Allele age estimators designed for whole genome datasets show only a modest decrease in accuracy 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 number of genomes, larger numbers of rare variants are being discovered and much work is being done to identify their 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 Relate, Genealogical Estimator of Variant Age, and time of coalescence, 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 and found that each provides usable estimates of allele age from whole-exome datasets. To test the robustness of these methods, analyses were undertaken to simulate data under a population expansion model and background selection. Relate performs the best amongst all three estimators with Pearson coefficients of 0.64 and 0.68 (neutral constant and expansion population model) with a 17 percent and 15 percent drop in accuracy 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 even Relate is 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 estimate the age of mutations with a modest decrease in performance.

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 accuracy in all three methods for 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 availability, genomic 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 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 particular, rare variants have been found to

70 be associated with complex human diseases and phenotypes (Momozawa and Mizukami 2021)
71 making variants at low frequency of particular 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 primarily utilizing genotype and exome data in their electronic health record collection
84 (Carey et al. 2016; Grzymski et al. 2019) with the United States launching the "All of Us"
85 initiative which just recently was expanded to include 250,000 whole genome sequences (The
86 "All of Us" Research Program 2019).

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

interest (Hateley et al. 2021). Many large population studies have worked to expand the database of WES sequences combating the problem of sample size and yielding very large datasets. This leads to the obvious avenue of measuring the accuracy of previously used methods on the growing amount of data.

Previous research has attempted to estimate the age of alleles in a dataset containing thousands of exomes using a coalescent method where researchers found that most deleterious variants arose recently in human history (Fu et al. 2013). In more recent years, improved methods to estimate allele age have been published that yield more accurate estimates utilizing haplotype-based estimates of genealogies (Platt et al. 2019; Albers and McVean 2020) and ancestral recombination graphs (Speidel et al. 2019). These more recent methods of allele age estimation have been shown to estimate the age of a mutation from WGS with relative accuracy, however little to no work has been done to establish accuracy of these methods when using WES data. Each of the three estimators of allele age (Genealogical Estimator of Variant Age, Relate, time of coalescence), rely on mutation accumulation on shared haplotypes between pairs of samples to accurately estimate the branch length of the branch containing the focal mutation. The missing mutations in the non-coding regions from estimates from WES data will introduce increased bias in comparison with WGS, but if there is enough mutation in the coding regions the error in estimates may be largely mitigated.

In the methodology for Relate (Speidel et al. 2019), the local genealogy underlying a focal site is reconstructed based on clustering from a distance matrix of mutation variation between haplotype pairs computed from a hidden Markov model (HMM). Then using a coalescent prior, Relate employs a Monte Carlo Markov Chain (MCMC) algorithm to estimate the time of the branch containing the focal mutation. Similarly, Genealogical Estimator of

Variant Age (GEVA) (Albers and McVean 2020) also employs an HMM to identify shared and non-shared haplotype pairs, but then subsequently samples from those pairs before calculating the posterior estimate of the age of the mutation from a joint clock model leveraging recombination and mutation. Time of coalescence (t_c) (Platt et al. 2019) however calculates the maximum shared haplotype (*msh*) which can be bounded by either a recombination or mutation event. Allele age can then be estimated based on a likelihood function leveraging the length of the shared haplotype.

In this study, three estimators of allele age (GEVA, relate, tc) are assessed whether accurate estimates of mutation age can be generated from exome data alone. Simulations under both a constant population size model and a more complex demographic model considering the expansion of human populations out of Africa are undertaken to test the robustness of these methods. In addition to simulations under a neutral scenario, data is also simulated with 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 will additionally assess whether these estimators of allele age can be used successfully on rare variants found in exome datasets.

Methods

Simulated Datasets

Using stdpopsim version 0.2.0 (Adrion et al. 2020), mutations were generated under both a simple, constant population model (Model id: PiecewiseConstant) and a complex demography model (Model id: OutOfAfrica_3G09) (Gutenkunst et al. 2009) with no selection. Both simulations were performed using msprime version 1.2.0 as the simulator (Kelleher et al. 2016) generating neutral mutations. From the simple model, a population of 7242 genomes was

