Joint estimates of heterozygosity and runs of homozygosity for modern and ancient samples

Supplementary Material

Gabriel Renaud, Kristian Hanghøj, Thorfinn Sand Korneliussen,
Eske Willerslev and Ludovic Orlando

Contents

1 Supplementary Results

1.1 Simulated data

1.1.1 Local estimates of heterozygosity

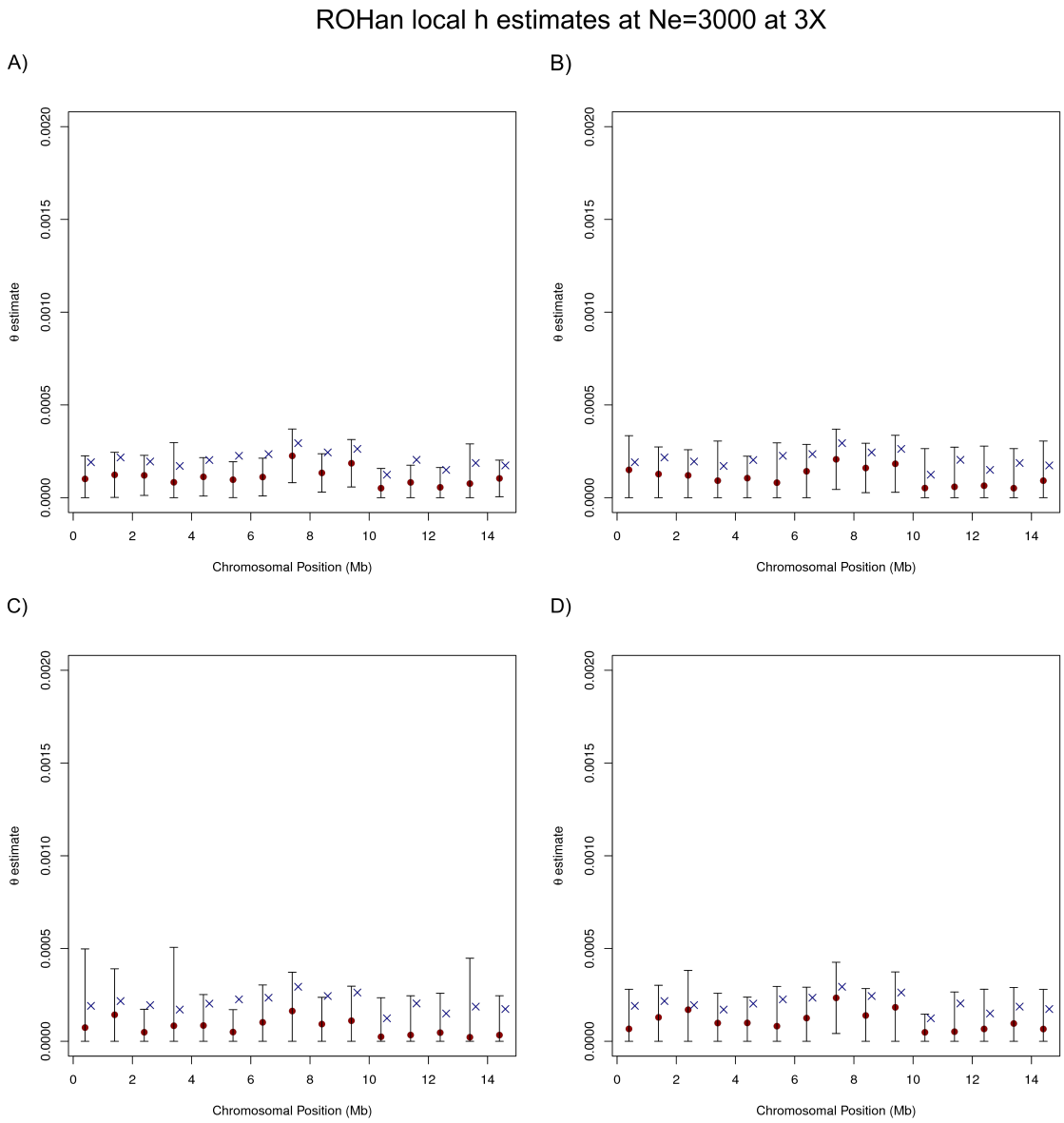

Figure 1: Comparison between the simulated local rates of heterozygosity versus the predicted one on windows of 1Mbp, $N_e = 3000$, a coverage of 3X using A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña. The red dot represents the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval and the dark blue crosses represent the simulated value.

