Allele age estimators designed for whole genome datasets show only a moderate reduction in performance when applied to whole exome datasets.

Alyssa Pivirotto (1), Noah Peles (1), Jody Hey (1)

(1) Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA USA

Corresponding Author: hey@temple.edu

Short Title: Allele age estimators perform well on exome data.

Keywords: allele age, whole exome sequencing, exome, WES, personalized genomics

Abstract

Personalized genomics in the healthcare system is becoming increasingly accessible as the costs of sequencing decreases. With the increase in the number of genomes, larger numbers of rare variants are being discovered, leading to important initiatives in identifying the functional impacts in relation to disease phenotypes. One way to characterize these variants is to estimate the time the mutation entered the population. However, allele age estimators such as those implemented in the programs Relate, Genealogical Estimator of Variant Age (GEVA), and Runtc, were developed based on the assumption that datasets include the entire genome. We examined the performance of each of these estimators on simulated exome data under a neutral constant population size model, as well as under population expansion and background selection models. We found that each provides usable estimates of allele age from whole-exome datasets. Relate performs the best amongst all three estimators with Pearson coefficients of 0.83 and 0.73 (with respect to true simulated values, for neutral constant and expansion population model, respectively) with a 12 percent and 20 percent decrease in correlation between whole genome and whole exome estimations. Of the three estimators, Relate is best able to parallelize to yield quick results with little resources, however, Relate is currently only able to scale to thousands of samples making it unable to match the hundreds of thousands of samples being currently released. While more work is needed to expand the capabilities of current methods of estimating allele age, these methods show a modest decrease in performance in the estimation of the age of mutations.

Article Summary

Increasing availability of whole exome sequencing yields large numbers of rare variants that have direct impact on disease phenotypes. Many methods of identifying the functional impact of mutations exist including the estimation of the time a mutation entered a population. Popular methods of estimating this time assume whole genome data in the estimate of the allele age based on haplotypes. We simulated genome and exome data under a constant and expansion population demography model and found that there is a decrease in performance in all three methods on exome data of 15-30% depending on the method. Testing the robustness of the best performing method, Relate, further simulations introducing background selection and varying the sample size were also undertaken with similar results.

Introduction

With the rapid advancement in sequencing technology and the availability of genomic data, such data has become integral to investigating the genetic cause of many major diseases. A major benefit of whole genome sequencing is the identification of mutations across the entire genome including non-coding regions, however it remains the costlier option in comparison with whole-exome sequencing (Warr et al. 2015; Schwarze et al. 2020; Almogy et al. 2022). Whole-exome sequencing (WES) targets the coding regions of the genome allowing for investigation into functional genomic mutations, which can more easily be associated with phenotypic changes and these mutations are found more commonly in databases of disease-causing variants (McLendon et al. 2008; Landrum et al. 2018; Karczewski et al. 2020; Piñero et al. 2020). This is more easily accomplished for monogenic diseases where just one gene contributes to a disease phenotype, however, improved methods have made understanding and identifying potential underlying mutations to polygenic diseases possible (Cano-Gamez and Trynka 2020). In

69 particular, rare variants have been found to be associated with complex human diseases and
70 phenotypes (Momozawa and Mizukami 2021) making variants at low frequency of particular
71 interest to researchers.

72 As increasing numbers of genomes are sequenced, researchers are finding more rare
73 variants, some of which have been shown to be causative of human disease (Yang et al. 2013). It
74 is becoming more common for healthcare providers to perform exome sequencing to better
75 understand the underlying cause of a patient's condition (Goh and Choi 2012; Ma et al. 2019;
76 Salfati et al. 2019). However, while these studies can identify potentially causative mutations
77 through association studies, the history of these mutations are not as easily able to be examined
78 except through large population studies (Grzymski et al. 2020). Currently, there are initiatives to
79 expand current genomic resources with corresponding medical information (Carey et al. 2016;
80 Grzymski et al. 2019; Backman et al. 2021). In the UK, researchers have released nearly 500,000
81 exome datasets corresponding to other participant information collected over the last fifteen
82 years (Backman et al. 2021). Other similar biobanks operated primarily by healthcare systems
83 are also utilizing genotype and exome data in their electronic health record collection (Carey et
84 al. 2016; Grzymski et al. 2019) with the United States launching the "All of Us" initiative which
85 was expanded to include 250,000 whole genome sequences (The "All of Us" Research Program
86 2019).

87 One way to better understand the history of a potential causative variant is to estimate
88 when that mutation was introduced to a population. Combining the approximate age of an allele
89 with allele frequency allows for a better understanding of the selective pressure the mutation has
90 undergone (Slatkin 2000). Studies that attempt to understand the selective pressure on rare
91 causative variants of disease rely on small whole-genome studies which focus on just a handful

92 of individuals who carry a particular mutation of interest (Hateley et al. 2021). A previous study
93 focused on candidate genes to utilize the absolute allele age to compare the ages of shared
94 variants in the population to private variants only found in the probands (Wilfert et al. 2021),
95 while a second study utilized relative allele age comparing the age of the candidate risk allele to
96 other variants only shared between the two carriers (Hateley et al. 2021). In both cases, the
97 studies utilized whole genome data limiting their analysis to just a small portion of all publicly
98 available genotype data as many large population studies have worked to expand the database of
99 WES sequences yielding very large datasets. One example of this is the gnomAD project which
100 contains over 750,000 exomes contributing to about 90% of the total data represented (Chen et
101 al. 2024).

102 In one previous study, exome data was used to estimate the age of alleles in a dataset of
103 several thousand exomes leveraging a coalescent-based method with results indicating that most
104 deleterious variants arose recently in human history (Fu et al. 2013). In more recent years,
105 improved methods to estimate allele age have been published that yield more accurate estimates
106 utilizing haplotype-based estimates of genealogies (Platt et al. 2019; Albers and McVean 2020)
107 and ancestral recombination graphs (Rasmussen et al. 2014; Speidel et al. 2019; Wohns et al.
108 2022; Zhang et al. 2023). These more recent methods of allele age estimation have been shown
109 to estimate the age of a mutation from WGS with relative accuracy, however little to no work has
110 been done to establish accuracy of these methods when using WES data. The three methods
111 included here, identified by the programs that implement them, are Relate, GEVA, and Runtc.

112 All three methods used here rely on mutation accumulation on shared haplotypes to
113 accurately estimate the branch length of the branch containing the focal mutation. When
114 estimating from WES data, the missing mutations in the non-coding regions will introduce

115 increased variance in comparison with WGS due to there being less data utilized in the estimates.
116 We may also expect there to be an increase in bias depending on the estimator's underlying
117 assumptions and methodology. For example, in the case of Runtc, the missing mutations would
118 mean that the estimated shared haplotypes are much longer than the true haplotypes. This would
119 then estimate a much shallower tree in comparison with its true deeper tree yielding an
120 underestimation of allele age.

