mapDamage2.0 fast approximate Bayesian estimates of ancient DNA damage parameters

Hákon Jónsson*, Aurélien Ginolhac, Mikkel Schubert, Philip Johnson and Ludovic Orlando

April 23, 2013

Contents

1 Base specific damage probabilities **2**

2 The mutation model **3**

3 mapDamage2.0 in the field of aDNA **3**
3.1 Comparison to other tools 6

4 Rescaling of base quality scores **16**

5 Usage documentation **18**

List of Figures

S1 mapDamage2.0 and the Briggs-Johnson model run time comparison 7
S2 mapDamage2.0 and the Briggs-Johnson model parameter estimate comparison . . . 8
S3 mapDamage2.0 and the Briggs-Johnson model parameter estimate comparison . . . 9
S4 Sample E520, posterior predictive intervals 10
S5 Sample E521, posterior predictive intervals 11
S6 Sample E522, posterior predictive intervals 12
S7 Sample E523, posterior predictive intervals 13
S8 Sample E524, posterior predictive intervals 14
S9 Sample E525, posterior predictive intervals 15
S10 An example variant that was filtered out during the rescaling 17

List of Tables

S1 Parameter posterior means for the collection of aDNA datasets 4
S2 Parameter posterior means for the dataset from Schuenemann et al., 2011 5
S3 Rank correlation between time and parameter posterior means 6
S4 SNP calling for the ancient Aboriginal Australian (Rasmussen et al., 2011) 16

*to whom correspondence should be addressed

S5 Type of variants filtered out in the rescaling 18

1 Base specific damage probabilities

Nucleotide misincorporations observed at position i within a read are governed by four key parameters the average length of overhangs (λ); nick frequency (ν), and cytosine deamination rates at both double stranded regions (δ_{d}) and overhangs (δ_{s}). For the Markov chain (figure 1) we use the following order of states:

$$\mathrm{C_{start}},\ \text{C to T},\ \text{Single s.},\ \text{Double s.},\ \mathrm{T_{end}}\ \text{and}\ \mathrm{C_{end}}$$

with resulting transition probability matrix

$$P = \begin{pmatrix} 0 & \nu_i & 0 & 0 & 0 & 1-\nu_i \\ 0 & 0 & \lambda_i & 1-\lambda_i & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta_s & 1-\delta_s \\ 0 & 0 & 0 & 0 & \delta_d & 1-\delta_d \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since there are no self loops in the Markov chain, the support for the distribution of the states will be at the end states for any initial distribution after the 3rd power of the transition probability matrix. The initial distribution for the cytosines for this Markov chain will have this form

$$\kappa = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Then

$$(\kappa P^3)^T = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \nu_i\ (\lambda_i\,\delta_s + \delta_d\ (1-\lambda_i)) \\ 1-\nu_i\ (\lambda_i\,\delta_s + \delta_d\ (1-\lambda_i)) \end{pmatrix}$$

The base specific damage probabilities are derived as

$$\begin{aligned} p_{ct}(\delta_d, \delta_s, \lambda, \nu, i) &= \nu_i\ (\lambda_i\,\delta_s + \delta_d\ (1-\lambda_i)) \\ p_{ga}(\delta_d, \delta_s, \lambda, \nu, i) &= (1-\nu_i)\ (\lambda_i\,\delta_s + \delta_d\ (1-\lambda_i))\,. \end{aligned}$$

Here the ν_i and λ_i are defined in the following fashion

$$\lambda_i = \left(1 - \sum_{k=0}^{i} \binom{k+r-1}{k} (1-\lambda)^r \lambda^k\right)/2$$

ν_i ={Probability of C>T versus G>A substitution due to DNA damage at position i}

The position specific probabilities for the overhangs λ_i are defined in this fashion[1] to get similar values for the parameters as described by Briggs et al. (2007). In the case of double stranded protocol we estimate ν_i with a smooth function using generalized additive model using

$$z_i = \frac{s_{C,T,i}/d_{C,i}}{s_{C,T,i}/d_{C,i} + s_{G,A,i}/d_{G,i}} \tag{1}$$

as a dependent variable and position as the predictor. For the single stranded protocol, the model simplifies as we set $\nu_i = 1$ for all i. Non-informative priors are used for the parameters, and parameter estimation is carried out using MCMC (Hastings, 1970) with Gibbs sampling (Geman and Geman, 1984).

2 The mutation model

The mutation/sequencing error rate matrix Q_Θ for Θ is parametrized in a HKY (Hasegawa et al., 1985) setup

$$Q_\Theta(\mu, \rho) = \mu \cdot \begin{pmatrix} - & \rho \cdot \pi_\mathrm{c} & \pi_g & \rho \cdot \pi_t \\ \rho \cdot \pi_a & - & \rho \cdot \pi_g & \pi_t \\ \pi_a & \rho \cdot \pi_c & - & \rho \cdot \pi_t \\ \rho \cdot \pi_a & \pi_c & \rho \cdot \pi_g & - \end{pmatrix}.$$

$\pi_\mathrm{a}, \pi_\mathrm{c}, \pi_\mathrm{g}$ and π_t refer to the base frequencies of the reference genome used for aligning the reads, μ is the overall substitution frequency and ρ is the transversion / transitition bias.

3 mapDamage2.0 in the field of aDNA

We applied mapDamage2.0 on a collection of aDNA sequence datasets from a range of time periods, source materials and environments (supplementary table S1). When BAM alignment files (Li et al., 2009) were unavailable, mapping was carried out using raw sequences and corresponding reference genome. The reads were first trimmed using AdapterRemoval v1.2 (Lindgreen, 2012) to remove adapter sequences (allowing a mismatch rate of 1/3), trimming low-quality bases (Ns and bases with a Phred score of 2) at read termini. Following (Green et al., 2010 and Reich et al., 2010), paired-ended reads that overlapped with at least 11bp were collapsed into a single sequence. Non-overlapping paired-ended reads and reads that were shorter than 25bp after trimming were discarded. The trimmed reads were mapped to a reference using BWA (Li and Durbin, 2009), with mapping carried out without the use of a seed-region (Schubert et al., 2012). Following mapping PCR duplicates were filtered using MarkDuplicates from the Picard tool-kit[2] and using a script kindly provided by Martin Kircher to filter collapsed reads (FilterUniqueBAM.py[3]). Finally, the BAMs were realigned using the GATK IndelRealinger tool (McKenna et al., 2010) to improve the local aligment around indels.

After alignment, mapDamage2.0 was used with default parameters with two exceptions: only the forward end of the sequences was used and 20 bases from that end to be more comparable across a wide range of datasets. In particular, some sequence data sets correspond to single end reads while others were paired-end reads that showed significant overlap and could be collapsed. As the latter

[1] For clarity we use site specific probabilities for forward-only overhangs.

[2] `http://picard.sf.net`

[3] `https://bioinf.eva.mpg.de/fastqProcessing/` for a SAM compatible version of this script

(paired-ends) showed a much higher quality than the former (single-ends) and provided full-length sequence information from DNA inserts (in contrast to single-end sequencing which might not reach the insert ends) we decided to consider only the first 20 base pairs from the 5' end. With this conservative approach, G→A misincorporations rates at sequencing ends will not be under-estimated due to the presence of inserts only partially sequenced. The following bash command shows the options used for the dataset.

```
mapDamage --length 100 --seq-length 20 --forward -i alignment.bam -r reference.fa
```

For the datasets from Bon et al. (2012) and Miller et al. (2012) we opted for taking the reverse end since the molecular protocol used for preparing and amplifying libraries resulted in the typical frequency of G→A substitutions at sequencing ends but not at sequencing starts. To be explicit about a potential source of bias, using the reverse end of the single ends in Bon et al. (2012) could downwardly bias the damage estimates. To account for this peculiarity a minor change was applied to the preceding bash command.

```
mapDamage --length 100 --seq-length 20 --reverse -i alignment.bam -r reference.fa
```

To avoid problems with low quality sequences from the Miller et al. (2012) dataset we discarded merged reads if one of the paired end was of low quality. More specifically, we removed collapsed reads using a minimum base quality threshold (Phred based) for the last 3' nucleotide: 40 for read length $\leq$ 101 bp and 38 for read length $\geq$ 101 bp. Further, as the reference genome for the polar bear (*Ursus maritimus*) is still at draft stage (Li et al., 2011), we restricted the analysis to contigs over 10,000 bases.

The distribution of cytosine deamination rates between double- and single-strand regions was estimated by drawing (50,000 times) from the posterior conditional distributions for δ_d and δ_s then calculating the ratio. The mean of the distribution of calculated ratios is reported in supplementary table S1 as column δ_d/δ_s.