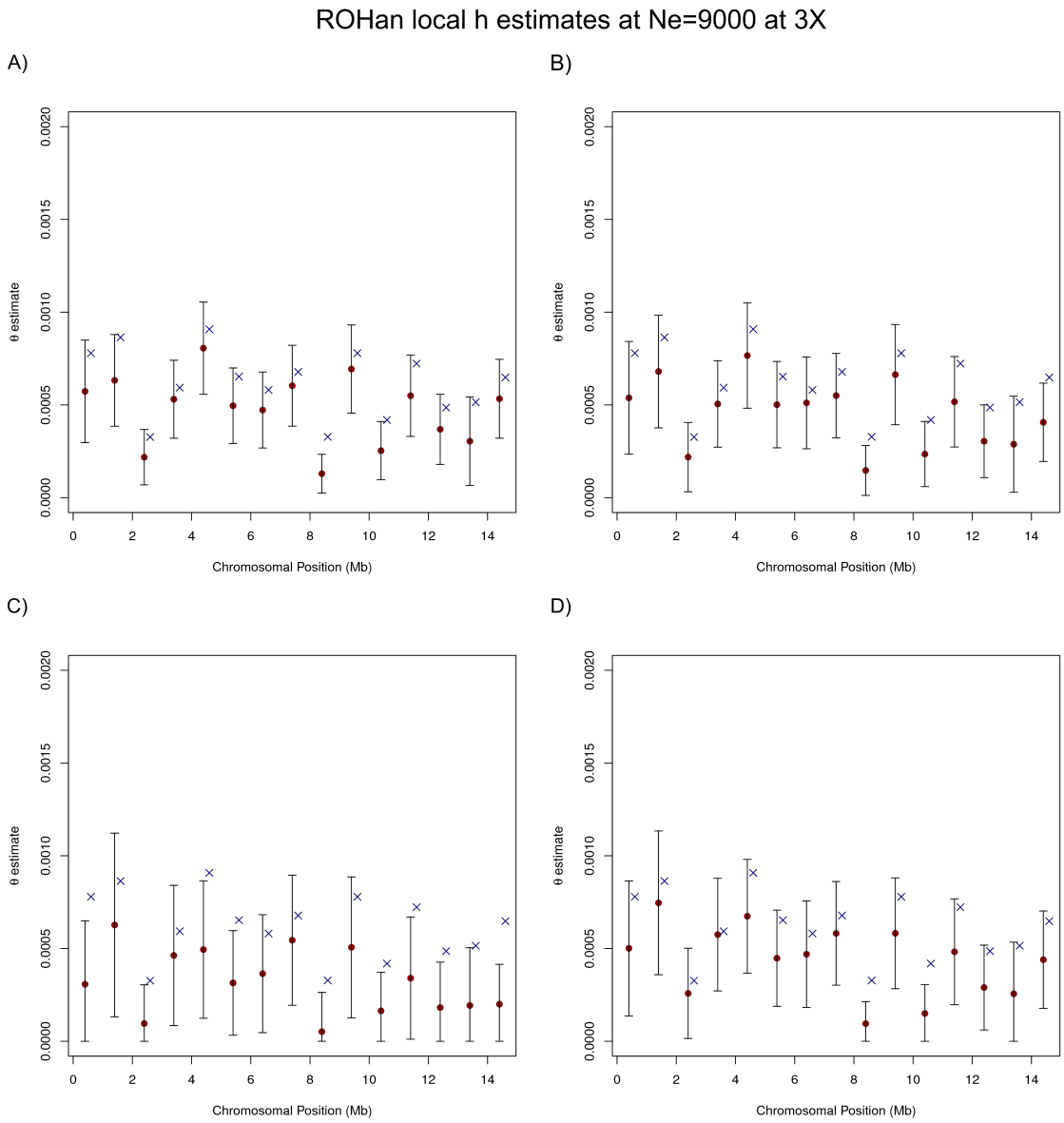

Figure 2: Comparison between the simulated local rates of heterozygosity versus the predicted one on windows of 1Mbp, $N_e = 9000$, a coverage of 3X using A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña. The red dot represents the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval and the dark blue crosses represent the simulated value.