121 Relate provides estimates of allele age as part of an estimate of the ancestral
122 recombination graph (ARG) and is one of several methods that do so (Rasmussen et al. 2014;
123 Ignatieva et al. 2021; Wohns et al. 2022; Zhang et al. 2023). Relate was chosen for this analysis
124 as it performed similarly in accuracy and scalability in comparison with other methods (Zhang et
125 al. 2023). In the methodology for Relate (Speidel et al. 2019), the local genealogy underlying a
126 focal site is reconstructed based on clustering from a distance matrix of mutation variation
127 between haplotype pairs computed from a hidden Markov model (HMM). Then using a
128 coalescent prior, Relate employs a Monte Carlo Markov Chain (MCMC) algorithm to estimate
129 the time of the branch containing the focal mutation. Similarly, Genealogical Estimator of
130 Variant Age (GEVA) (Albers and McVean 2020) also employs an HMM to identify shared and
131 non-shared haplotype pairs, but then subsequently samples from those pairs before calculating
132 the posterior estimate of the age of the mutation from a joint clock model leveraging
133 recombination and mutation. Estimates of the time of first coalescence (t_c) by the Runtc program
134 (Platt et al. 2019) are based upon the maximum shared haplotype (msh) which can be bounded
135 by either a recombination or mutation event. First coalescent time is used as a proxy for allele
136 age and is estimated using the likelihood of the msh.

To assess performance, data sets were simulated under a constant population size model, a more complex demographic model considering the expansion of human populations out of Africa, and a model that included background selection. For each simulation, the true age of each mutation is compared to the estimates of the age generated from each program. Because rare variants are so integral to understanding human disease (Wang et al. 2021), we also examined specifically of how estimators performed on rare variants.

Methods

Simulated Datasets

Data sets were simulated with neutral mutations under both a simple, constant population model (Model id: PiecewiseConstant) and a complex demography model (Model id: OutOfAfrica_3G09) (Gutenkunst et al. 2009) using msprime version 1.2.0 (Kelleher et al. 2016) and stdpopsim version 0.2.0 (Adrion et al. 2020). Simulations were run to generate sequences that are a facsimile of human chromosome 22. They were 51megabases in length and were generated using the GRCh38 HapMap genetic map for chromosome 22 (The International HapMap Consortium et al. 2007). For the simple model, a cohort of 7242 genomes was sampled, and from the complex model, 7242 CEU genomes were sampled. Full parameters for each simulation model can be found in the supplement (Tables S1 and S2). From each simulation, segregating sites were output in VCF format using tskit version 0.5.0 (Kelleher et al. 2018) without any mispolarization or sequencing errors introduced. The output from the simulation will serve as the whole genome dataset for chromosome 22 for downstream analyses. Sites with multiple mutations were excluded from the analysis. True mutation ages are extracted from the simulated ancestral recombination graphs (ARGs) using the tree file. The entire simulation pipeline is described in Supplemental Figure 1.

Conversion to Exome Data

For each model, the simulated sequences were filtered for the coding regions of chromosome 22 using a bed file of exonic regions generated from a bed file containing all autosome coding regions from the UK Biobank (available at: https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=3803) (Backman et al. 2021). Using the BCFtools version 1.10.2-27-g9d66868 (Danecek et al. 2021) filter function, the VCF outputs from stdpopsim were filtered just for the segregating variants that fell within the regions in the supplied bed file. These filtered VCF files serve as the WES data in the analyses. Each of these original simulation runs are annotated as WGS or WES, and then either SIMPLE for constant population size or COMPLEX for the population expansion model hereafter. In total, four VCF files were prepared, one for each of the possible conditions of simple and complex demography, and exome and whole genome (identified as WES_Simple, WES_Complex, WGS_Simple, and WGS_Complex).

Allele Age Estimates

Genealogical Estimate of Variant Age (GEVA) Estimates (Albers and McVean 2020)

The chromosome column of each of the four VCF files first were converted to number format using BCFtools (Danecek et al. 2021) as required by GEVA. Each VCF file was then converted to a binary file using the GRCh38 HapMap genetic map (The International HapMap Consortium et al. 2007) supplied from stdpopsim (Adrion et al. 2020) annotated with the additional map position column in centimorgans (cM). To run the estimator program for all sites of a frequency of 2 copies or more, a list of all sites (k>= 2) was extracted and split into batches of 300 variants due to memory constraints as recommended in the documentation. For the simple model, the

185 mutation rate was 1e-8 and the effective population size was 10000. For the complex model, the
186 mutation rate was 2.35e-8 and the effective population size was 7300, based on the values used
187 for the simulations. Estimates from the Joint Clock model that passed the supplied heuristic filter
188 were used. This includes the mode of the posterior distribution of the concordant and discordant
189 sampled haplotype pairs. Conversion and estimator programs are available at
190 https://github.com/pkalbers/geva.
191

192 **Time of Coalescence (t_c) Estimates (Platt et al. 2019)**

193 Time of coalescence, t_c, maximizes the likelihood of recombination and mutation events over the
194 maximum shared haplotype, msh, as determined by the shared haplotypes tract surrounding a
195 focal allele in a sample of genomes. However, in the case of missing data due to only
196 sequencing the coding regions, the msh may start or end due to mutation or recombinational
197 events in the intronic or intergenic regions. In the case of exome sequence data, the dataset is
198 missing mutations between the exons. Since this data is missing, the end of the shared haplotype
199 would not be identified until the next closest exon. To account for this we implemented a
200 modified estimator in which the shared haplotype terminates at a random position between the
201 exon where the haplotype was found to end and the next closest exon to the focal mutation. With
202 modified msh value measured to include this random distance, the maximum likelihood
203 estimator of t_c is calculated as originally described (Platt et al. 2019). Also following Platt et al.
204 (2019) prior to estimating coalescent times of singletons (variants with just one copy in a
205 population) VCF files were first phased for singletons by placing the mutation on the longer of
206 the two possible haplotypes for each singleton in the sample.
207

208 The Runtc program was run with groups based on their k (copy frequency number) value using
209 the flag --k-range to speed up the estimation pipeline. For the simple and complex models, the
210 mutation rates and effective population size values were based on the values used for
211 simulations, and the same as used for GEVA. The same genetic map as used for the simulations
212 and other estimates for GRCh38 was used with the --map flag. This was done on all variants
213 across the frequency spectrum. Phasing and estimator programs available at
214 https://github.com/jaredgk/runtc.
215
216 **Relate Estimates (Speidel et al. 2019)**
217 Using Relate version 1.1.9, each of the four VCF files were converted to haps/sample format.
218 Each analysis was run using 12 threads and Relate's supplied parallelization script. For the
219 simple and complex models, the mutation rates and effective population size values were based
220 on the values used for simulations, and the same as used for GEVA. Additionally, for the
221 complex model, the full pipeline was run to estimate the allele age based on branch lengths re-
222 estimated after estimating the historical population size. The same genetic map as used for the
223 simulations and other estimates for GRCh38 was used with the --map flag. Relate is available at
224 https://myersgroup.github.io/relate/index.html.
225
226 *Statistics*
227 For every combination of estimator and simulation, four statistics were calculated: Root Mean
228 Square Log Error (RMSLE) between estimated and true ages, and bias, Pearson's R, and
229 Spearman's R between log transformed estimated mutational ages and log transformed true ages
230 calculated from the simulation.