Supplementary table S1: Parameter posterior means (δ_d, δ_s, λ, ρ, μ and δ_d/δ_s) for various datasets. Radiocarbon dating or indirect age estimate are in years from the present day (BP). The column Spe. indicates which species was used for the alignment with the following abbreviations: HS *Homo sapiens*, CC *Crocuta crocuta spelaea*, MC *Mammuthus columbi*, DG *Dinornis giganteus* and UM *Ursus maritimus*. Ref. column is the reference used for mapping: P is the *Yersinia pestis* pPCP1 plasmid, M and N stand for the mitochondria and nuclear genomes of the species considered. Samples that were realigned are marked by an asterisk in the id column.

Id	δ_d	δ_s	λ	ρ	μ	δ_d/δ_s	Spe.	Ref.	Art.	Nr. of seq.	Radiocarbon date
BRA1	0.024	0.519	0.327	0.063	0.009	0.050	HS	M	22	720	7,000
BRA2	0.020	0.521	0.301	0.353	0.009	0.042	HS	M	22	250	7,000
E520*	0.031	0.453	0.193	0.049	0.007	0.070	HS	M	24	5,261	672
E521*	0.033	0.597	0.289	0.024	0.013	0.055	HS	M	24	25,736	672
E522*	0.028	0.507	0.237	0.043	0.013	0.055	HS	M	24	21,039	672
E523*	0.018	0.454	0.244	0.033	0.008	0.040	HS	M	24	4,893	672
E524*	0.017	0.486	0.283	0.025	0.009	0.036	HS	M	24	17,381	672
E525*	0.019	0.725	0.258	0.031	0.013	0.026	HS	M	24	6,048	672
SLVi33.16	0.031	1.000	0.235	0.102	0.054	0.031	HS	M	8	113,700	38,310
SLVi33.25	0.023	0.997	0.276	0.138	0.050	0.023	HS	M	8	9,229	/
SLVi33.26	0.027	0.998	0.239	0.088	0.050	0.027	HS	M	8	27,752	/

SLFeld1	0.021	0.821	0.392	0.067	0.054	0.027	HS	M	8	251	39,900
SLMez1	0.019	0.856	0.352	0.121	0.049	0.023	HS	M	8	5,093	65,000
SLSid1253	0.034	0.976	0.190	0.138	0.047	0.035	HS	M	8	277	3,8790
Ajv52	0.118	0.777	0.255	0.137	0.014	0.152	HS	M	25	2,848	4,360-4,005
Ajv70	0.082	0.587	0.279	0.050	0.019	0.143	HS	M	25	2,391	4,360-4,005
Gok4	0.036	0.570	0.205	0.327	0.011	0.065	HS	M	25	816	4,971-4,871
Ire8	0.114	0.636	0.267	0.036	0.026	0.181	HS	M	25	2,107	4,280-3,850
WB16A	0.025	0.813	0.251	0.631	0.053	0.032	HS	M	11	10,332	727-712
WB18	0.055	0.768	0.250	2.766	0.014	0.074	HS	M	11	11,880	727-712
WB1	0.047	0.631	0.204	1.388	0.009	0.075	HS	M	11	12,806	727-712
WB21	0.059	0.852	0.219	1.162	0.028	0.070	HS	M	11	39,860	727-712
Ajv52	0.077	0.688	0.247	0.496	0.019	0.111	HS	N	25	859,548	4,360-4,005
Ajv70	0.053	0.651	0.344	0.484	0.017	0.082	HS	N	25	1,307,016	4,360-4,005
Gok4	0.021	0.630	0.363	0.575	0.018	0.033	HS	N	25	1,058,571	4,971-4,871
Ire8	0.043	0.635	0.270	0.639	0.019	0.067	HS	N	25	562,685	4,280-3,850
BRA1	0.009	0.230	0.292	0.565	0.009	0.038	HS	N	22	728,160	7,000
BRA2	0.007	0.230	0.236	0.502	0.010	0.029	HS	N	22	364,258	7,000
SLVi33.16	0.021	1.000	0.302	0.679	0.016	0.021	HS	N	8	32,784,524	38,310
SLVi33.25	0.018	1.000	0.370	0.718	0.016	0.018	HS	N	8	25,813,140	/
SLVi33.26	0.023	1.000	0.336	0.646	0.013	0.023	HS	N	8	27,207,294	/
SLFeld1	0.025	0.995	0.278	0.663	0.029	0.025	HS	N	8	43,863	39,900
SLMez1	0.019	1.000	0.299	0.644	0.016	0.019	HS	N	8	1,261,335	65,000
SLSid1253	0.013	0.999	0.263	0.678	0.021	0.013	HS	N	8	48,537	38,790
E520*	0.011	0.804	0.503	0.152	0.002	0.014	HS	P	24	2,983	672
E521*	0.010	0.625	0.285	0.169	0.005	0.017	HS	P	24	3,804	672
E522*	0.015	0.441	0.270	48.326	0.000	0.034	HS	P	24	2,532	672
E523*	0.004	0.533	0.333	0.158	0.002	0.007	HS	P	24	1,862	672
E524*	0.010	0.769	0.423	0.127	0.003	0.014	HS	P	24	3,016	672
E525*	0.010	0.813	0.429	1.689	0.001	0.012	HS	P	24	1,240	672
CC8*	0.003	0.997	0.318	0.968	0.002	0.003	CC	M	2	11,266	22,700-22,480
CC9*	0.005	0.997	0.316	1.362	0.005	0.005	CC	M	2	22,563	22,700-22,480
Hunt*	0.015	0.063	0.174	0.095	0.021	0.364	MC	M	5	7,066	11,220
Moa7	0.016	0.213	0.297	0.135	0.011	0.091	DG	M	1	878	1,175
Moa9	0.027	0.281	0.260	0.074	0.011	0.099	DG	M	1	4,032	988
polar*	0.031	0.701	0.223	1.315	0.013	0.044	UM	M	19	15,472	110,000-130,000
polar*	0.022	0.739	0.253	1.147	0.015	0.030	UM	N	19	3,844,349	110,000-130,000

Supplementary table S2: Parameter posterior means and running times (in *min*) for running mapDamage2.0 on the ancient plague dataset from Schuenemann et al., 2011.

Id	Nr. of seq.	Time	δ_d	δ_s	λ	ρ	μ	δ_d/δ_s
E520	5,261	16	0.031	0.453	0.193	0.049	0.007	0.070
E521	25,736	15	0.033	0.597	0.289	0.024	0.013	0.055
E522	21,039	16	0.028	0.507	0.237	0.043	0.013	0.055
E523	4,893	14	0.018	0.454	0.244	0.033	0.008	0.040
E524	17,381	17	0.017	0.486	0.283	0.025	0.009	0.036
E525	6,048	17	0.019	0.725	0.258	0.031	0.013	0.026

Supplementary table S3: Rank correlation between age from radiocarbon dating and parameter posterior means. The asterisk is on values in the p-val column which are lower than 0.05. No correction for multiple testing was applied.

	p-val.	Spearman ρ
δ_d	0.75908	0.04816
δ_s	0.00349*	0.43575
λ	0.89471	-0.02079
ρ	0.06141	0.28767
μ	0.00024*	0.53268