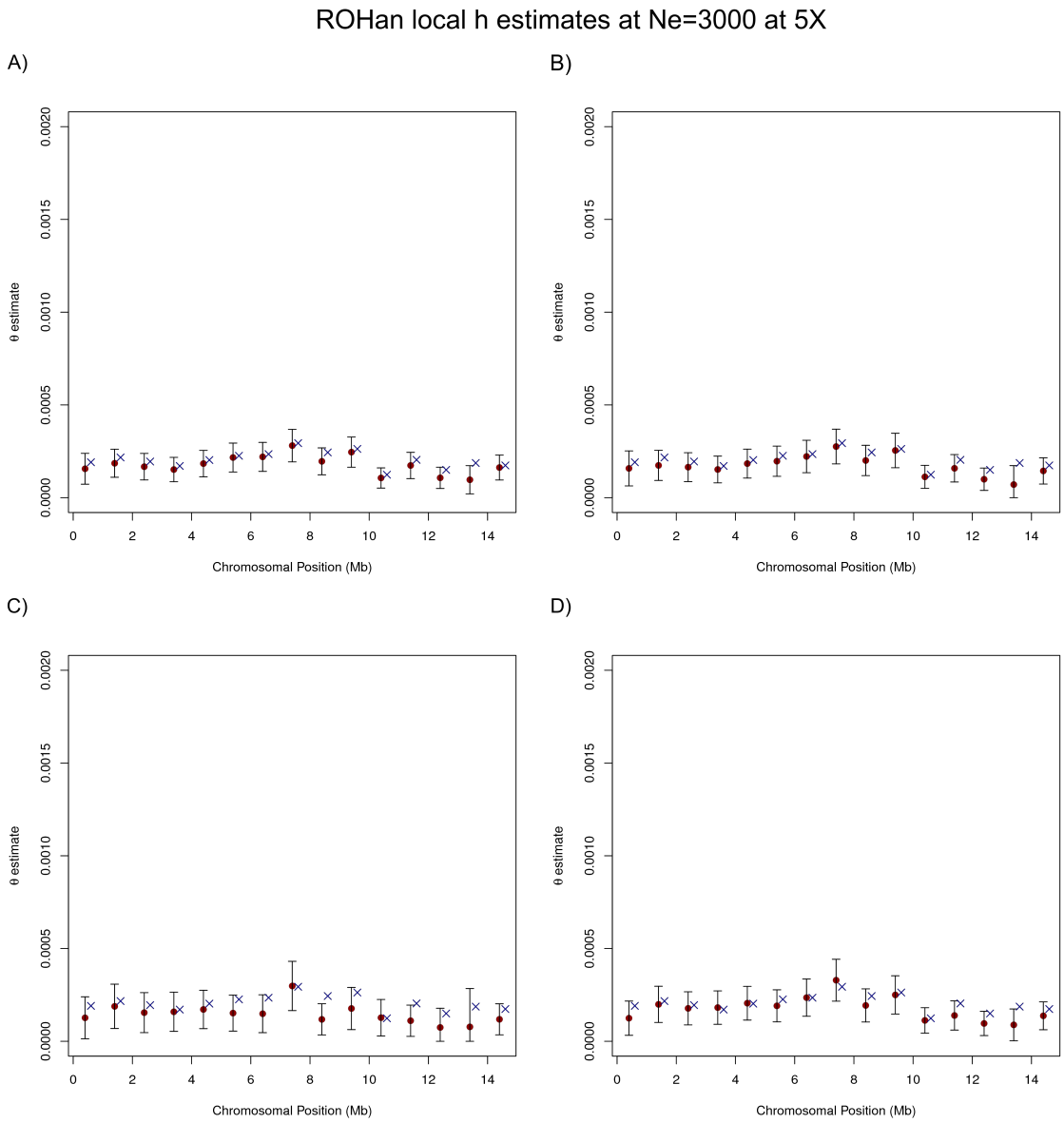

Figure 3: Comparison between the simulated local rates of heterozygosity versus the predicted one on windows of 1Mbp, $N_e = 3000$, a coverage of 5X using A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña. The red dot represents the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval and the dark blue crosses represent the simulated value.