231

232 *Comparisons across Sample Size*

233 For both the simple and complex models described above, a spectrum of sampled genomes was
234 extracted to compare accuracy across sample sizes. Sample sizes of 100, 500, 1000, 2500, 5000,
235 7500, 10000, and 15000 genomes were used. The same parameters as described above for each
236 simple and complex model were used. Following the same pipeline as above (Figure S1), the
237 Tree file for each simulation was extracted and converted to a VCF which was then filtered for
238 exon sites. Using just Relate, allele ages for the WES versions of the simulation samples were
239 estimated.

240

241 *Simulating sites with background selection*

242 Following the pipeline described in Figure S1, stdpopsim version 0.2.0 (Adrion et al. 2020) was
243 used to simulate mutations under background selection using SLiM version 4.1 (Haller et al.
244 2019) and the Gamma distribution of fitness effects based on Kim et al (Kim et al. 2017). Relate
245 was then used to estimate the ages of mutations from both the original entire chromosome output
246 from the simulation (WGS) and the filtered exome version (WES) with the same parameters as
247 described above.

248

249 ## Results

250 To better understand the accuracy of current methods of estimating allele age on whole-
251 exome data, the ages of neutral mutations were simulated under both a simple, constant
252 population size model and a more complex model. The SNP data directly generated from these
253 simulations serve as the whole-genome dataset (WGS). For the simulated whole-exome data
254 (WES), the generated simulated data was filtered based on the exome regions sequenced from

255 the UK Biobank (Backman et al. 2021). For each of the simulations (simple and complex, WES
256 and WGS), three estimators of the age of mutations were generated using the programs Relate
257 (Speidel et al. 2019), GEVA(Albers and McVean 2020), and Runtc (Platt et al. 2019).

258 Each method provided estimates for a different subset of variants depending on the
259 filtering mechanism implemented. Relate generates estimates for sites that pass the filtering
260 threshold with all sites identified as non-mapping or flipped being excluded. The t_c estimator
261 (implemented in Runtc) can be applied to all mutations, including those found just once (i.e.
262 singletons). GEVA generates estimates for mutations of a frequency of 2 copies or more, and in
263 addition required that some sites near the chromosome termini be excluded.

264 Relate estimated allele age the fastest, completing the estimation of allele age for the
265 WES dataset in less than one hour for both demography models, and for the WGS in less than 12
266 hours for the simple model and around 13.5 hours for the complex model (Table S3). Both
267 GEVA and Runtc took much longer as neither can easily parallelize. In the case of both
268 estimators, sites can be separated into batches to allow for each batch to be run separately,
269 however if forced to run sequentially both estimators would take on the magnitude of days to
270 complete even for the much smaller WES dataset. Both GEVA and Relate (when run in parallel
271 with multiple threads) required large amounts of memory and required dedicated computational
272 resources.

273 ***Relate Outperforms with a Simple, Constant Population Size Model of Neutral Variation.***

274 For each estimator method and simulated dataset, the estimated age of each variant was
275 compared to the variants' corresponding true age. The true mutation age was extracted from the
276 tree files from the simulation output. For each comparison, bias, Pearson's correlation, and

Spearman's correlation was calculated between the log transformed estimated ages and the log transformed true ages.

For mutations simulated under a simple model, a comparison of the true age to the estimated age from WES dataset showed a Pearson's correlation coefficient of 0.83 for the Relate method with both GEVA and t_c with correlations of 0.79 and 0.57 respectively (Figure 1A, 1B, 1C). All three estimators underestimated mutations that were exceptionally old, with t_c having the largest discrepancy in allele age. As would be expected, each estimator did perform better estimating the ages of the mutations from WGS dataset with correlations of 0.94 for Relate (WES: 0.83), 0.92 for GEVA (WES: 0.79), and 0.86 for t_c (WES: 0.56) (Figure 1D, 1E, 1F). t_c saw the biggest shift in correlation in performance between WGS estimates and WES with a 33% drop in the Pearson's correlation value (Figure 1C, 1F) with GEVA and Relate showing similar shifts in the Pearson's correlation with decrease of 14% for GEVA and 12% for Relate between the estimates based on the WGS and WES datasets.