140 sampled, and from the complex model, 7242 CEU genomes were sampled. Full parameters for
141 each simulation model can be found in the supplement (Tables S1 and S2). From each
142 simulation, segregating sites were output in VCF format using tskit version 0.5.0 (Kelleher et al.
143 2018). The output from the simulation will serve as the whole genome dataset for chromosome
144 22 for downstream analyses. Entire simulation pipeline described in Supplemental Figure 1.
145
146 *Convert to Exome Data*
147 For each of the simulations (simple and complex), the simulated chromosome 22 files were
148 filtered for the coding regions. A bed file containing the coding regions for chromosome 22 was
149 generated from a bed file containing all autosome coding regions from the UK Biobank
150 (available at: https://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=3803) (Backman et al. 2021).
151 Using the BCFtools version 1.10.2-27-g9d66868 (Danecek et al. 2021) filter function, the VCF
152 outputs from stdpopsim were filtered just for the segregating variants that fell within the regions
153 in the supplied bed file. These filtered VCF files serve as the WES data in the analyses. Each of
154 these original simulation runs are annotated as WGS or WES, and then either SIMPLE for
155 constant population size or COMPLEX for the population expansion model hereafter. In total,
156 four VCF files were prepared, one for each of the possible conditions of simple and complex
157 demography, and exome and whole genome (identified as WES_Simple, WES_Complex,
158 WGS_Simple, and WGS_Complex).
159
160 *True Values*
161 From the tree output file from stdpopsim (Adrion et al. 2020), corresponding Nodes, Sites,
162 Edges, and Mutations files were extracted using tskit (Kelleher et al. 2018). Using the extracted

163 tree and genomic information from these output files, for each site, the position, the time of the
164 mutation, and the time corresponding to both the parent and child node of the branch containing
165 the mutation was extracted.
166
167 *Allele Age Estimates*
168 **Genealogical Estimate of Variant Age (GEVA) Estimates (Albers and McVean 2020)**
169 The chromosome column of each of the four VCF files first were converted to number format
170 using BCFtools version 1.10.2-27-g9d66868 (Danecek et al. 2021) as required by GEVA. Each
171 VCF file was then converted to a binary file using the GRCh38 HapMap genetic map (The
172 International HapMap Consortium et al. 2007) supplied from stdpopsim (Adrion et al. 2020)
173 annotated with the additional map position column in centimorgans (cM). To run the estimator
174 program for all sites of a frequency of 2 copies or more, a list of all sites (k>= 2) was extracted
175 and split into batches of 300 variants due to memory constraints as recommended in the
176 documentation. The estimator was run with default parameters of 10000 for the effective
177 population size and a mutation rate of 1e-8. Estimates from the Joint Clock model that passed the
178 supplied heuristic filter were used. This includes the mode of the posterior distribution of the
179 concordant and discordant sampled haplotype pairs. Conversion and estimator programs
180 available at https://github.com/pkalbers/geva.
181
182 **Time of Coalescence (t_c) Estimates (Platt et al. 2019)**
183 Time of coalescence, t_c, maximizes the likelihood of recombination and mutation events over
184 the maximum shared haplotype, msh, as determined by the shared haplotypes tract surrounding a
185 focal allele in a sample of genomes. However, in the case of missing data due to only
186 sequencing the coding regions, the msh may start or end due to mutation or recombinational

events in the intronic or intergenic regions. In the case of exome sequence data, the dataset is missing mutations between the exons. Since this data is missing, the end of the shared haplotype would not be identified until the next closest exon. To account for this we implemented a modified estimator in which the shared haplotype terminates at a random position between the exon where the haplotype was found to end and the next closest exon to the focal mutation. With modified msh value measured to include this random distance, the maximum likelihood estimator of t_c is calculated as originally described (Platt et al. 2019). Also following Platt et al. (2019) prior to estimating coalescent times of singletons (variants with just one copy in a population) VCF files were first phased for singletons by placing the mutation on the longer of the two possible haplotypes for each singleton in the sample.

The estimator was run with groups based on their k (copy frequency number) value using the flag --k-range to speed up the estimation pipeline, and was run with default values of 1e-8 for mutation rate and 10000 for effective population size. The same genetic map as used for the simulations and other estimates for GRCh38 was used with the --map flag. This was done on all variants across the frequency spectrum. Phasing and estimator programs available at https://github.com/jaredgk/runtc.