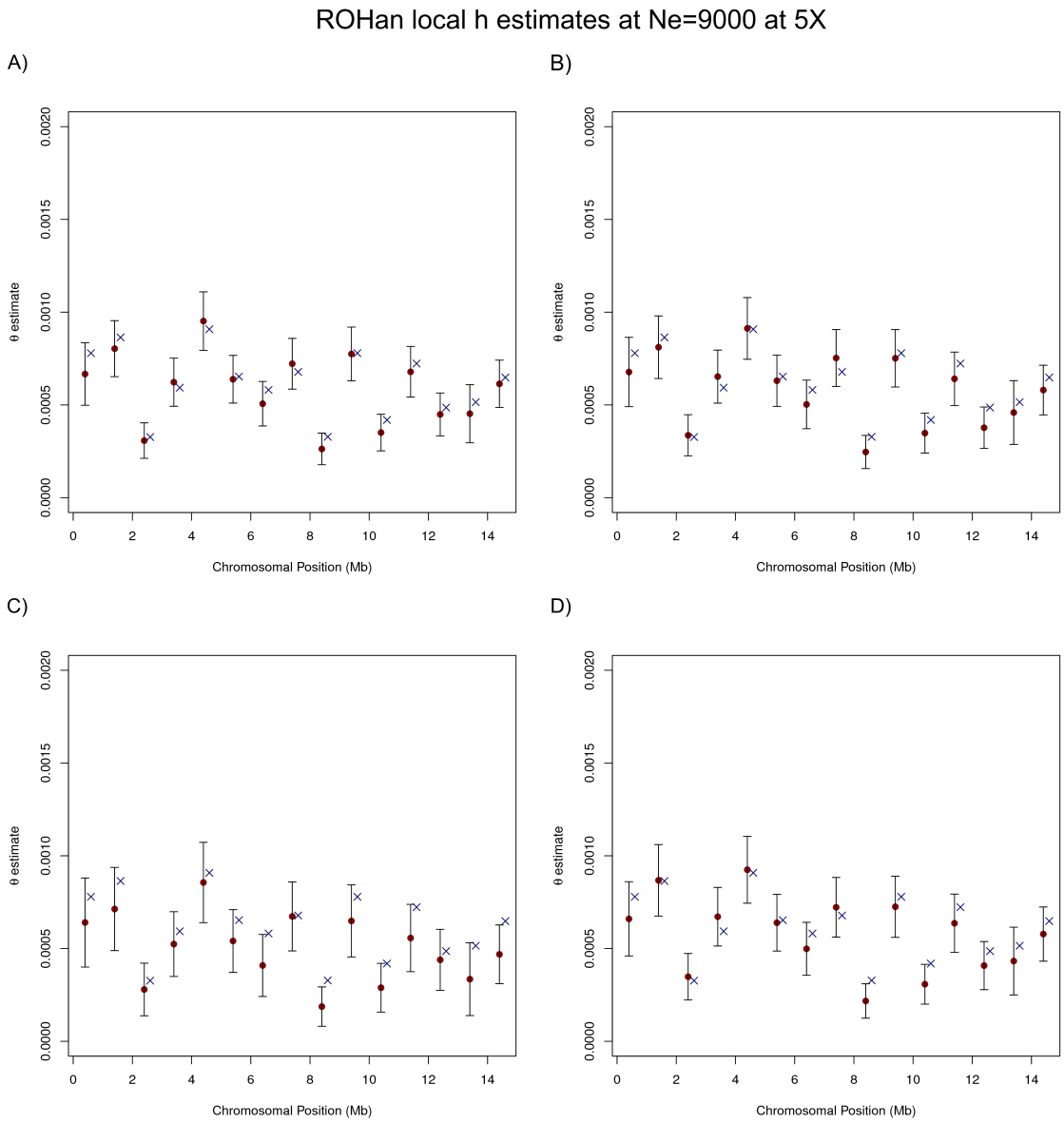

Figure 4: Comparison between the simulated local rates of heterozygosity versus the predicted one on windows of 1Mbp, $N_e = 9000$, a coverage of 5X using A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña. The red dot represents the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval and the dark blue crosses represent the simulated value.