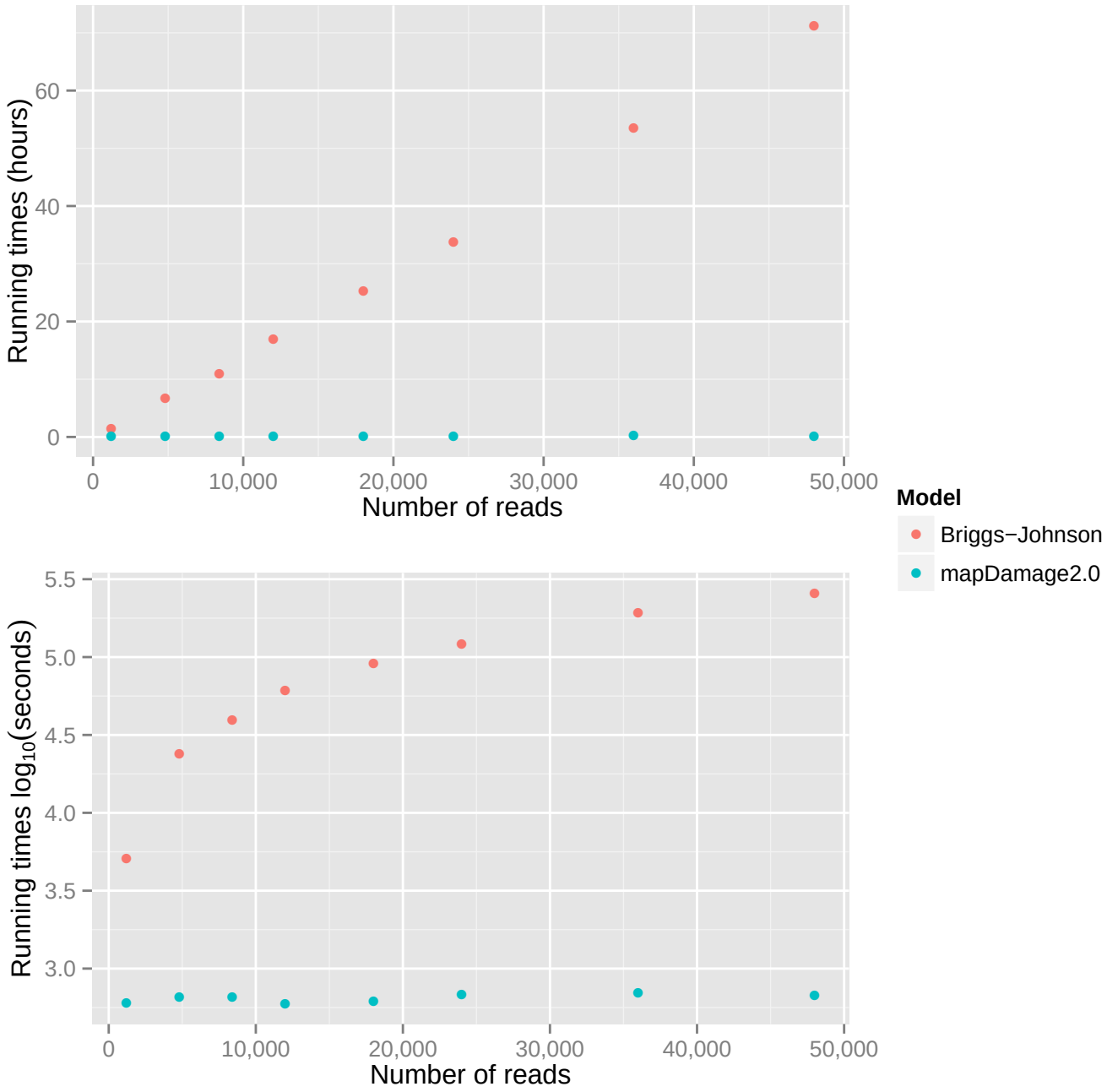
3.1 Comparison to other tools

The only implemented statistical model capable of inferring DNA damage parameters is the Briggs-Johnson model described in Briggs et al. (2007) to the best of our knowledge. We restricted our comparison to only a subset of the datasets presented in supplementary table S1 (namely, the dataset from Green et al., 2010 and Schuenemann et al., 2011) as the Briggs-Johnson program only works for the DNA library building procedure from Meyer and Kircher, 2010.

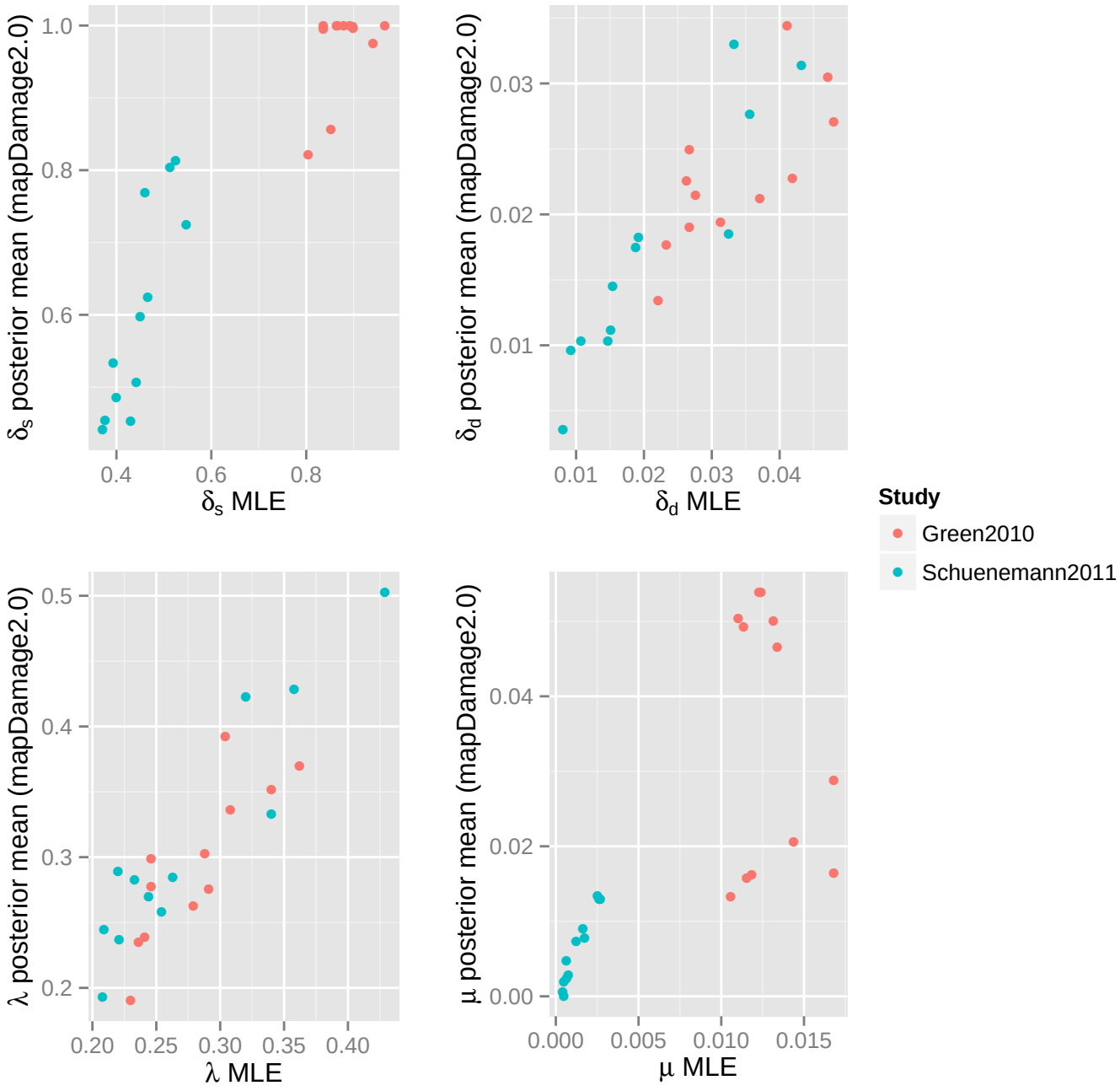
We first measured the difference in performance of the two methods by randomly selecting a subset of reads from SLVi33.16 (supplementary table S1) at each nuclear chromosome (X, Y and chr. 1 to chr. 22). Various sizes for subsets were considered (50, 200, 350, 500, 750, 1,000, 1,500 and 2,000 reads per chromosome). Given the extensive running time of the Briggs-Johnson model, only a sub-sample of total 12,000 reads (500 reads from each chromosome) was considered if the dataset contained more than 40,000 reads. Finally, mapDamage2.0 was restricted to the Jukes-Cantor mutation model (with option –jukes-cantor while the default is HKY) in order to match the Briggs-Johnson program. The same system setup was used for running both programs (AMD Opteron(TM) Processor 6276, 2.3 GHz).

The difference in performance for the methods is striking (see supplementary figure S1), with the summary model approach in mapDamage2.0 outcompeting the Briggs-Johnson model by several orders of magnitude in terms of running-time. We found that using the Briggs-Johnson model on the full sequence dataset from the Neandertal SLVi33.16 sample (32,784,524 hits in the original bam file) would be virtually impossible as the analysis would run over 5 years. However, the same analysis took no more than two hours with mapDamage2.0, including rescaling. This demonstrates that mapDamage2.0 is capable of analyzing large sequence datasets, such as those generated by high-throughput sequencing platforms.