Relate Estimates (Speidel et al. 2019)

Using Relate version 1.1.9, each of the four VCF files were converted to haps/sample format. Using Relate's supplied parallelization script, each simulation was run using 12 threads with a mutation rate of 1e-8 and an effective population size of 10000. The same genetic map as used for the simulations and other estimates for GRCh38 was used with the --map flag. Relate is

210 available at https://myersgroup.github.io/relate/index.html.

211

212 *Statistics*

213 For every combination of estimator and simulation, four statistics were calculated: bias,
214 Pearson's R, Spearman's R, and Root Mean Square Log Error (RMSLE) for each set of
215 mutational ages in comparison with the true age of the mutation as calculated from the
216 simulation.

217

218 *Comparisons across Sample Size*

219 For both the simple and complex models described above, a spectrum of sampled genomes was
220 extracted to compare accuracy across sample sizes. Sample sizes of 100 genomes, 1000
221 genomes, and 10000 genomes were used in addition to the 7242 genomes sampled in the first set
222 of simulations. The same parameters as described above for each simple and complex model
223 were used. Following the same pipeline as above (Figure S1), the Tree file for each simulation
224 was extracted and converted to a VCF which was then filtered for exon sites. Using just Relate,
225 allele ages for the WES versions of the simulation samples were estimated.

226

227 *Simulating sites with background selection*

228 Following the pipeline described in Figure S1, stdpopsim version 0.2.0 (Adrion et al. 2020) was
229 utilized to simulate mutations under background selection using SLiM version 4.1 (Haller et al.
230 2019) and the Gamma distribution of fitness effects based on Kim et al (Kim et al. 2017). Relate
231 was then used to estimate the ages of mutations from both the original entire chromosome output

232 from the simulation (WGS) and the filtered exome version (WES) with the same parameters as
233 described above.
234

235 ## Results

236 To better understand the accuracy of current methods of estimating allele age on whole-
237 exome data, mutations were simulated under both a simple, constant population size model and a
238 more complex model based on Gutenkunst's out-of-Africa demography (Gutenkunst et al. 2009)
239 with no selection. Simulations were generated using msprime (Kelleher et al. 2016) from
240 stdpopsim (Adrion et al. 2020) using HapMap GRCh38 genetic map (The International HapMap
241 Consortium et al. 2007). Complete simulation parameters can be found in the methods and
242 Supplemental Tables 1 and 2. The SNP data directly generated from these simulations serve as
243 the whole-genome dataset (WGS). For the simulated whole-exome data (WES), the generated
244 simulated data was filtered based on the exome regions sequenced from the UK Biobank
245 (Backman et al. 2021). For each of the simulations (simple and complex, WES and WGS), the
246 ages of mutations were estimated with Relate (Speidel et al. 2019), Genealogical Estimation of
247 Variant Age (GEVA) (Albers and McVean 2020), and time of coalescence (t_c) (Platt et al. 2019).

248 Each method provided estimates for a different subset of variants depending on the
249 filtering mechanism implemented. Relate estimated the age of all mutations that passed the
250 filtering threshold with all sites identified as non-mapping or flipped being excluded. t_c estimated
251 the age of all mutations including both singletons and derived mutations found in all but one
252 genome. GEVA estimates the age of mutations of a frequency of 2 copies or more, and
253 additionally several other sites were excluded due to proximity to the end of the chromosome.

254 Relate estimated allele age the quickest, completing the estimation of allele age for the
255 WES dataset in less than one hour for both demography models, and for the WGS in less than 12

256 hours for the simple model and around 13.5 hours for the complex model (Table S3). Both
257 GEVA and t_c took much longer as neither can easily parallelize. In the case of both estimators,
258 sites can be separated into batches to allow for each batch to be run separately, however if forced
259 to run sequentially both estimators would take on the magnitude of days to complete even for the
260 much smaller WES dataset. Both GEVA and Relate (when run in parallel with multiple threads)
261 required large amounts of memory and required dedicated computational resources.