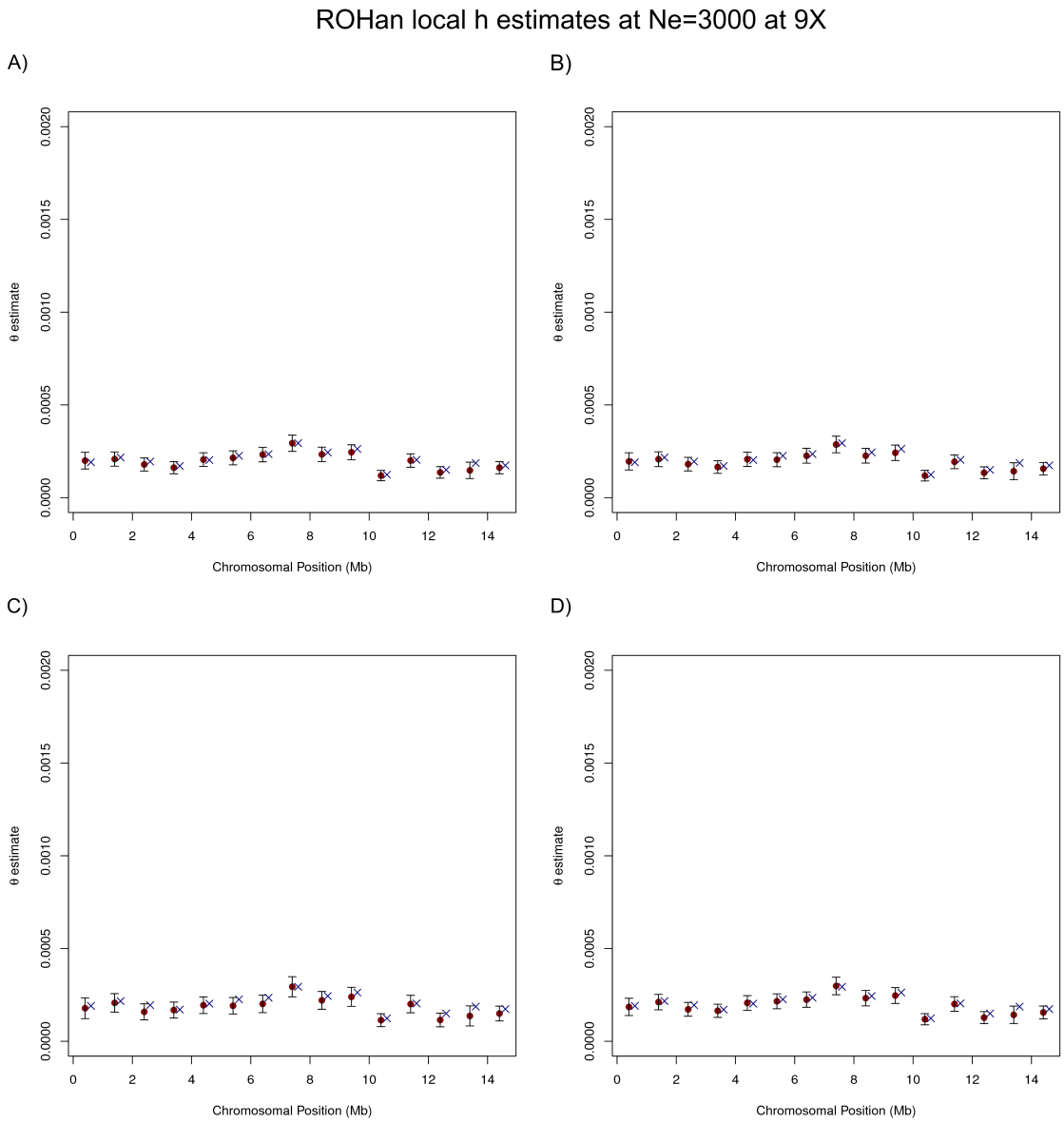

Figure 5: Comparison between the simulated local rates of heterozygosity versus the predicted one on windows of 1Mbp, $N_e = 3000$, a coverage of 9X using A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña. The red dot represents the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval and the dark blue crosses represent the simulated value.