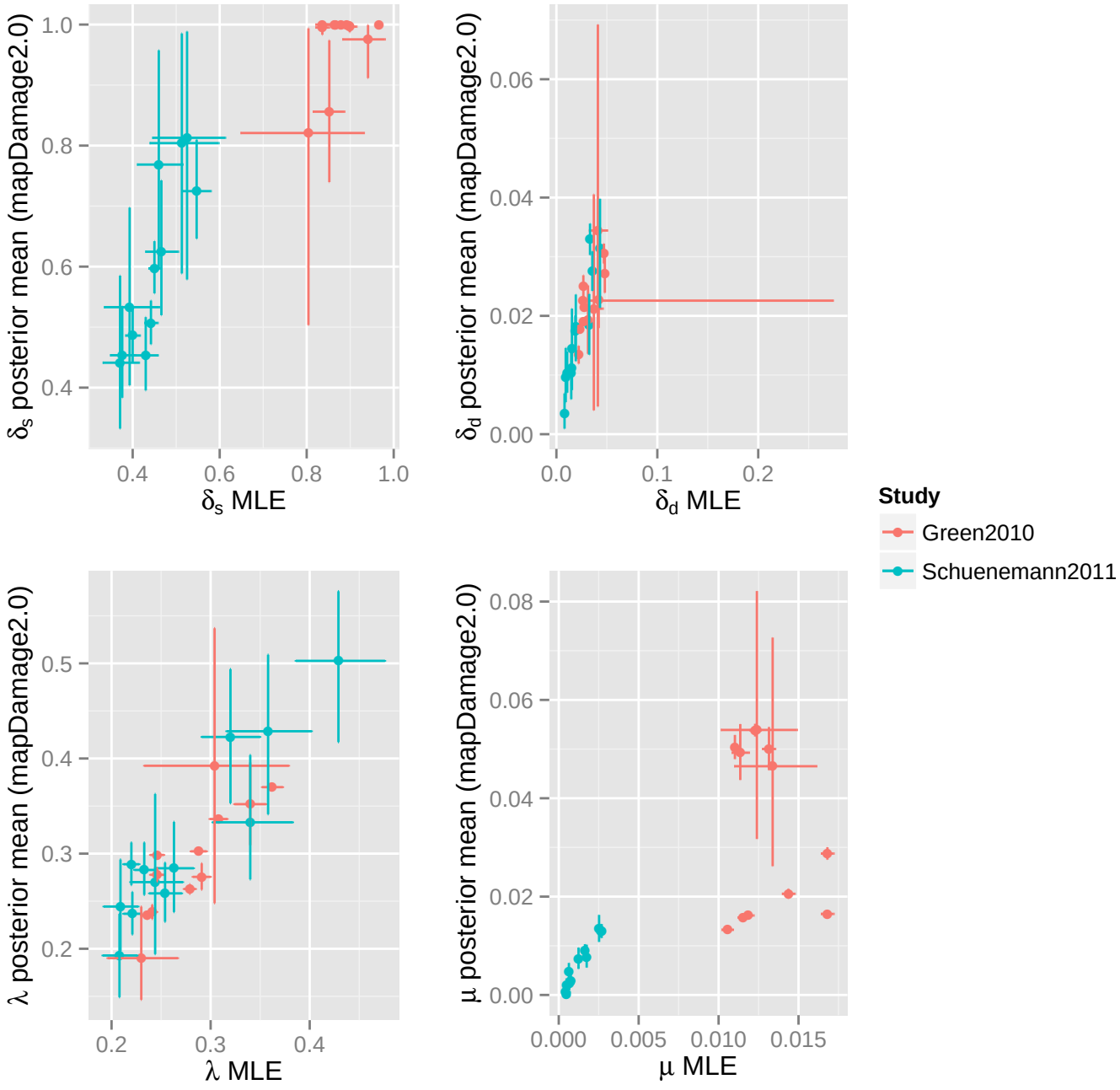
The parameter estimates for the different models are provided in supplementary figures S2 and S3 were found to be consistent, illustrating that the underlying simplifying assumptions of mapDamage2.0 provide reliable results. We note however, a significant departure in posterior estimates for μ, with mapDamage2.0 providing greater values than the Briggs-Johnson model. This is most likely to the fact that we restricted mapDamage2.0 to the 20 first bases (while the full read length was used with the Briggs-Johnson method) at sequencing starts. This region is more prone to misalignments *e.g.* indels are very difficult to align correctly close to the ends. Finally, the position specific nick frequency is estimated with a generalized additive model in mapDamage2.0 (see equation 1) consequently the nick frequency parameter in the Briggs-Johnson model is incomparable with our model.



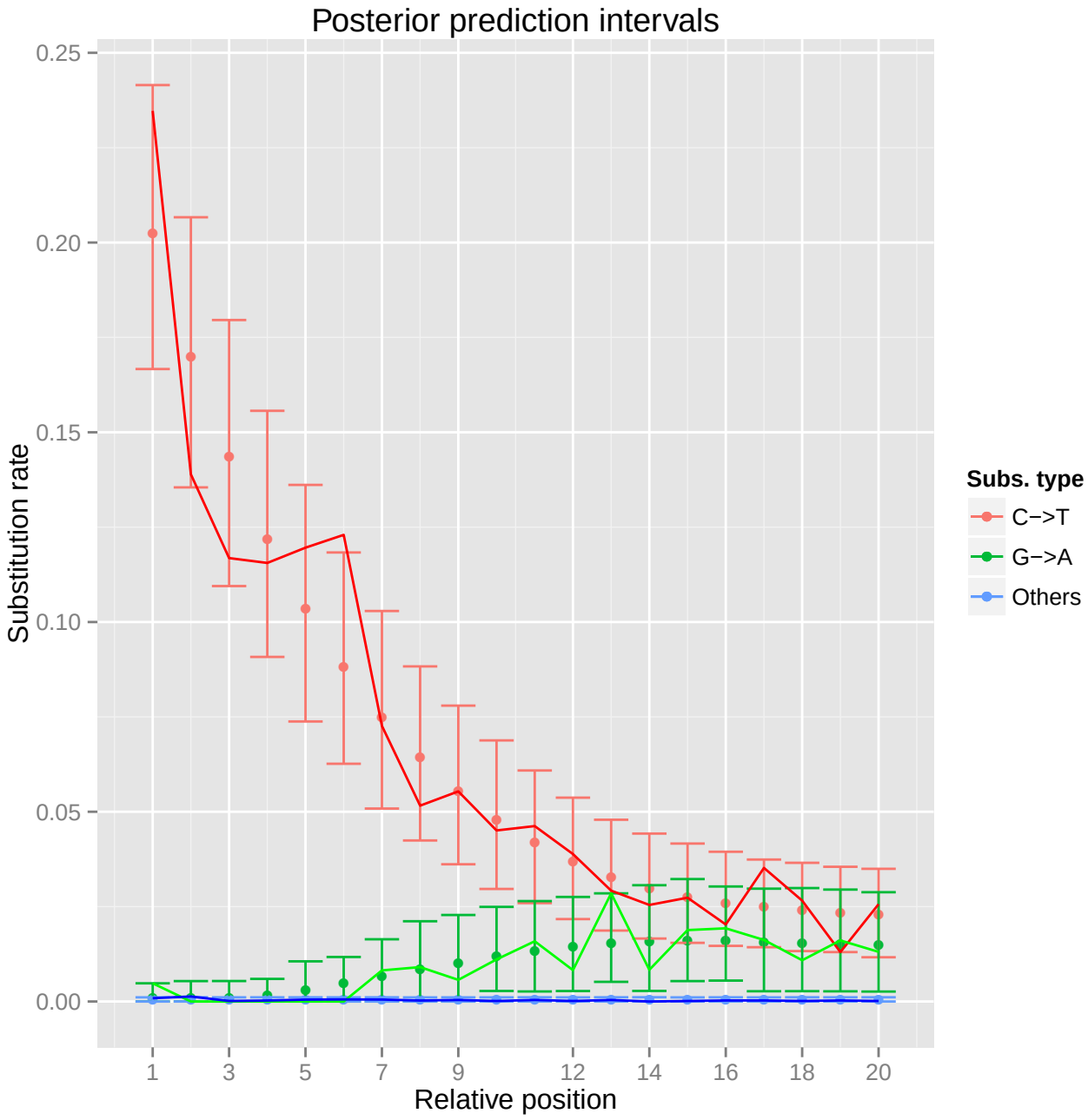
Supplementary figure S1: mapDamage2.0 and the Briggs-Johnson model (Briggs et al., 2007) run time comparison, using size varying subsets of sample SLVi33.16 from Green et al., 2010.