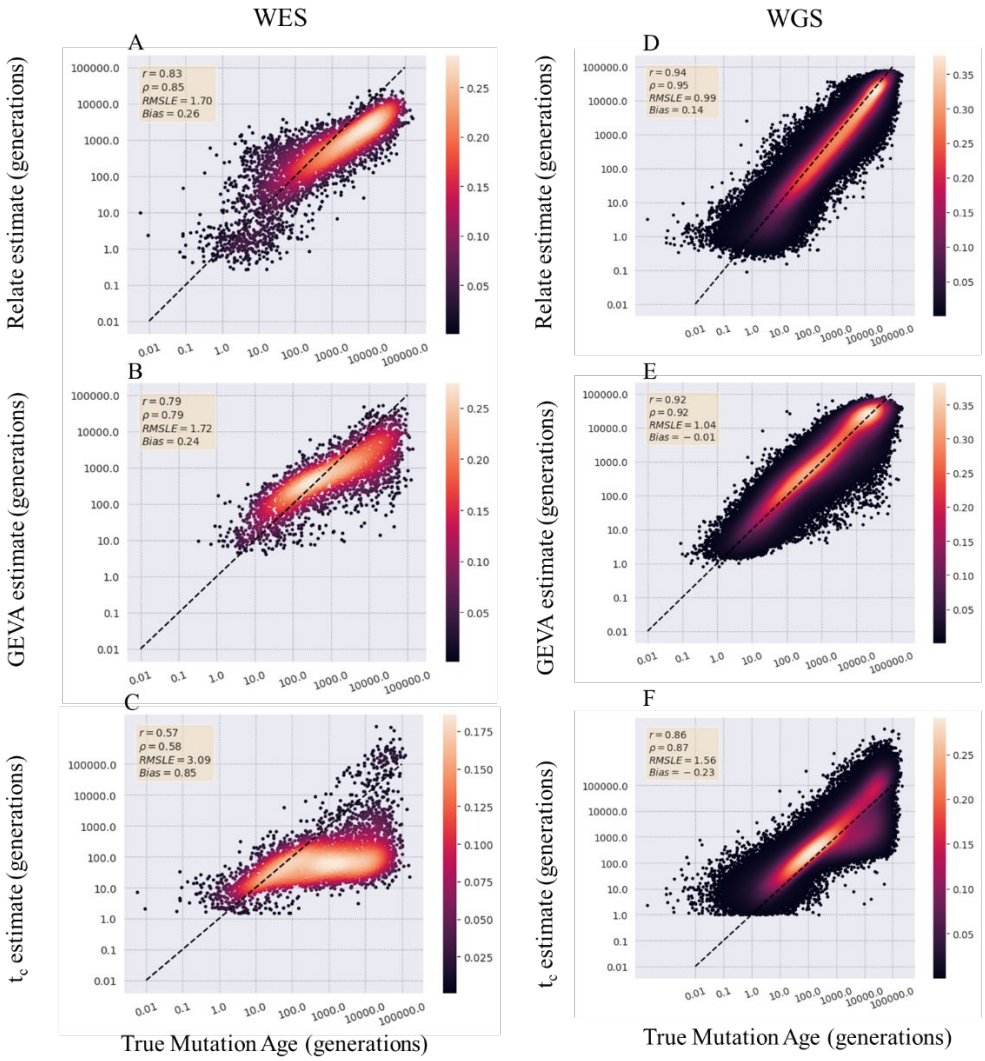

Figure 1. Estimator comparison on WES and WGS data simulated under a simple model. The estimated age of mutations was compared to true age values from variants simulated under a simple constant population size model. Values are reported estimates from WES datasets for Relate (A), GEVA (B), and time of coalescence (C) and from the WGS datasets for Relate (D), GEVA (E), and time of coalescence (F). Results are plotted on a log scale with the dotted black line representing perfect recapitulation of true age values. Points are colored by density calculated by a Gaussian density gradient. Pearson's r, Spearman's ρ, RMSLE, and Bias are reported for each comparison.

291 ***All Three Estimators Show Relative Robustness to Demography and Selection.***

292 To investigate whether estimators of mutational age are robust to more complicated
293 demography models, a more complex out-of-Africa model was simulated (Gutenkunst et al.
294 2009). Again, ages of all polymorphic mutations were estimated using the three estimators and
295 compared with true values. A similar pattern of estimates from the simple model emerged with
296 all three methods underestimating old mutations. Relate and GEVA show similar levels of
297 correlation between the true and estimated ages with Pearson's correlation coefficients of 0.73
298 for both with a lower correlation in the t_c estimated ages with a Pearson's of 0.61 (Figure 2A, 2B,
299 and 2C). Again, Relate and GEVA were better able to estimate the ages of the mutations from
300 WGS data with correlations of 0.91 for Relate (WES: 0.73) and 0.86 for GEVA (WES: 0.73)
301 (Figure 2D, 2E). Relate showed a 20% decrease in correlation between estimates from WGS and
302 WES datasets while GEVA had a slightly smaller decrease in correlation at only 15%.

303 Further testing the robustness of allele age estimation on exome data, the estimator Relate
304 was used to estimate the age of mutations in a sample size of 100, 500, 1000, 2500, 5000, 7500,
305 10000, and 15000 genomes. Relate was used as it was the most tractable for the larger sample
306 size of 15000 genomes, however estimation of mutations in a sample of 100000 genomes was
307 also attempted but required over 15 TB of storage and high memory requirements making it not
308 easily tractable even with dedicated computing resources. It should be noted that an alternative
309 ARG estimate, ARGNeedle, is designed to scale to hundreds of thousands of samples with
310 similar accuracy to Relate and may be utilized with datasets of this magnitude (Zhang et al.
311 2023). With increasing sample size, the Pearson coefficient between estimated age of mutations
312 and true age increased (Figure 3, Figure S2). This was previously reported for GEVA and Relate

313 in which cases it was attributed to the increase in the surrounding mutations around a focal site to

314 estimate the age of the mutation of interest.