262 For each method, the estimated age of each variant was compared to the variants'
263 corresponding true value for variants under both the simple and complex model. True mutation
264 time was extracted from the tree files from the simulation output. For the simple model,
265 comparing the true age of the mutation to the estimated age from WES data for each of the three
266 methods, a Pearson's correlation coefficient of 0.64 was found for the Relate method with both
267 GEVA and t_c following with correlations of 0.45 and 0.26 respectively (Figure 1). All three
268 estimators underestimated mutations that were exceptionally old, with t_c having the largest
269 discrepancy. Each estimator did perform better estimating the ages of the mutations from WGS
270 dataset with correlations of 0.77 for Relate (WES: 0.64), 0.65 for GEVA (WES: 0.45), and 0.36
271 for t_c (WES: 0.26) (Figure S2). GEVA saw the biggest drop in performance between WGS
272 estimates and WES with a 31% drop (Table 1) with t_c seeing the next biggest drop in
273 performance at a 29% drop and Relate with lowest drop in performance at a 17% deficit.
274

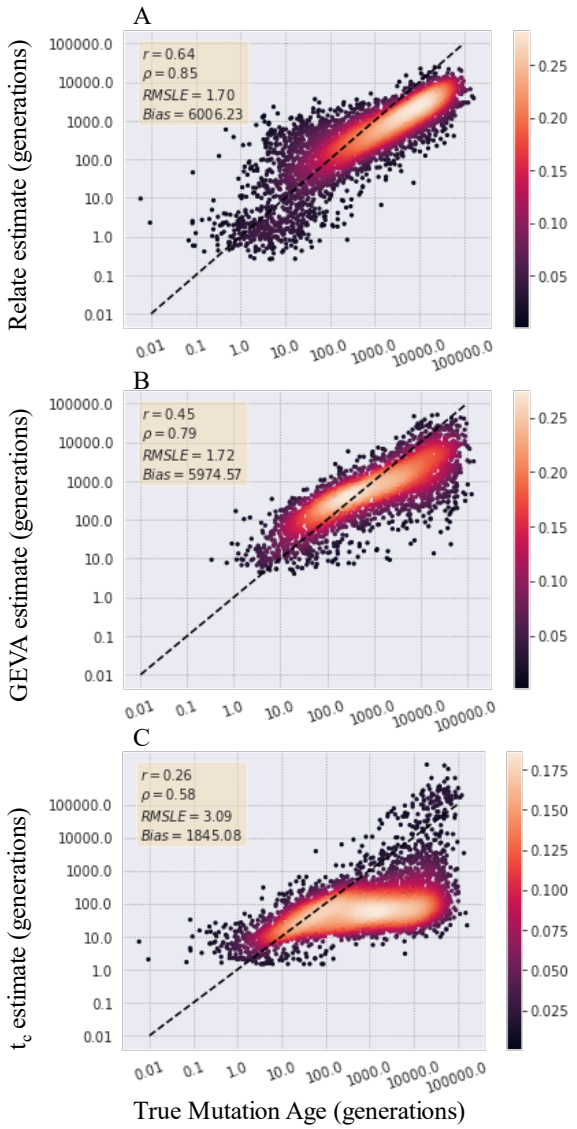

Figure 1. Estimator comparison on WES data for simple model.

The estimated age of mutations for each of the three estimators Relate (A), GEVA (B), and time of coalescence (C) were compared to true age values for a simple constant population size simulation. 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. WGS comparison found in Figure S2.

275 Table 1. Pearson's correlation coefficients for simple and complex model for WES and WGS datasets.

	Simple Model		Complex Model	
	WES	WGS	WES	WGS
Relate (12 threads)	0.6394	0.7743	0.6799	0.7994
GEVA	0.4467	0.6476	0.4510	0.6290
t_c	0.2564	0.3623	0.3691	0.3693

Each of the three allele age estimators (Relate, GEVA, t_c) were run on neutral simulations of the simple constant population size model and the out-of-Africa expansion population size model for both the whole exome sequences (WES) and the whole genome sequences (WGS). For each comparison between estimator and true value, the Pearson's correlation coefficient was calculated.

276

277 To investigate whether estimators of mutational age are robust to more complicated
278 demography models, a more complex out-of-Africa model was simulated (Gutenkunst et al.
279 2009). Again, ages of all polymorphic mutations were estimated using the three estimators and
280 compared with true values. A similar pattern of estimates from the simple model emerged with
281 all three methods underestimating old mutations. Relate remained the most accurate with a
282 Spearman's correlation coefficient of 0.68 with GEVA and t_c with correlations of 0.45 and 0.37
283 respectively (Figure 2). Again, Relate and GEVA were better able to estimate the ages of the
284 mutations from WGS data with correlations of 0.80 for Relate (WES: 0.68) and 0.63 for GEVA
285 (WES: 0.45) (Figure S3). However, t_c performed similarly with Pearson's correlation
286 coefficients of 0.37 for both WES and WGS datasets with the smallest drop in performance at
287 only 0.05% (Table 1). GEVA had the largest drop in performance again at 28% with Relate
288 having a 15% drop in performance.