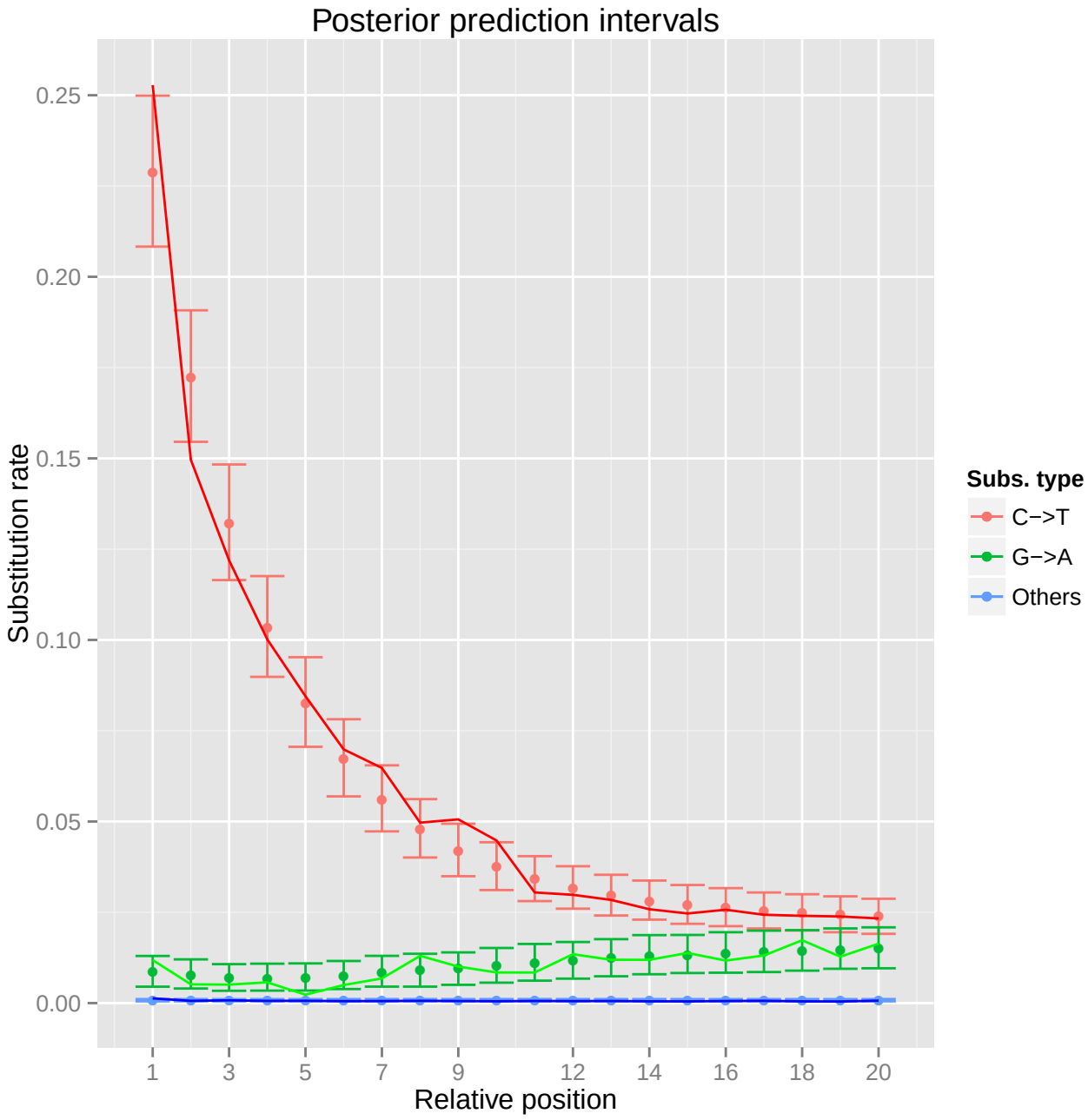
Supplementary figure S2: mapDamage2.0 and the Briggs-Johnson model (MLE) parameter estimate comparison, using downsampled subsets from Green et al., 2010 and Schuenemann et al., 2011.



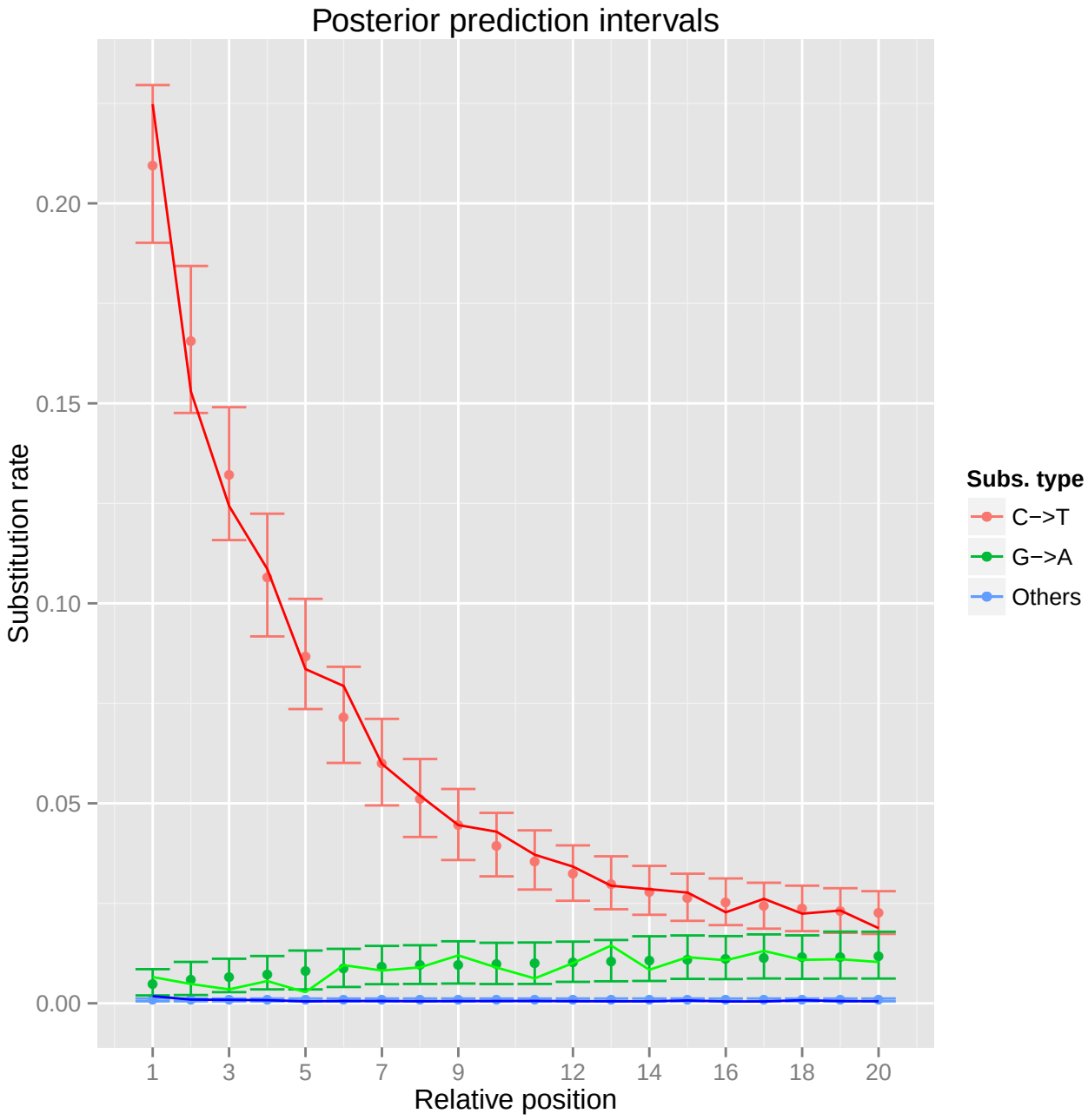
Supplementary figure S3: mapDamage2.0 and the Briggs-Johnson model (MLE) parameter estimate comparison, using downsampled subsets from Green et al., 2010 and Schuenemann et al., 2011. Horizontal error bars are the asymptotic χ^2 confidence intervals (95%) and the vertical error bars are the quantiles (2.5% and 97.5%) for the estimated posterior distribution.



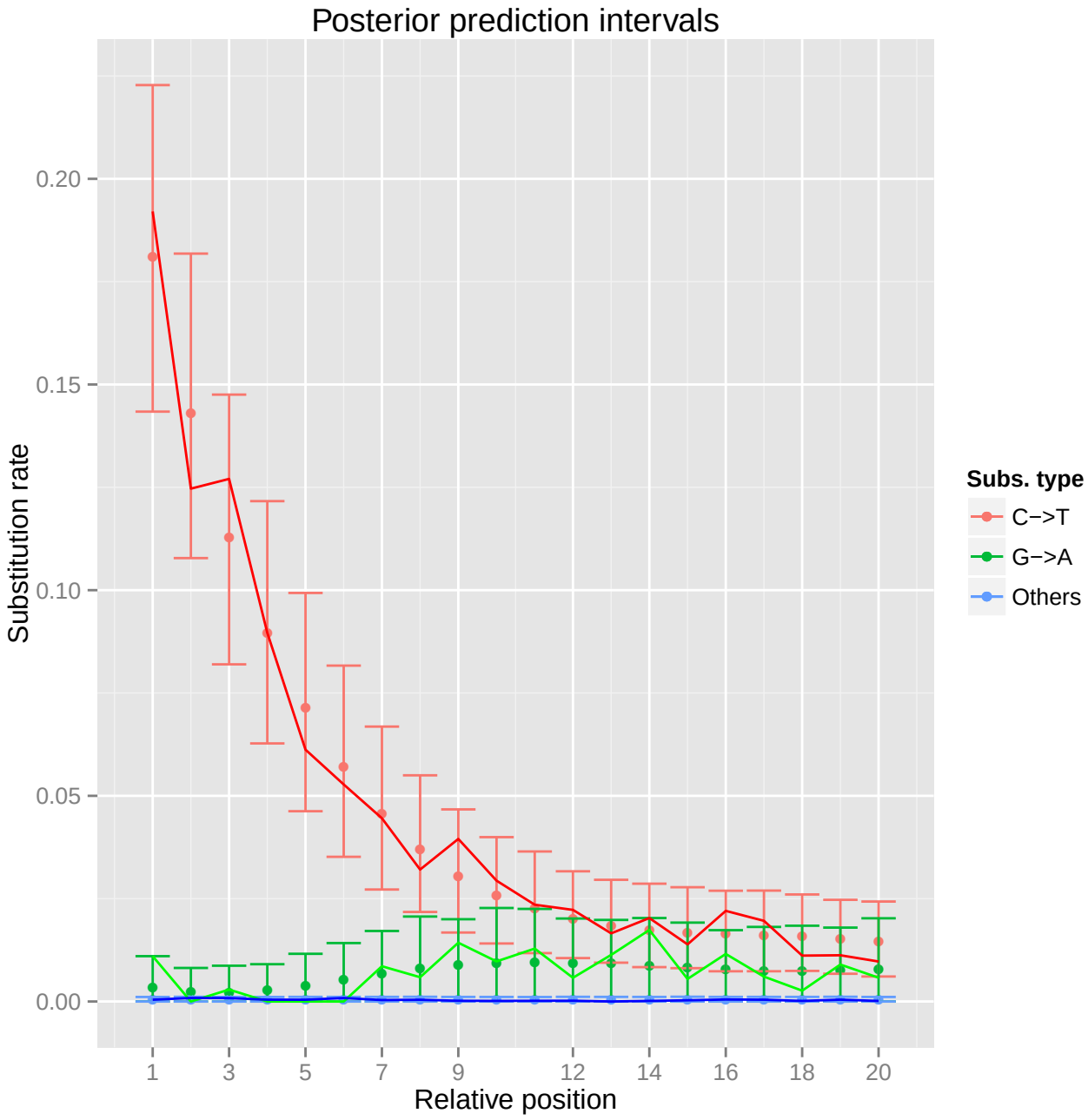
Supplementary figure S4: Sample E520, 95% posterior predictive intervals for the substitution frequencies and the solid line is the empirical frequency. The x-axis is the position from the 5' end of the template meanwhile the y-axis is the subsitution frequency. See Schuenemann et al. (2011) for more detailed information regarding the original data.