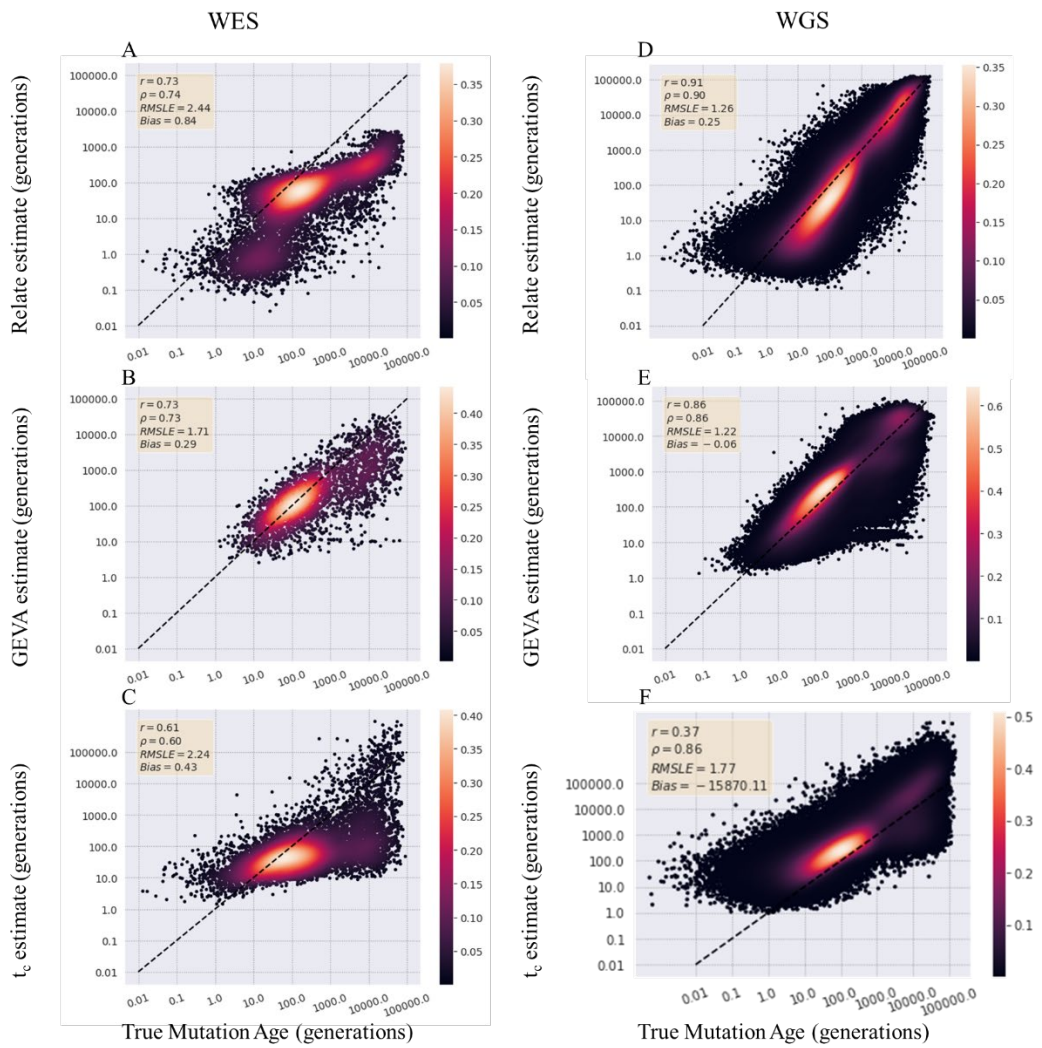

Figure 2. Estimator comparison on WES and WGS data simulated under a complex model. The estimated age of mutations was compared to true age values from variants simulated under a complex population expansion model. Values are reported estimates from WES datasets for Relate (A), GEVA (B), and time of coalescence (C) and from the WGS datasets for Relate (D), GEVA (E), and time of coalescence (F). Results are plotted on a log scale with the dotted black line representing perfect recapitulation of true age values. Points are colored by density calculated by a Gaussian density gradient. Pearson's r, Spearman's ρ, RMSLE, and Bias are reported for each comparison.

In nature, mutations are not just experiencing genetic drift, but often also natural selection, particularly negative selection because of a mutation being deleterious. To simulate this, keeping with the simple and complex model, each model is simulated using SLiM with background selection with a distribution of selection coefficients pulled from a gamma distribution. Relate was then used to estimate the ages of the mutations both using filtered exon dataset (WES) and the entire simulated chromosome (WGS). Relate performed slightly worse on data generated from a simulation with background selection, relative to its performance with no selection, for both the simple and complex models (Figure 4). Relate had an 18.6% decrease in Pearson's correlation for mutations estimated from the WES dataset simulated under a simple model comparison with the WGS dataset. For mutations estimated under the complex model, there was an even higher decrease at 26.7% in Pearson's correlation between the estimates. This is on average a 22% decrease in correlation for estimates under a non-neutral model, higher than the 16% average decrease in correlation between mutations estimated from datasets simulated with no selection.

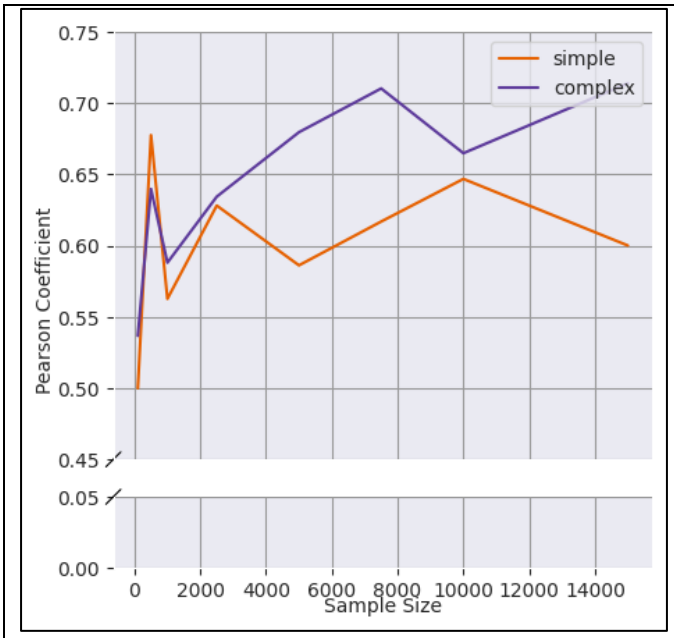

Figure 3. Correlation between true and estimated allele age increases with sample size. Allele ages are estimated from samples of 100, 500, 1000, 2500, 5000, 7500, 10000, and 15000 genomes from simulations of constant population size (purple) and population expansion model (orange). The true age of mutations is compared with the estimated age of mutations estimated with Relate for each of the three sampled set of mutations.

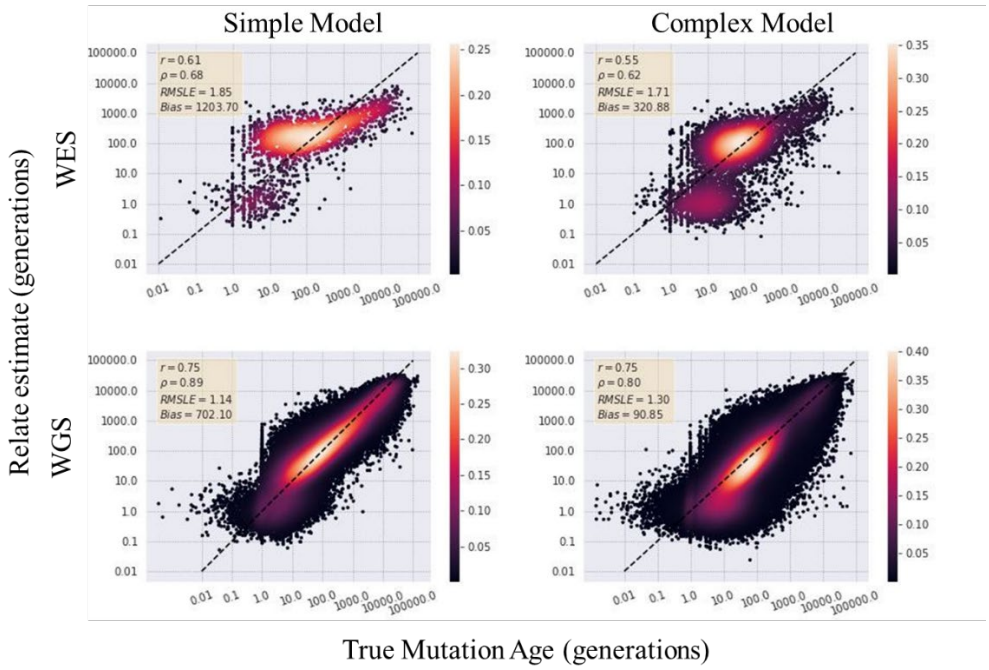

Figure 4. Relate allele age estimates compared to true values under background selection. The allele age of mutations from simulations of sampled genomes under a model which incorporates background selection are estimated using Relate. Upper panels have estimates from whole-exome datasets while lower panels contain estimates from the entire chromosome of mutations.