289

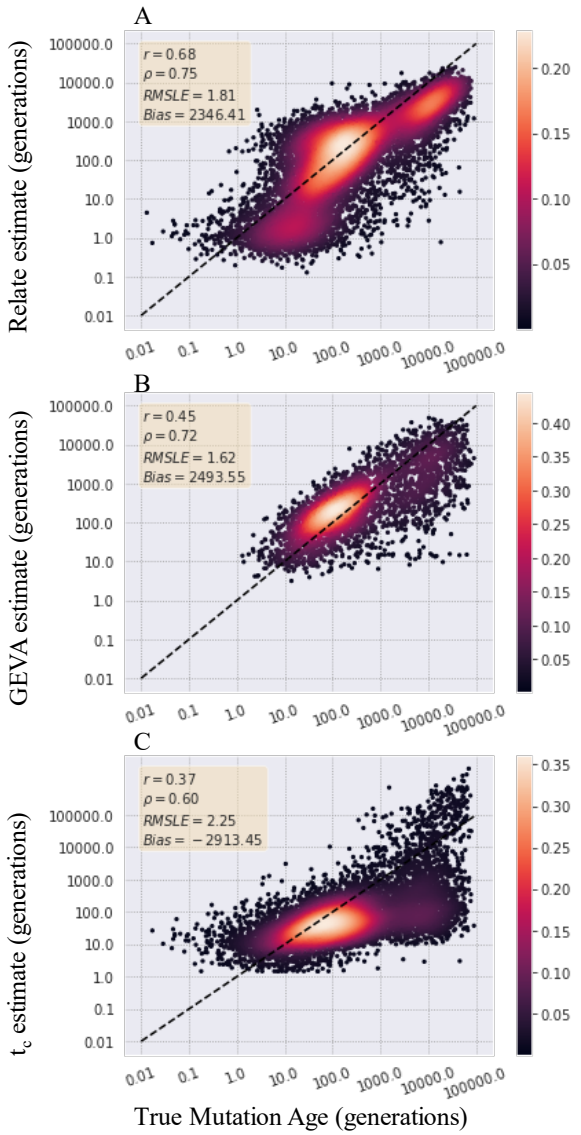

Figure 2. Estimator comparison on WES data for a complex model.

Estimated age of mutations for each of the three estimators Relate (A), GEVA (B), and time of coalescence (C) were compared to true age values for a complex demography simulation. 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. WGS comparison found in Figure S3.

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 and then normalized by the average true age of the bin (Figure 3). While it was found that the RMSLE was much lower for rare variants without the normalization for true age (Figure S4), once normalized it was found that all methods did better at estimating the age of more common variants in comparison with rare variants (less than a 1% frequency) (Figure 3A). Similar trends in error rates were found in the complex model (Figure S5). 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 3B).

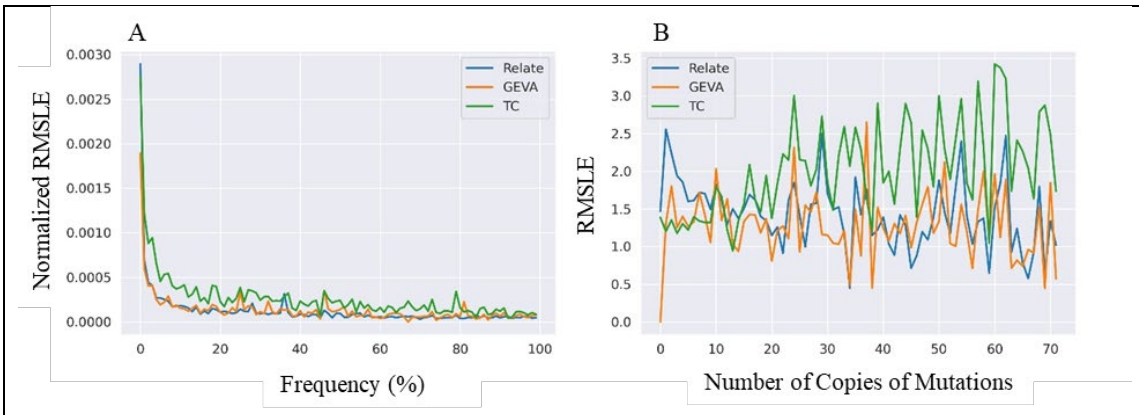

Figure 3. Error in estimates across spectrum of frequency values for the simple model.