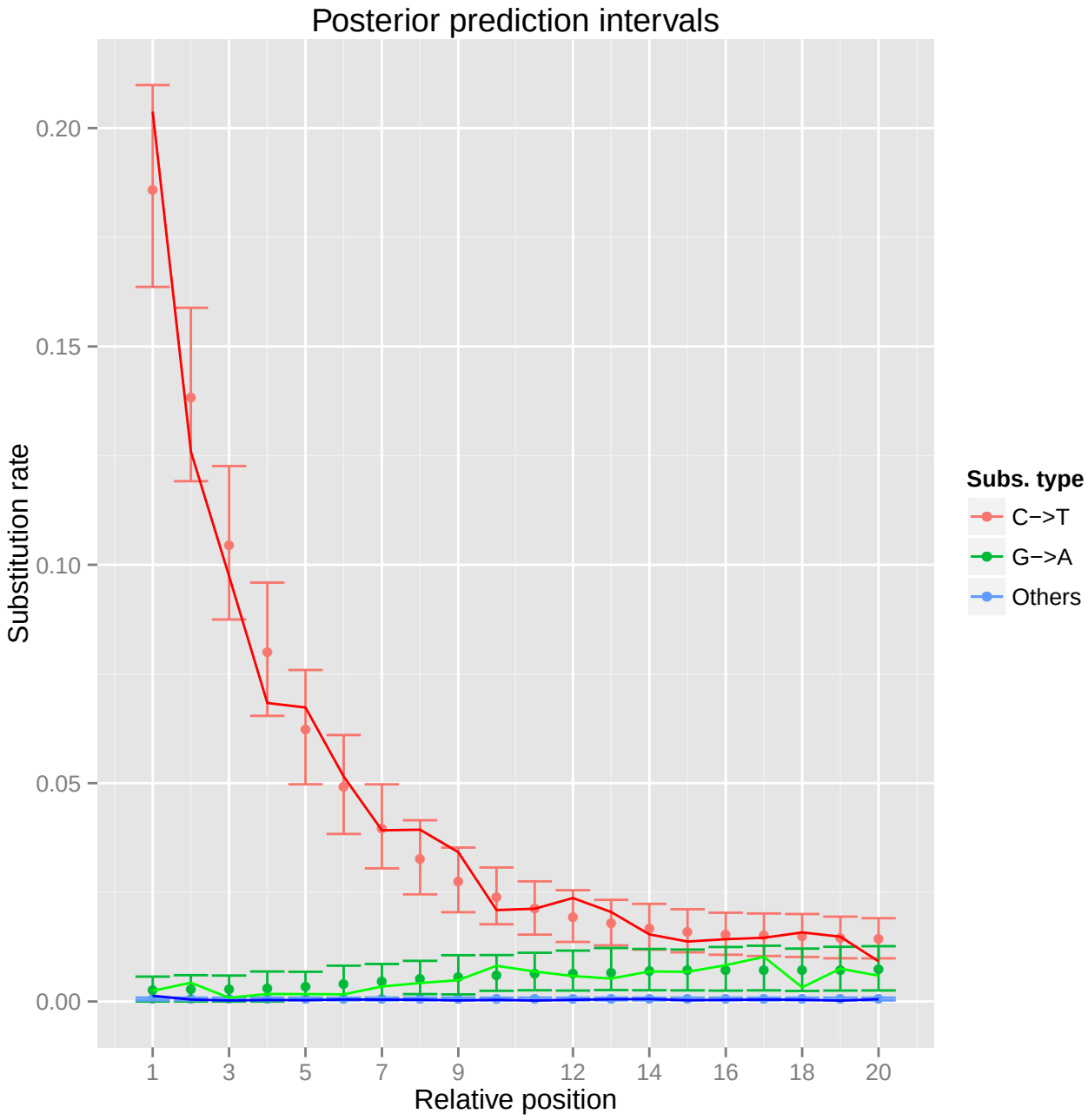
Supplementary figure S5: Sample E521, 95% posterior predictive intervals for the substitution frequencies and the solid line is the empirical frequency. The x-axis is the position from the 5' end of the template meanwhile the y-axis is the subsitution frequency. See Schuenemann et al. (2011) for more detailed information regarding the original data.



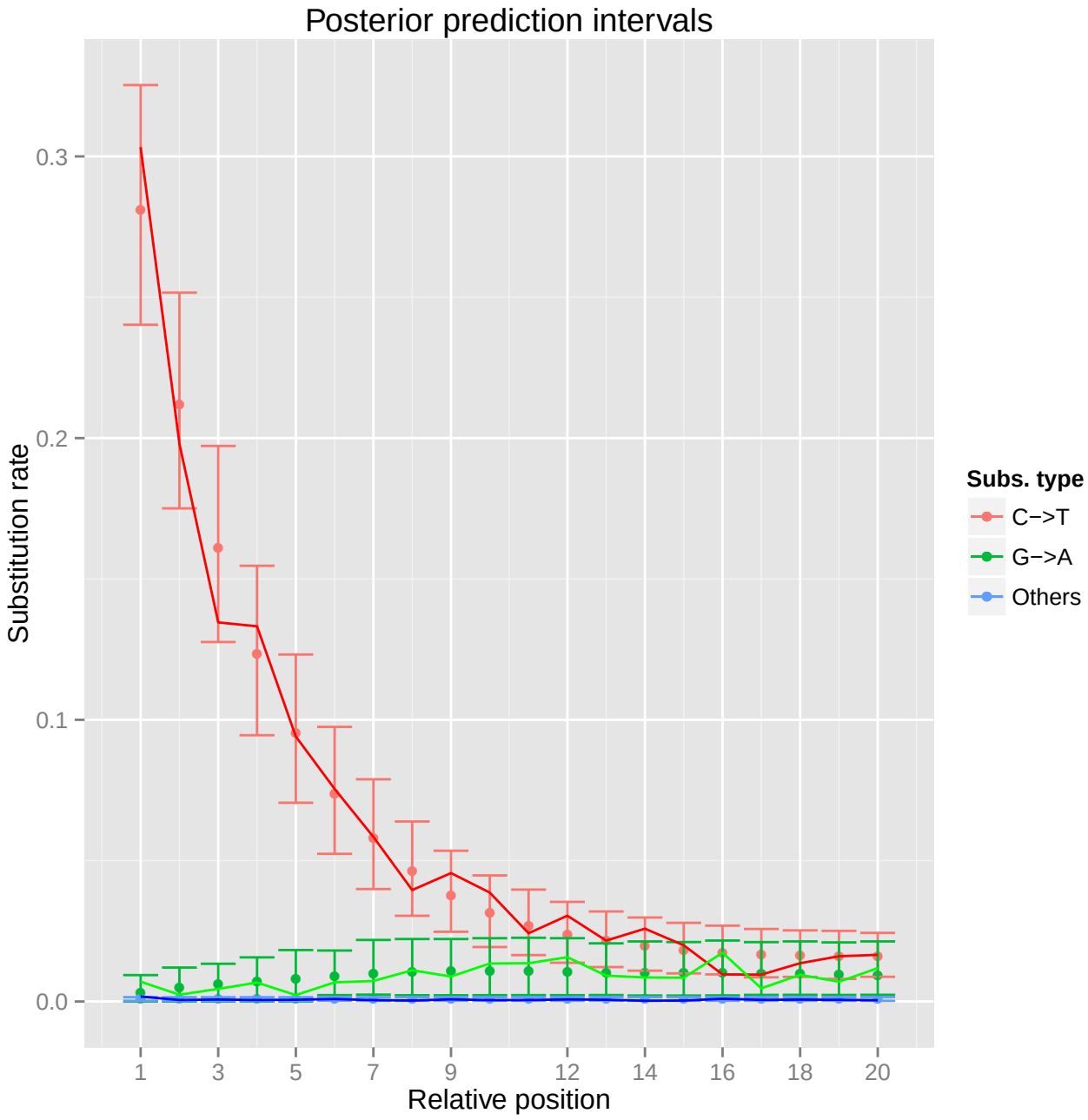
Supplementary figure S6: Sample E522, 95% posterior predictive intervals for the substitution frequencies and the solid line is the empirical frequency. The x-axis is the position from the 5' end of the template meanwhile the y-axis is the subsitution frequency. See Schuenemann et al. (2011) for more detailed information regarding the original data.