Rare Variants Show No Decrease in Accuracy of Allele Age Estimates in Comparison with Common Variation.

To address the accuracy of allele age estimators for rare alleles in WES datasets in comparison to more common variants, an analysis of the root mean square log error (RMSLE) averaged across frequency bins was undertaken. Variants were binned into 1% frequency bins and for each estimator method the RMSLE was calculated for that bin (Figure 5). GEVA and Relate had relatively similar RMSLE values for rare variants in comparison with common variants with larger amounts of variation between frequency bins for more common mutations

(Figure 5A). However, time of coalescence showed much higher error rates for common variants (Figure 5A). Similar trends in error rates were found in the complex model (Figure S5). Focusing on rare variation, looking at the changes in error in mutations that just fall within the first 1 percent frequency bin, all three estimators show similar levels of error (Figure 5B). While GEVA is unable to estimate the age of singletons, both Relate and t_c were able to estimate the age of mutations appearing once in the population as well as other low frequency mutations (Figure 5B) with similar accuracy.

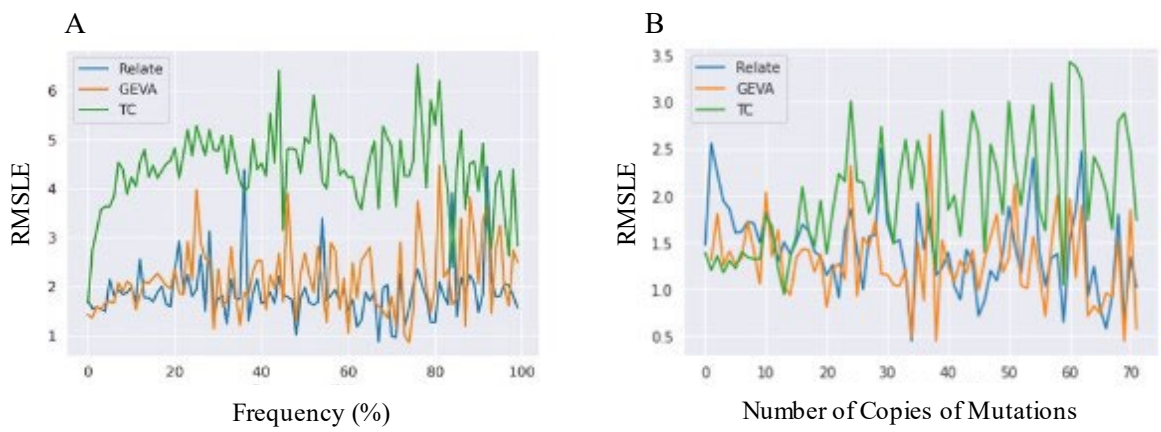

Figure 3. Error in estimates across spectrum of frequency values for the simple model.
For each estimator, sites generated under the simple model are binned by 1% and an average root mean square log error (RMSLE) for that bin was normalized by the average true mutation age of that bin (A). Focusing on just k number of copies for 1% or less, the RMSLE is average for k values ranging from 0 to 72.

Discussion

As sequencing becomes more cost efficient and thus more accessible (Schwarze et al. 2020; Almogy et al. 2022), personalized genomics will also become increasingly accessible. In personalized genomics, a large portion of resources is spent identifying and describing causative

353 mutations of disease (Landrum et al. 2018; Karczewski et al. 2020), and in doing so, it becomes
354 important to understand the functional impacts of the mutation. Currently, there are many
355 methods to characterize a mutation (McLendon et al. 2008; Eilbeck et al. 2017; Landrum et al.
356 2018; Biddanda et al. 2020), with one approach being the estimation of the time of the mutation
357 entered the population (Slatkin and Rannala 2000; Platt et al. 2019; Speidel et al. 2019; Albers
358 and McVean 2020). As these methods are modeled and tested with whole genome sequence data
359 little is known about their ability to estimate from whole exome data. As WES data remains the
360 more cost-efficient option, it is more likely to be used when costs may be the prohibitive factor
361 such as in personalized medicine (Goh and Choi 2012; Salfati et al. 2019). In one potential use
362 case in characterizing a candidate variant of a disease phenotype, researchers estimated the ages
363 of similar frequency variants to compare to the age of the candidate causative variant (Hateley et
364 al. 2021). In studies such as this, allele age can be an additional indicator of the allele's effect
365 with harmful alleles having an expected age that is younger neutral alleles, conditional on
366 frequency. However, it should be noted that these are noisy estimates and would only be useful
367 in conjunction with other data. As age estimates can aid in understanding an allele's effect
368 especially in increasingly large datasets it's important to identify the ability of current estimators
369 to perform this task. Thus, here we conduct a comparison of several methods of allele age
370 estimation to ascertain whether the methods can be successfully applied to whole exome
371 sequence data.

372 Of the three estimators, as expected, all had the highest accuracy in identifying the ages
373 of mutations from WGS data for both the simple and complex demography models with Relate
374 performing the best across all estimators. For WES data, we find that of all three methods Relate
375 again performs best with the highest correlation and lowest bias with a decrease of between 12 to

376 20 percent in correlation in comparison of the estimates on the WGS data. However, all three
377 methods most underestimate the oldest of alleles. Time of coalescence underestimates allele age
378 to the highest degree with the RMSLE tripling between 100 generations to 10000 generations.
379 This indicates that the modified t_c estimator implemented here, that attempts to account for when
380 a haplotype ends in non-coding regions, does not totally mitigate the issue of haplotype tracks
381 ending in the intergenic or intronic regions. As a trend, each method slightly overestimated the
382 youngest of alleles as the bias flips in signs from the youngest (less than 100 generations) to the
383 oldest (greater than 100 generations). This is expected for older mutations, as coalescent events
384 that happen further back in time allow for more mutations to accumulate along that branch. Thus,
385 in estimating allele age from a dataset missing some substantial proportion of the mutations
386 means that the estimators are observing fewer mutations on a branch and estimating it shorter
387 than its true length. The overestimation of the youngest variants and underestimation of the
388 oldest variants has been previously identified in comparison of whole genome sequence data
389 estimates from both Relate and GEVA (Brandt et al. 2022; Ragsdale and Thornton 2023). In the
390 comparison examining the robustness of the methods on simulations expanding beyond just a
391 neutral model to also include deleterious mutations, Relate had a larger decrease in accuracy in
392 the estimates between exome and genome data in comparison with the neutral model (simple:
393 19%, complex: 27%) (Figure 5).

