the effects of mutation. Modern sequencing technologies have led to an increased interest in MA experiments. 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. Studies that combine MA with whole genome sequencing use custom bioinformatic pipelines that implement heuristic rules to identify putative mutations.

**Results**: We describe accuMUlate, a program that is designed to detect mutations from MA experiments. accuMUlate implements a probabilistic model that reflects the design of a typical MA experiments while being flexible enough to accommodate properties unique to any particular experiment. For each putative mutation identified from this model accMUlate calculates a set of summary statistics that can be used to filter sites that may be false positives. A companion tool, denominate, can be used to apply filtering rules based on these statistics to simulated mutations and thus identify the number of callable sites per sample.

**Availability**: Source code and releases available from https://github.com/dwinter/accuMUlate.

## Introduction

Mutations are important contributors to diverse processes such as adaptation, genome evolution, and the development of disease. Mutation accumulation (MA) experiments are 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 selection, allowing all but the most deleterious mutations to drift to fixation over the course of an experiment. 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 genotype. Studies combining MA experiments with whole-genome sequencing have provided key insights into the evolution of mutation rates [Long et al., 2016, Lynch et al., 2016, Sung et al., 2012], genome evolution [Tenaillon et al., 2016], the molecular basis of mutation [Ness et al., 2015, Zhu et al., 2014, Behringer and Hall, 2016], and the distribution of fitness effects among spontaneous mutations [Heilbron et al., 2014, 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 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. The particular filtering approaches used differ among studies. Putative mutations can then be validated by single-locus sequencing.

In this paper 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.

## Approach

accMUlate 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 (including sequencing depth and mean mapping quality), 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 genotype. Accurate estimates of mutation rates require both a numerator (the number of sites at which a mutation was 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 straightforward 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 filtering criteria should count towards the denominator for mutation rate calculations [Long et al., 2016].

## Implementation

The accuMUlate package is written in C++ and contains two executable files. The main program, accMUlate, takessingle 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. This information is provided in a tabular format, the first three columns of which correspond to the location of the putative mutation on the reference genome in standard BED format.

A second executable , denominate, can be used to calculate the number of sites at which mutation could have been detected if one was present by using the same parameters used in mutation calling and applying filtering criteria based on the statistics reported by accuMUlate. An additional repository, accuMUlate-tools (https://github.com/dwinter/accuMUlate-tools) provides a number of smaller utilities that may help users prepare files for accuMUlate and analyze the results.

The software is provided under an MIT license and is available from https://github.com/dwinter/accuMUlate/. The wiki at this repository (https://github.com/dwinter/accuMUlate/wiki) provides detailed instructions on how to compile the software, how to prepare files for analysis and how to interpret output files.

## Demonstration

We demonstrate the use of accMUlate by reanalysing data generated from a previously published MA experiment. Shaw et al. [2000] accumulated mutations in several lines of *Arabidopsis thaliana*, which 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 accuMU-

LATE and the degree to which the various summary statistics differ among validated mutations and other putative mutants.

A summary of this demonstration is described in a document provided as a supplementary file. As described in that supplement, 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, suggesting these statistics may help to filter out false-positive mutations. A software repository including all of the code used to perform these analyses is provided in the supplementary files associated with this paper. This repository demonstrates how the output files produced by ACCUMULATE can be used to produce results of the sort required for the publication of MA experiments. Moreover, the repository shows how ACCMULATE can contribute to reproducible analyses of sequencing data arising from MA lines.

# References


Claude Becker, Jörg Hagmann, Jonas Müller, Daniel Koenig, Oliver Stegle, Karsten Borgwardt, and Detlef Weigel. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, 480(7376):245–249, 2011.

Megan G Behringer and David W Hall. Genome-wide estimates of mutation rates and spectrum in *Schizosaccharomyces pombe* indicate CpG sites are highly mutagenic despite the absence of DNA methylation. *G3: Genes| Genomes| Genetics*, 6(1):149–160, 2016.

Marcus M. Dillon and Vaughn S. Cooper. The fitness effects of spontaneous mutations nearly unseen by selection in a bacterium with multiple chromosomes. *Genetics*, 204(3):1225–1238, 2016. ISSN 0016-6731. doi: 10.1534/genetics.116.193060. URL http://www.genetics.org/content/204/3/1225.

Karl Heilbron, Macarena Toll-Riera, Mila Kojadinovic, and R. Craig MacLean. Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics*, 197(3):981–990, 2014. ISSN 0016-6731. doi: 10.1534/genetics.114.163147. URL http://www.genetics.org/content/197/3/981.

Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843, 2014. doi: 10.1093/bioinformatics/btu356. URL +http://dx.doi.org/10.1093/bioinformatics/btu356.

Hongan Long, David J Winter, Allan Y-C Chang, Way Sung, Steven H Wu, Mariel Balboa, Ricardo BR Azevedo, Reed A Cartwright, Michael Lynch, and Rebecca A Zufall. Low base-substitution mutation rate in the germline genome of the ciliate *Tetrahymena thermophila*. *Genome Biology and Evolution*, page evw223, 2016.

Michael Lynch, Matthew S Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W Kelley Thomas, and Patricia L Foster. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714, 2016.

Rob W. Ness, Andrew D. Morgan, Radhakrishnan B. Vasanthakrishnan, Nick Colegrave, and Peter D. Keightley. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Research*, 25(11):1739–1749, 2015. doi: 10.1101/gr.191494.115. URL http://genome.cshlp.org/content/25/11/1739.abstract.

Stephan Ossowski, Korbinian Schneeberger, José Ignacio Lucas-Lledó, Norman Warthmann, Richard M Clark, Ruth G Shaw, Detlef Weigel, and Michael Lynch. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *science*, 327(5961):92–94, 2010.

Ruth G Shaw, Diane L Byers, and Elizabeth Darmo. Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics*, 155(1):369–378, 2000.

Way Sung, Abraham E Tucker, Thomas G Doak, Eunjin Choi, W Kelley Thomas, and Michael Lynch. Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. *Proceedings of the National Academy of Sciences*, 109(47):19339–19344, 2012.

Olivier Tenaillon, Jeffrey E Barrick, Noah Ribeck, Daniel E Deatherage, Jeffrey L Blanchard, Aurko Dasgupta, Gabriel C Wu, Sebastien Wielgoss, Stephane Cruveiller, Claudine Medigue, et al. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536(7615):165–170, 2016.

Wen Wei, Lu-Wen Ning, Yuan-Nong Ye, Shi-Jie Li, Hui-Qi Zhou, Jian Huang, and Feng-Biao Guo. Smal: A resource of spontaneous mutation accumulation lines. *Molecular Biology and Evolution*, 31(5):1302, 2014. doi: 10.1093/molbev/msu073. URL +http://dx.doi.org/10.1093/molbev/msu073.

Steven H Wu, Rachel S Schwartz, David J Winter, Don Conrad, and Reed A Cartwright. Estimating error models for whole genome sequencing using mixtures of dirichlet-multinomial distributions. *Bioinformatics*, 2017. doi: 10.1093/bioinformatics/btx133.

Yuan O. Zhu, Mark L. Siegal, David W. Hall, and Dmitri A. Petrov. Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences*, 111(22):E2310–E2318, 2014. doi: 10.1073/pnas.1323011111. URL http://www.pnas.org/content/111/22/E2310.abstract.