Supplementary figure S7: Sample E523, 95% posterior predictive intervals for the substitution frequencies and the solid line is the empirical frequency. The x-axis is the position from the 5' end of the template meanwhile the y-axis is the subsitution frequency. See Schuenemann et al. (2011) for more detailed information regarding the original data.



Supplementary figure S8: Sample E524, 95% posterior predictive intervals for the substitution frequencies and the solid line is the empirical frequency. The x-axis is the position from the 5' end of the template meanwhile the y-axis is the subsitution frequency. See Schuenemann et al. (2011) for more detailed information regarding the original data.



Supplementary figure S9: Sample E525, 95% posterior predictive intervals for the substitution frequencies and the solid line is the empirical frequency. The x-axis is the position from the 5' end of the template meanwhile the y-axis is the subsitution frequency. See Schuenemann et al. (2011) for more detailed information regarding the original data.

4 Rescaling of base quality scores

The Aboriginal aDNA sequence analysis (Rasmussen et al., 2011) was carried out using reads that were trimmed for the first and last three nucleotides prior mapping in order to reduce the effect of DNA damage related nucleotide misincorporations. Here we used a model based approach which aligns them untrimmed against the reference and then downscales likely damaged positions in the aligned reads.

Similar alignment pipeline as Rasmussen et al. (2011) was used for the untrimmed reads and mapDamage2.0 was used to rescale base qualities in the BAM file. Only the 20 first bases were considered from each end, the position dependent nick probability was assumed to be 1 at the 5' end and 0 at the 3' end, the following bash command shows the exact usage of mapDamage2.0.

```
mapDamage -i Aborigine.bam \
          -r hg19.fa       \
          -d AboOut        \
          --verbose        \
          --rescale        \
          --length=100     \
          --seq-length=20  \
          --fix-nicks
```

Fixing the nick probability in this fashion provided a decent fit for low damage datasets. Note the –rescale flag which outputs the rescaled BAM file to the output folder. For comparison the variant sites were filtered before and after rescaling according to the pipeline by Rasmussen et al. (2011) (supplementary table S4), kindly provided by Simon Rasmussen.

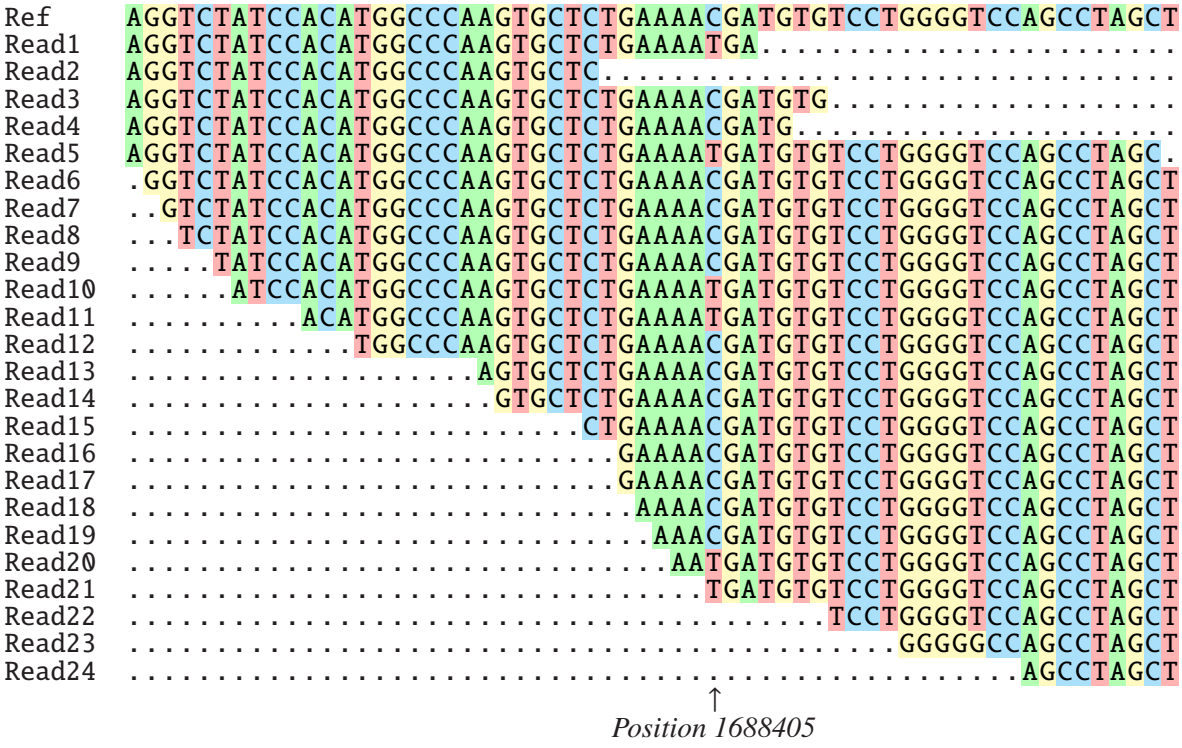
Supplementary table S4: SNP calling at chromosome 10 for the ancient Aboriginal Australian (Rasmussen et al., 2011).

Metric	Before rescaling	After rescaling
Total nr. of SNPs	23,455	22,025
SNP-db overlap	19,393 (82.7%)	18,253 (82.9%)
Not present in SNP-db	4,062	3,772
SNPs filtered out	-	1,586
SNPs kept	-	156

To explore in greater detail the nature of the variants that were removed in the rescaling process, we counted the reference and alternative alleles combinations from the VCF file (supplementary table S5). As expected from the *post-mortem* deamination process, the vast majority (98.8%) of the variants filtered corresponded to situations where an C (G) was found in the reference and a T (A) was observed in the sequence reads, significantly different ($p < 2.2 \cdot 10^{-16}$, Fisher's exact test) from the frequency in db-SNP (38.2%). One typical example is provided in supplementary figure S10 (at position 1,688,405 in chromosome 10), where the minor allele frequency (MAF) for this position is low (6/20) but still above the MAF frequency cutoff in the filtering pipeline described by Rasmussen et al., 2011 and therefore contributes to the final genotype calls. While DNA damage is likely a contributing factor (especially for Read20 and Read21 as both show C→T misincorporation at the sequencing starts) we must caution that the variant site could still correspond to a true polymorphism or a misincorporation due to a repeat structure in the genome, misalignment, sequencing errors and other sources of bias. Increasing the sequencing effort and/or using specific molecular tools limiting the impact of *post-*

mortem cytosine deamination (*e.g.* Briggs et al., 2009) could provide further tests for the validity of this supposedly variable site.

We counted the occurrences of C→T mismatches found at Cs located in a dinucleotide context (CpA, CpC, CpG and CpT), and the total of number of the same dinucleotides in the alignment against the reference. We found elevated frequency of C→T at CpG compared to other dinucleotide contexts (ca. 0.0098 vs 0.0029, respectively, ie. ≈3.4X difference). This could suggest that our approach is over-aggressive and corrects more often than it should for T variants found in a CpG context. Alternatively, this difference could originate from the observed biased distribution of CpG towards read starts (data not shown; the %CpG in the reference genome for the first 20 nucleotidic positions within reads was found to be 2X greater than for the last 20 nucleotidic positions). This suggests that the position specific CpG composition of the reads could account for >58% of rescaling performed at sequencing starts, leaving a minority of cases where over-aggressive rescaling is performed. A simple partition scheme, where the original BAM file is splitted into two or more sub data sets (eg. CpG Islands vs no CpG Islands; read classes of %CpG etc), could be further used to account for the specificities of any region of interest during rescaling.



Supplementary figure S10: Alignment region around the example variant site (position 1,688,405, chromosome 10 and build hg19) that was filtered out during the rescaling.