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.

Further testing the robustness of allele age estimation on exome data, the estimator Relate was used to estimate the age of mutations in a sample size of 100 genomes, 1000 genomes, and 10000 genomes. Relate was used as it was the most tractable for the larger sample size of 10000 genomes, however estimation of mutations in a sample of 100000 genomes was also attempted but required over 15 TB of storage and high memory requirements making it not easily tractable even with dedicated computing resources. With increasing sample size, the Pearson coefficient between estimated age of mutations and true age increased (Figure 4, Figure S6). This is expected as Relate depends on the surrounding mutations around a focal site to estimate the age of the mutation of interest, and as the sample size increases as does the number of mutations present.

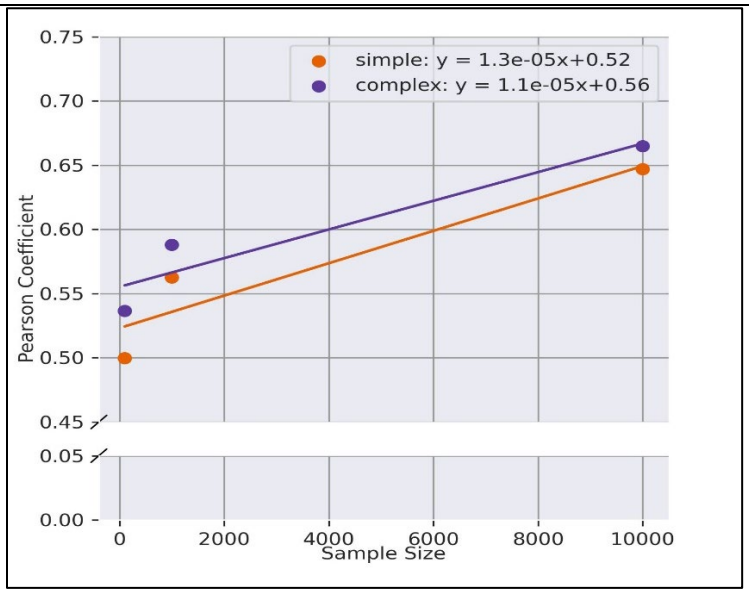

Figure 4. Correlation between true and estimated allele age increases with sample size. Samples of 100, 1000, and 10000 genomes are sampled from simulations. The true age of mutations is compared with the estimated age of mutations estimated with Relate for each of the three sampled set of mutations. A regression line is fit for both the simple (orange) and complex (purple) simulation estimates.

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, realtive to its performance with no selection, for both the simple and complex models (Table 1, Figure 5). Relate had an 18.6% decrease in accuracy for the simple model with the WES dataset in comparison with the WGS and a 26.7% decrease in accuracy with the WES dataset for the complex model. This is on average a 22% decrease in accuracy, higher than the 16% average decrease in accuracy on mutations under a simulation with no selection.