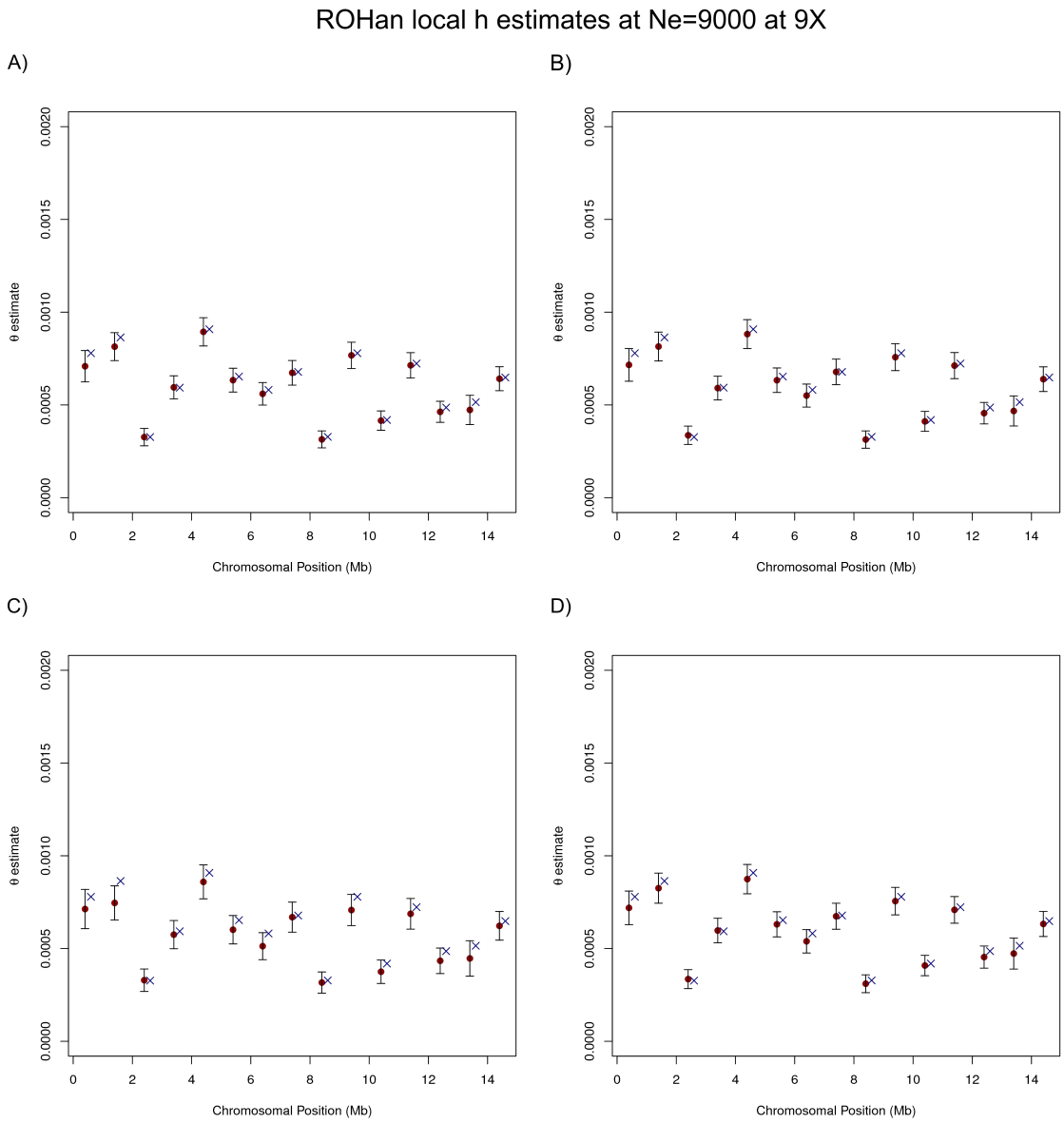

Figure 6: Comparison between the simulated local rates of heterozygosity versus the predicted one on windows of 1Mbp, $N_e = 9000$, a coverage of 9X using A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña. The red dot represents the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval and the dark blue crosses represent the simulated value.