Supplementary table S5: Type of variants filtered in the ancient Aboriginal Australian rescaling (Rasmussen et al., 2011).

Variant type	Frequency
C→T	786
G→A	781
Others	19

5 Usage documentation

Let assume the user has a BAM file called seq.bam aligned against the reference in ref.fa, then the minimal options supplied to mapDamage would be the following[4]

```
mapDamage -i seq.bam -r ref.fa
```

The output will be in the folder result_seq, the directory name can be changed by the -d option. By default, the following files will be created in this folder.

```
3pGtoA_freq.txt
5pCtoT_freq.txt
dnacomp_genome.csv
dnacomp.txt
Fragmisincorporation_plot.pdf
Length_plot.pdf
lgdistribution.txt
misincorporation.txt
Runtime_log.txt
Stats_out_MCMC_correct_prob.csv
Stats_out_MCMC_hist.pdf
Stats_out_MCMC_iter.csv
Stats_out_MCMC_iter_summ_stat.csv
Stats_out_MCMC_post_pred.pdf
Stats_out_MCMC_trace.pdf
```

We suggest a quick look at Fragmisincorporation_plot.pdf and Length_plot.pdf to check if there are any problems in sequencing or mapping. The next step is to assess the fit by checking if the empirical substitution frequencies are generally contained in the posterior predictive distribution intervals. If confirmed, try to detect any evidence of non-equilibrium in the stochastic process by looking at worrying trends in the Stats_out_MCMC_trace.pdf file. More rigorous approach, is to run mapDamage2.0 multiple times and use convergence tests[5] on the output in Stats_out_MCMC_iter.csv from parallel chains. The acceptance ratio in Stats_out_MCMC_iter.csv should be about 0.22 but some deviation (0.1-0.3) is acceptable. To improve the fit, increasing the numbers in the following options could be useful.

```
mapDamage -i seq.bam -r ref.fa --rand=30 --burn=10000 --adjust=10 --iter=50000
```

Where –rand is the number of starting points in the likelihood estimation, –burn is the number of iterations in a burning period, –adjust is the number of burning periods and –iter is the number of

[4] A more detailed documention for the parameters is provided in the README.md file.

[5] For example the one presented by Gelman and Rubin, 1992.

iterations in the samples used for the parameter estimation. Note that the total number of MCMC iterations is burn · adjust + iter.

If the sample is from a single strand library build preparation as described in Meyer et al., 2012 then we can run the mapDamage using the following command.

```
mapDamage -i seq.bam -r ref.fa --single_stranded
```

You should see elevated C→T substitutions frequency at both ends in the posterior predictive plot with this option. If using the single ends instead of merged paired ends then we suggest using only the forward part of the sequences.

```
mapDamage -i seq.bam -r ref.fa --forward
```

This could also be addressed by allowing for difference in the mean overhangs lengths at the 5' and 3' end.

```
mapDamage -i seq.bam -r ref.fa --diff-hangs
```

If you are satisfied with the fit then you can append the rescale option to make a rescaled BAM file in the output folder[6].

```
mapDamage -i seq.bam -r ref.fa --rescale
```

You can run the statistical part without reparsing the BAM file. This is useful for convergence tests or model tweaking.

```
mapDamage -d valid_result_folder --stats-only
```

[6]Depending on the dataset it could be worthwhile to explore possible sources of rescaling bias such as heterogeneity in diversity or context dependent sequencing errors, by bipartioning the BAM file and explore the difference in the scaling for the two parts.

References

M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, J. A. Samaniego, M. T. P. Gilbert, E. Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, and M. Bunce. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings. Biological sciences / The Royal Society*, 279(1748):4724–33, Oct. 2012.

C. Bon, V. Berthonaud, F. Maksud, K. Labadie, J. Poulain, F. Artiguenave, P. Wincker, J.-M. Aury, and J.-M. Elalouf. Coprolites as a source of information on the genome and diet of the cave hyena. *Proceedings. Biological sciences / The Royal Society*, 279(1739):2825–30, July 2012.

A. W. Briggs, U. Stenzel, P. L. F. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, and S. Pääbo. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (37):14616–21, Sept. 2007.

A. W. Briggs, J. M. Good, R. E. Green, J. Krause, T. Maricic, U. Stenzel, C. Lalueza-Fox, P. Rudan, D. Brajković, v. Kućan, I. Gušić, R. Schmitz, V. B. Doronichev, L. V. Golovanova, M. de la Rasilla, J. Fortea, A. Rosas, and S. Pääbo. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science (New York, N.Y.)*, 325(5938):318–21, July 2009.

J. Enk, A. Devault, R. Debruyne, C. E. King, T. Treangen, D. O'Rourke, S. L. Salzberg, D. Fisher, R. MacPhee, and H. Poinar. Complete Columbian mammoth mitogenome suggests interbreeding with woolly mammoths. *Genome biology*, 12(5):R51, Jan. 2011.

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, Nov. 1992.

S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, Nov. 1984.

R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo. A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, 328(5979):710–22, May 2010.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2):160–174, 1985.

W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

M. Knapp, K. A. Horsburgh, S. Prost, J.-A. Stanton, H. R. Buckley, R. K. Walter, and E. A. Matisoo-Smith. Complete mitochondrial DNA genome sequences from the first New Zealanders. *Proceedings of the National Academy of Sciences of the United States of America*, 109(45):18350–18354, 2012.

B. Li, G. Zhang, E. Willerslev, J. Wang, and J. Wang. Genomic data from the polar bear (Ursus maritimus). 2011.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, July 2009.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, Aug. 2009.

S. Lindgreen. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes*, 5:337, Jan. 2012.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–303, Sept. 2010.

M. Meyer and M. Kircher. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor protocols*, 2010(6):pdb.prot5448, June 2010.

M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. D. Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, 338(6104):222–226, 2012.

W. Miller, S. C. Schuster, A. J. Welch, A. Ratan, O. C. Bedoya-Reina, F. Zhao, H. L. Kim, R. C. Burhans, D. I. Drautz, N. E. Wittekindt, L. P. Tomsho, E. Ibarra-Laclette, L. Herrera-Estrella, E. Peacock, S. Farley, G. K. Sage, K. Rode, M. Obbard, R. Montiel, L. Bachmann, O. Ingólfsson, J. Aars, T. Mailund, O. Wiig, S. L. Talbot, and C. Lindqvist. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 109(36):E2382–90, Sept. 2012.

M. Rasmussen, X. Guo, Y. Wang, K. E. Lohmueller, S. Rasmussen, A. Albrechtsen, L. Skotte, S. Lindgreen, M. Metspalu, T. Jombart, T. Kivisild, W. Zhai, A. Eriksson, A. Manica, L. Orlando, F. M. De La Vega, S. Tridico, E. Metspalu, K. Nielsen, M. C. Ávila Arcos, J. V. Moreno-Mayar, C. Muller, J. Dortch, M. T. P. Gilbert, O. Lund, A. Wesolowska, M. Karmin, L. A. Weinert, B. Wang, J. Li, S. Tai, F. Xiao, T. Hanihara, G. van Driem, A. R. Jha, F.-X. Ricaut, P. de Knijff, A. B. Migliano, I. C. Romero, K. Kristiansen, D. M. Lambert, S. Brunak, P. Forster, B. Brinkmann, O. Nehlich, M. Bunce, M. Richards, R. Gupta, C. D. Bustamante, A. Krogh, R. A. Foley, M. M. Lahr, F. Balloux, T. Sicheritz-Pontén, R. Villems, R. Nielsen, J. Wang, and E. Willerslev. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science (New York, N.Y.)*, 334(6052): 94–8, Oct. 2011.

D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, and S. Pääbo. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–60, Dec. 2010.

F. Sánchez-Quinto, H. Schroeder, O. Ramirez, M. C. Ávila Arcos, M. Pybus, I. Olalde, A. M. V. Velazquez, M. E. P. Marcos, J. M. V. Encinas, J. Bertranpetit, L. Orlando, M. T. P. Gilbert, and C. Lalueza-Fox. Genomic affinities of two 7,000-year-old Iberian hunter-gatherers. *Current biology*, 22(16):1494–9, Aug. 2012.

M. Schubert, A. Ginolhac, S. Lindgreen, J. F. Thompson, K. A. Al-Rasheid, E. Willerslev, A. Krogh, and L. Orlando. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13(1):178, 2012.

V. J. Schuenemann, K. Bos, S. DeWitte, S. Schmedes, J. Jamieson, A. Mittnik, S. Forrest, B. K. Coombes, J. W. Wood, D. J. D. Earn, W. White, J. Krause, and H. N. Poinar. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. *Proceedings of the National Academy of Sciences of the United States of America*, 108(38): E746–52, Sept. 2011.

P. Skoglund, H. Malmström, M. Raghavan, J. Storå, P. Hall, E. Willerslev, M. T. P. Gilbert, A. Götherström, and M. Jakobsson. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science (New York, N.Y.)*, 336(6080):466–9, Apr. 2012.