394 All three estimators assume an infinite-site model where only a single mutation is able to
395 arise at each site, however in nature this is not always the case (Karlin and McGregor 1967;
396 Kimura 1969). For simulation data, this is not a concern as sites where a site had more than one
397 mutation occur can be excluded or not allowed to occur. However, for empirical data, repeated
398 mutations may occur especially in regions of the genome where the mutation rate is particularly

399 high such as in CpG sites (Lek et al. 2016). If the infinite-site model is violated with repeated
400 mutation at a given loci with multiple common ancestral haplotypes, then haplotypes with
401 differing genealogical histories would then be pooled together artificially inflating the estimated
402 ages. Relate deals with this issue by excluding sites where mutations cannot be mapped onto the
403 local genealogy (Speidel et al. 2019) while t_c takes a composite of the estimated age of each copy
404 of allele allowing for repeated mutations at a single locus not to skew the estimated ages too high
405 (Platt et al. 2019). However, it is likely that the majority of human variant have arisen by way of
406 single mutations, with violations of the infinite-site model by multiple mutations constituting a
407 very small proportion of sites, given the few known examples (McVean et al. 2002; Blumenfeld
408 and Patnaik 2004; Tishkoff et al. 2007).

409 While common variants have long been identified as potential pools of variants affecting
410 human phenotypes (Burton et al. 2007; Claussnitzer et al. 2020), rare variants are often less well
411 understood and harder to be studied and characterized. This is particularly important because
412 these mutations are much less likely to be found in published large population data such as
413 1KGP (Auton et al. 2015) leading to very few opportunities to learn more about their history
414 except through individualized studies or exceptionally large population studies. While rare
415 variants have a lower error in comparison with common variants across all three estimators, it's
416 important to consider the range of ages for these low frequency variants. Common variants will
417 have a wider range of variation in estimated values for alleles of the same frequency leading to
418 higher error values. Thus, while the absolute age of a rare variant has a smaller error due to
419 smaller variation, the relative position of the age of a rare variant within a given frequency bin
420 may be more error prone to fluctuations in the estimators.

This analysis highlights both the abilities and limitations of utilizing allele age estimators on whole exome data to better understand the history of focal mutations. Expanding the methods of characterization for mutations of interest remains an important question as clinicians and researchers are still attempting to understand mechanisms of many disease phenotypes (Claussnitzer et al. 2020). These estimators demonstrate lower accuracy in absolute age estimates from whole exome data in comparison with whole genome estimates. However, all three estimators perform with similar accuracy on rare variants which introduces a potential opportunity to examine these previously hard to characterize variants. Relate which performs with the highest accuracy, has the potential to offer valuable insights into the history of potentially causative alleles to human disease through the comparison of relative ages among mutations of similar frequency. As the number of sequenced exomes increase with the expansion of personalized genomics, these methods can play a vital role in advancing our understanding of the genetic basis of disease.

Data Availability Statement

Simulations and code are available at https://github.com/ampivirotto/wes_alleleAge.

Acknowledgments

We would like to thank Drs. Alexander Lucaci and Jordan Zehr for their thoughtful editorial insights.

Funding

This work was supported in part by grant R01 GM144468 from the National Institutes of Health to Jody Hey and Sergei Pond and grant 1564659 from the National Science Foundation to Jody

Hey and Arun Sethuraman. This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189.

Conflict of Interest

Authors declare no conflicts of interest.

References

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, et al. 2020. A community-maintained standard library of population genetic models. Coop G, Wittkopp PJ, Novembre J, Sethuraman A, Mathieson S, editors. eLife. 9:e54967. doi:10.7554/eLife.54967.

Albers PK, McVean G. 2020. Dating genomic variants and shared ancestry in population-scale sequencing data. PLOS Biology. 18(1):e3000586. doi:10.1371/journal.pbio.3000586.

Almogy G, Pratt M, Oberstrass F, Lee L, Mazur D, Beckett N, Barad O, Soifer I, Perelman E, Etzioni Y, et al. 2022. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. :2022.05.29.493900. doi:10.1101/2022.05.29.493900. [accessed 2024 Jan 19]. https://www.biorxiv.org/content/10.1101/2022.05.29.493900v4.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. Nature. 526(7571):68–74. doi:10.1038/nature15393.

Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, Benner C, Liu D, Locke AE, Balasubramanian S, et al. 2021. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 599(7886):628–634. doi:10.1038/s41586-021-04103-z.

Biddanda A, Rice DP, Novembre J. 2020. A variant-centric perspective on geographic patterns of human allele frequency variation. Goldberg A, Perry GH, editors. eLife. 9:e60107. doi:10.7554/eLife.60107.

Blumenfeld OO, Patnaik SK. 2004. Allelic genes of blood group antigens: A source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. Human Mutation. 23(1):8–16. doi:10.1002/humu.10296.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447(7145):661–678. doi:10.1038/nature05911.

Cano-Gamez E, Trynka G. 2020. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Frontiers in Genetics. 11. [accessed 2023 Nov 3]. https://www.frontiersin.org/articles/10.3389/fgene.2020.00424.

Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, Murray MF, Smelser DT, Gerhard GS, Ledbetter DH. 2016. The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. Genet Med. 18(9):906–913. doi:10.1038/gim.2015.187.

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, et al. 2024. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 625(7993):92–100. doi:10.1038/s41586-023-06045-0.

Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, et al. 2020. A brief history of human disease genetics. Nature. 577(7789):179–189. doi:10.1038/s41586-019-1879-7.

490 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA,
491 Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience. 10(2):giab008.
492 doi:10.1093/gigascience/giab008.

493 Eilbeck K, Quinlan A, Yandell M. 2017. Settling the score: variant prioritization and Mendelian disease.
494 Nat Rev Genet. 18(10):599–612. doi:10.1038/nrg.2017.52.

495 Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J,
496 et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.
497 Nature. 493(7431):216–220. doi:10.1038/nature11690.

498 Goh G, Choi M. 2012. Application of Whole Exome Sequencing to Identify Disease-Causing Variants in
499 Inherited Human Diseases. Genomics Inform. 10(4):214–219. doi:10.5808/GI.2012.10.4.214.

500 Grzymski JJ, Elhanan G, Morales Rosado JA, Smith E, Schlauch KA, Read R, Rowan C, Slotnick N, Dabe S,
501 Metcalf WJ, et al. 2020. Population genetic screening efficiently identifies carriers of autosomal
502 dominant diseases. Nat Med. 26(8):1235–1239. doi:10.1038/s41591-020-0982-5.

503 Grzymski JJ, Elhanan G, Smith E, Rowan C, Slotnick N, Dabe S, Schlauch K, Read R, Metcalf WJ, Lipp B, et
504 al. 2019. Population Health Genetic Screening for Tier 1 Inherited Diseases in Northern Nevada: 90% of
505 At-Risk Carriers are Missed. [accessed 2022 Jan 7]. https://www.biorxiv.org/content/10.1101/650549v1.

506 Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic
507 History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genetics.
508 5(10):e1000695. doi:10.1371/journal.pgen.1000695.