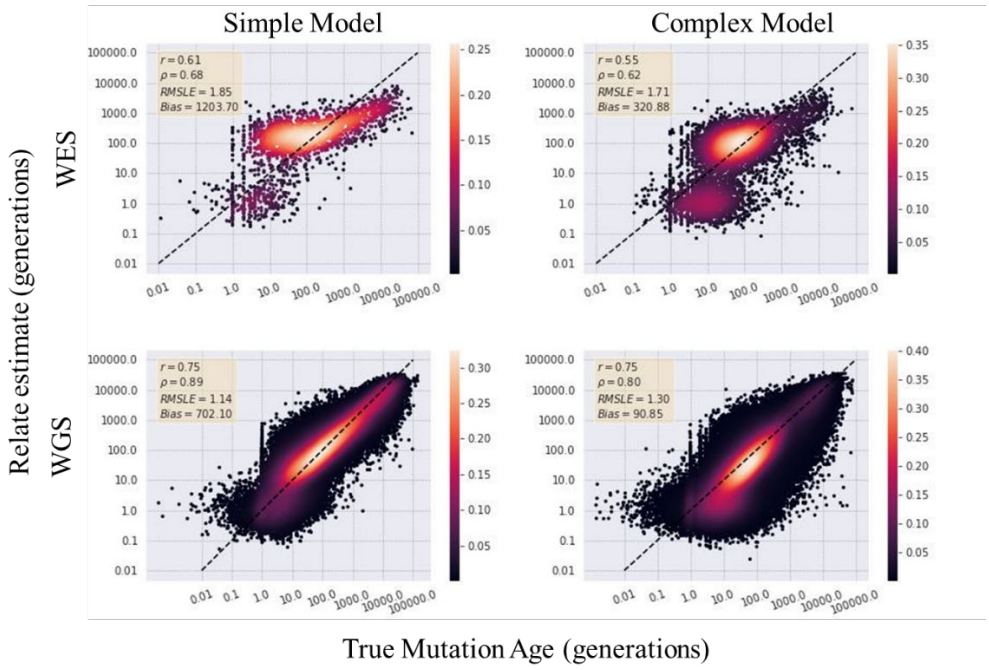

Figure 5. Relate allele age estimates compared to true values under background selection. The allele age of mutations from simulations of sampled genomes under a model which incorporates background selection are estimated using Relate. Upper panels have estimates from whole-exome datasets while lower panels contain estimates from the entire chromosome of mutations.

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 causative mutations for disease (Landrum et al. 2018; Karczewski et al. 2020), and in doing so, it becomes important to not only understand the functional impacts of the mutation but also the demographic history. Currently, there are many methods to characterize a mutation (McLendon et al. 2008; Eilbeck et al. 2017; Landrum et al. 2018; Biddanda et al. 2020), with one subset being the estimation of the

335 time of the mutation entered the population (Slatkin and Rannala 2000; Platt et al. 2019; Speidel
336 et al. 2019; Albers and McVean 2020). These methods are modeled and tested with whole
337 genome sequence data. However, since WES data remains the more cost-efficient option making
338 it more likely to be used in routine clinical diagnostics (Goh and Choi 2012; Salfati et al. 2019),
339 a comparison of several methods of allele age estimation to ascertain whether the methods can
340 accurately be applied to whole exome sequence data is required.

341 Of the three estimators, all had the highest accuracy in identifying the ages of mutations
342 from WGS data for both the simple and complex demography models with Relate performing the
343 best across all estimators. For WES data, we find that of all three methods Relate again performs
344 best with the highest correlation and lowest bias with a loss of between 15 to 17 percent in
345 comparison of the estimates on the WGS data. However, all three methods slightly underestimate
346 the oldest of alleles with t_c underestimating them to the highest degree with the RMSLE tripling
347 between 100 generations to 10000 generations. t_c greatly underestimates the ages of very old
348 variants indicating that the modified estimator, that attempts to account for when a haplotype
349 ends in non-coding regions, does not totally mitigate the issue of haplotype tracks ending in
350 intergenic or intronic regions. As a trend, each method slightly overestimated the youngest of
351 alleles as the bias values flip in signs from the youngest (less than 100 generations) to the oldest
352 (greater than 100 generations). This is expected, as coalescent events that happen further back in
353 time allow for more mutations to accumulate along that branch. Thus, in estimating allele age
354 from a dataset missing some substantial proportion of the mutations means that the estimators are
355 observing fewer mutations on a branch and estimating it shorter than its true length. The
356 overestimation of the youngest variants and underestimation of the oldest variants has been
357 previously identified in comparison of whole genome sequence data estimates from both Relate