1.1.2 Global estimates of heterozygosity

1.1.2.1 ROHan

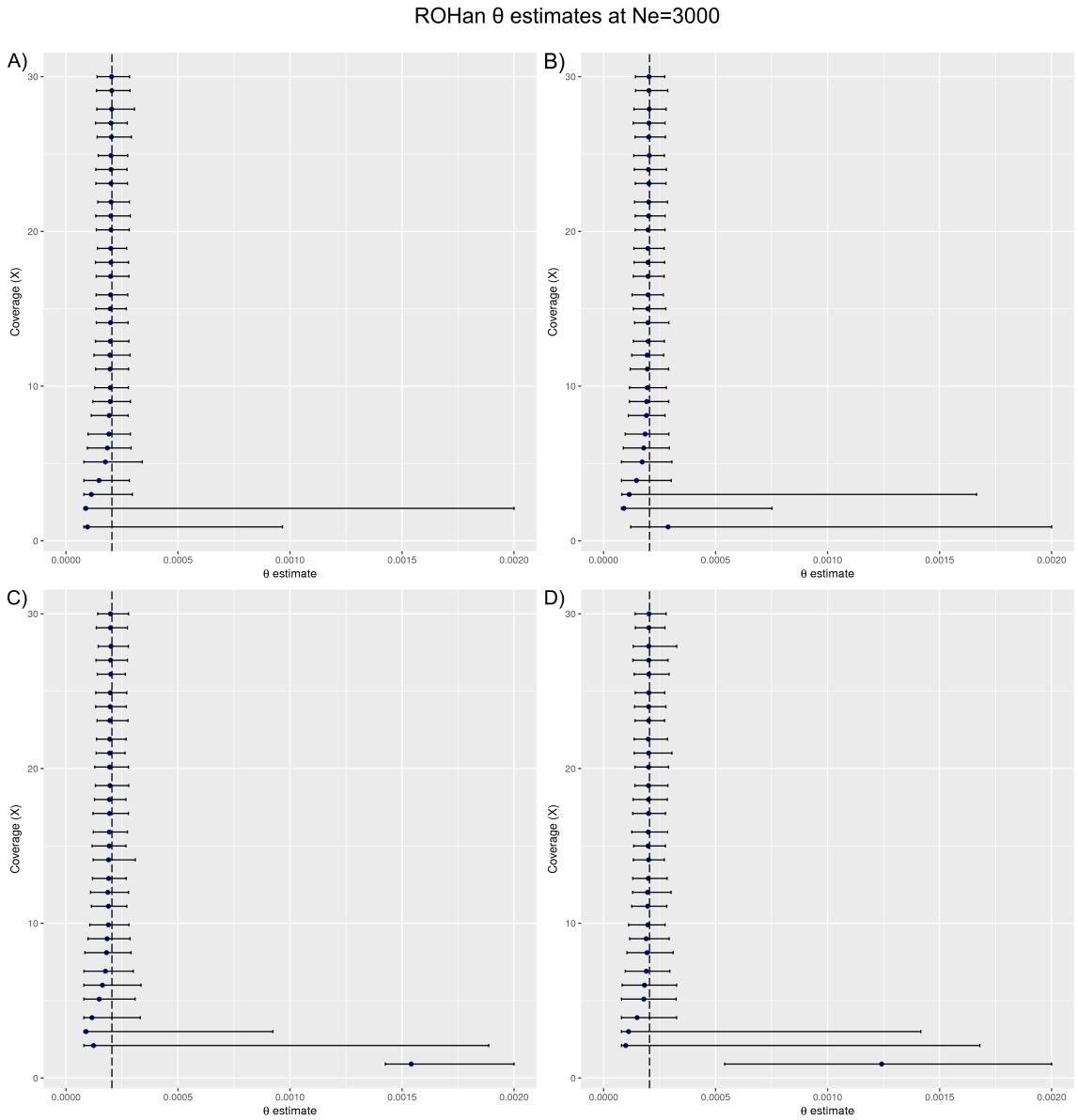

Figure 7: Simulated versus predicted genome-wide θ by ROHan for a simulated chromosome of 15Mbp and an effective population size of 3000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