509 Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. 2019. Tree-sequence recording in SLiM opens
510 new horizons for forward-time simulation of whole genomes. Molecular Ecology Resources. 19(2):552–
511 566. doi:10.1111/1755-0998.12968.

512 Hateley S, Lopez-Izquierdo A, Jou CJ, Cho S, Schraiber JG, Song S, Maguire CT, Torres N, Riedel M, Bowles
513 NE, et al. 2021. The history and geographic distribution of a KCNQ1 atrial fibrillation risk allele. Nat
514 Commun. 12(1):6442. doi:10.1038/s41467-021-26741-7.

515 Ignatieva A, Lyngsø RB, Jenkins PA, Hein J. 2021. KwARG: parsimonious reconstruction of ancestral
516 recombination graphs with recurrent mutation. Bioinformatics. 37(19):3277–3284.
517 doi:10.1093/bioinformatics/btab351.

518 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A,
519 Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456
520 humans. Nature. 581(7809):434–443. doi:10.1038/s41586-020-2308-7.

521 Karlin S, McGregor J. 1967. The number of mutant forms maintained in a population. In: Proceedings of
522 the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 4.

523 Kelleher J, Etheridge AM, McVean G. 2016. Efficient Coalescent Simulation and Genealogical Analysis for
524 Large Sample Sizes. PLOS Computational Biology. 12(5):e1004842. doi:10.1371/journal.pcbi.1004842.

525 Kelleher J, Thornton KR, Ashander J, Ralph PL. 2018. Efficient pedigree recording for fast population
526 genetics simulation. PLOS Computational Biology. 14(11):e1006581. doi:10.1371/journal.pcbi.1006581.

527 Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New
528 Nonsynonymous Mutations Using Large Samples. Genetics. 206(1):345–361.
529 doi:10.1534/genetics.116.197145.

530 Kimura M. 1969. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to
531 Steady Flux of Mutations. Genetics. 61(4):893–903.

532 Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al.
533 2018. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids
534 Research. 46(D1):D1062–D1067. doi:10.1093/nar/gkx1153.

535 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ,
536 Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature.
537 536(7616):285–291. doi:10.1038/nature19057.

538 Ma Y, Jun GR, Zhang X, Chung J, Naj AC, Chen Y, Bellenguez C, Hamilton-Nelson K, Martin ER, Kunkle BW,
539 et al. 2019. Analysis of Whole-Exome Sequencing Data for Alzheimer Disease Stratified by APOE
540 Genotype. JAMA Neurology. 76(9):1099–1108. doi:10.1001/jamaneurol.2019.1456.

541 McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M. Mastrogianakis G, Olson JJ, Mikkelsen T,
542 Lehman N, Aldape K, et al. 2008. Comprehensive genomic characterization defines human glioblastoma
543 genes and core pathways. Nature. 455(7216):1061–1068. doi:10.1038/nature07385.

544 McVean G, Awadalla P, Fearnhead P. 2002. A Coalescent-Based Method for Detecting and Estimating
545 Recombination From Gene Sequences. Genetics. 160(3):1231–1241. doi:10.1093/genetics/160.3.1231.

546 Momozawa Y, Mizukami K. 2021. Unique roles of rare variants in the genetics of complex diseases in
547 humans. Journal of Human Genetics. 66(1):11–23. doi:10.1038/s10038-020-00845-2.

548 Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. 2020. The
549 DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Research.
550 48(D1):D845–D855. doi:10.1093/nar/gkz1021.

551 Platt A, Pivirotto A, Knoblauch J, Hey J. 2019. An estimator of first coalescent time reveals selection on
552 young variants and large heterogeneity in rare allele ages among human populations. PLOS Genetics.
553 15(8):e1008340. doi:10.1371/journal.pgen.1008340.

554 Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-Wide Inference of Ancestral
555 Recombination Graphs. PLOS Genetics. 10(5):e1004342. doi:10.1371/journal.pgen.1004342.

556 Salfati EL, Spencer EG, Topol SE, Muse ED, Rueda M, Lucas JR, Wagner GN, Campman S, Topol EJ,
557 Torkamani A. 2019. Re-analysis of whole-exome sequencing data uncovers novel diagnostic variants and
558 improves molecular diagnostic yields for sudden death and idiopathic diseases. Genome Medicine.
559 11(1):83. doi:10.1186/s13073-019-0702-2.

Schwarze K, Buchanan J, Fermont JM, Dreau H, Tilley MW, Taylor JM, Antoniou P, Knight SJL, Camps C, Pentony MM, et al. 2020. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. Genetics in Medicine. 22(1):85–94. doi:10.1038/s41436-019-0618-7.

Slatkin M. 2000. Allele age and a test for selection on rare alleles. Charlesworth B, Harvey PH, editors. Phil Trans R Soc Lond B. 355(1403):1663–1668. doi:10.1098/rstb.2000.0729.

Slatkin M, Rannala B. 2000. Estimating Allele Age. Annual Review of Genomics and Human Genetics. 1(1):225–249. doi:10.1146/annurev.genom.1.1.225.

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics. 51(9):1321–1329. doi:10.1038/s41588-019-0484-x.

The “All of Us” Research Program. 2019. New England Journal of Medicine. 381(7):668–676. doi:10.1056/NEJMsr1809937.

The International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature. 449(7164):851–861. doi:10.1038/nature06258.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 39(1):31–40. doi:10.1038/ng1946.

Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, Vitsios D, Deevi SVV, Mackay A, Muthas D, et al. 2021. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature. 597(7877):527–532. doi:10.1038/s41586-021-03855-y.

Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. 2015. Exome Sequencing: Current and Future Perspectives. G3 Genes|Genomes|Genetics. 5(8):1543–1550. doi:10.1534/g3.115.018564.

Wilfert AB, Turner TN, Murali SC, Hsieh P, Sulovari A, Wang T, Coe BP, Guo H, Hoekzema K, Bakken TE, et al. 2021. Recent ultra-rare inherited variants implicate new autism candidate risk genes. Nat Genet. 53(8):1125–1134. doi:10.1038/s41588-021-00899-8.

Wohns AW, Wong Y, Jeffery B, Akbari A, Mallick S, Pinhasi R, Patterson N, Reich D, Kelleher J, McVean G. 2022. A unified genealogy of modern and ancient genomes. Science. 375(6583):eabi8264. doi:10.1126/science.abi8264.

Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, et al. 2013. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. New England Journal of Medicine. 369(16):1502–1511. doi:10.1056/NEJMoa1306555.

Zhang BC, Biddanda A, Gunnarsson ÁF, Cooper F, Palamara PF. 2023. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nat Genet. 55(5):768–776. doi:10.1038/s41588-023-01379-x.