358 and GEVA (Brandt et al. 2022; Ragsdale and Thornton 2023). In the comparison examining the
359 robustness of the methods on simulations expanding beyond just a neutral model to also include
360 deleterious mutations, Relate had a larger decrease in accuracy in the estimates between exome
361 and genome data in comparison with the neutral model (simple: 19%, complex: 27%) (Figure 5).
362 All three estimators assume an infinite-site model where only a single mutation is able to
363 arise at each site, however in nature this is not always the case (Karlin and McGregor 1967;
364 Kimura 1969). For simulation data, this is not a concern as sites where a site had more than one
365 mutation occur can be excluded or not allowed to occur. However, for empirical data, repeated
366 mutations may occur especially in regions of the genome where the mutation rate is particularly
367 high such as in CpG sites (Lek et al. 2016). If the infinite-site model is violated with repeated
368 mutation at a given loci with multiple common ancestral haplotypes, then haplotypes with
369 differing genealogical histories would then be pooled together artificially inflating the estimated
370 ages. Relate deals with this issue by excluding sites where mutations cannot be mapped onto the
371 local genealogy (Speidel et al. 2019) while t_c takes a composite of the estimated age of each copy
372 of allele allowing for repeated mutations at a single locus not to skew the estimated ages too high
373 (Platt et al. 2019). However, it is likely that the majority of human variant have arisen by way of
374 single mutations, with violations of the infinite-site model by multiple mutations constituting a
375 very small proportion of sites, given the few known examples (McVean et al. 2002; Blumenfeld
376 and Patnaik 2004; Tishkoff et al. 2007).
377 While common variants have long been identified as potential pools of variants affecting
378 human phenotypes (Burton et al. 2007; Claussnitzer et al. 2020), rare variants are often less well
379 understood and harder to be studied and characterized. Investigation of rare alleles using
380 methods such as mutation age estimators allows for these mutations to better characterized in a

population. This is particularly important because these mutations much less likely to be found in published large population data such as 1KGP (Auton et al. 2015) leading to very few opportunities to learn more about their history except through individualized studies. When normalized for the true age, rare variants are less able to be accurately aged, however since many rare variants are extremely young the overall error in estimates is small for these sites especially in comparison to common variants which may be extremely old and have a wide range of variation in estimated values for alleles of the same frequency.

This analysis highlights both the potential uses 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). While these estimators, in particular Relate, demonstrate reasonable accuracy in estimates based on whole exome data in comparison with whole genome estimates, users of these methods must also acknowledge the biases that appear in estimating the oldest and youngest of the mutations. However, utilization of these tools is still able to offer valuable insights into the history of potentially causative alleles to human disease. As the number of sequenced exomes increase as the field of personalized genomics becomes more widespread, these methods can play a vital role in advancing our understanding of the genetic basis of disease, enabling more precise diagnostics.

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.

Brandt DYC, Wei X, Deng Y, Vaughn AH, Nielsen R. 2022. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. Genetics. 221(1):iyac044. doi:10.1093/genetics/iyac044.

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.

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

459 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
460 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience.
461 10(2):giab008. doi:10.1093/gigascience/giab008.

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

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

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

470 Grzymski JJ, Elhanan G, Morales Rosado JA, Smith E, Schlauch KA, Read R, Rowan C,
471 Slotnick N, Dabe S, Metcalf WJ, et al. 2020. Population genetic screening efficiently identifies
472 carriers of autosomal dominant diseases. Nat Med. 26(8):1235–1239. doi:10.1038/s41591-020-
473 0982-5.

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

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

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

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

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

490 Karlin S, McGregor J. 1967. The number of mutant forms maintained in a population. In:
491 Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol.
492 4.

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

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

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

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

504 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle
505 M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature.
506 409(6822):860–921. doi:10.1038/35057062.

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

510 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH,
511 Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in
512 60,706 humans. Nature. 536(7616):285–291. doi:10.1038/nature19057.

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

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

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

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

526 Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI.
527 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids
528 Research. 48(D1):D845–D855. doi:10.1093/nar/gkz1021.

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

532 Ragsdale AP, Thornton KR. 2023. Multiple Sources of Uncertainty Confound Inference of
533 Historical Human Generation Times. Molecular Biology and Evolution. 40(8):msad160.
534 doi:10.1093/molbev/msad160.

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

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

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

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

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

549 Takahata N. 1993. Allelic genealogy and human evolution. Molecular Biology and Evolution.
550 10(1):2–22. doi:10.1093/oxfordjournals.molbev.a039995.

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

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

556 Tian X, Browning BL, Browning SR. 2019. Estimating the Genome-wide Mutation Rate with
557 Three-Way Identity by Descent. The American Journal of Human Genetics. 105(5):883–893.
558 doi:10.1016/j.ajhg.2019.09.012.

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

562 Tremblay M, Vézina H. 2000. New Estimates of Intergenerational Time Intervals for the
563 Calculation of Age and Origins of Mutations. The American Journal of Human Genetics.
564 66(2):651–658. doi:10.1086/302770.

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

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

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

574