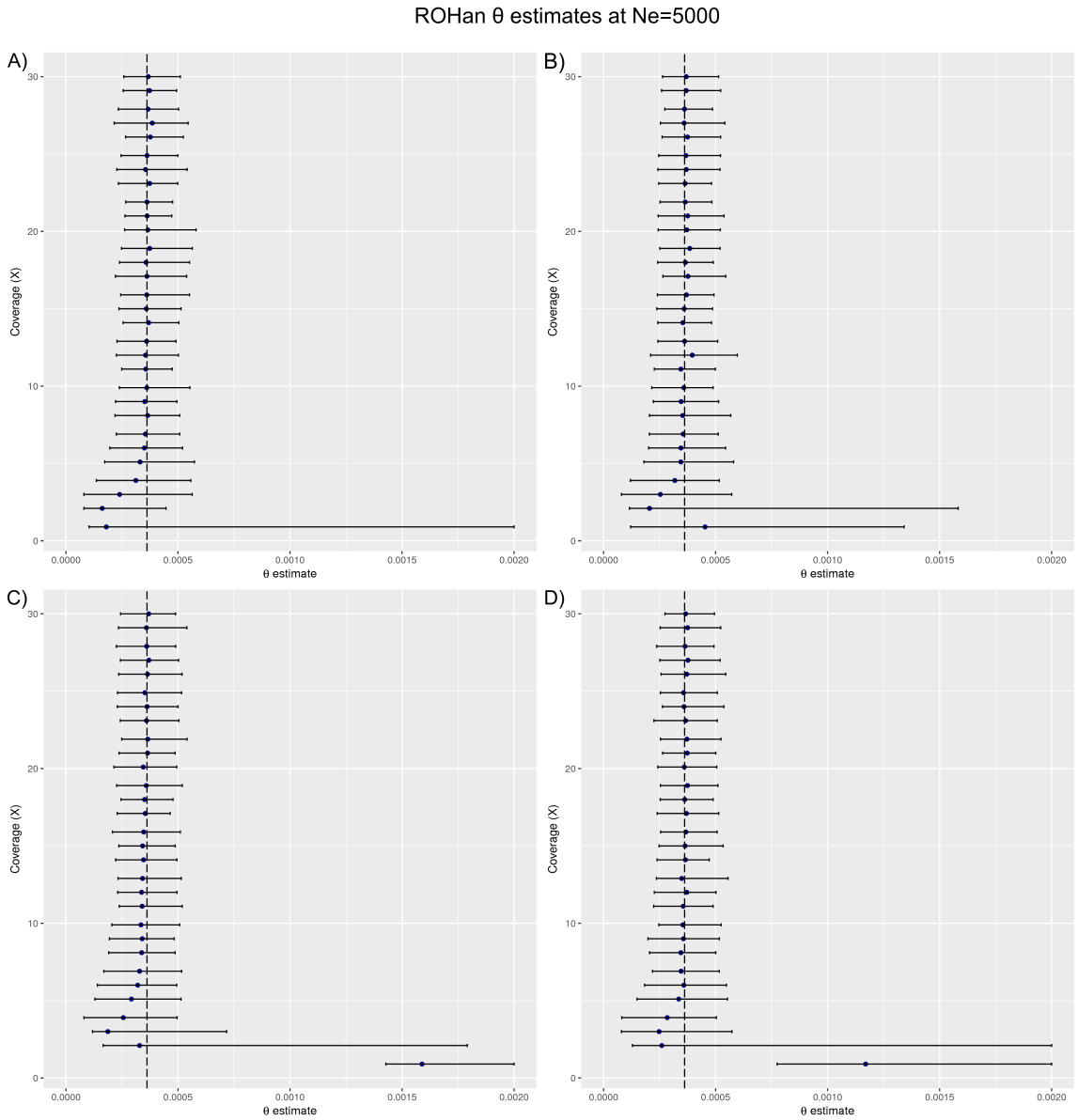

Figure 8: Simulated versus predicted genome-wide θ by ROHan for a simulated chromosome of 15Mbp and an effective population size of 5000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

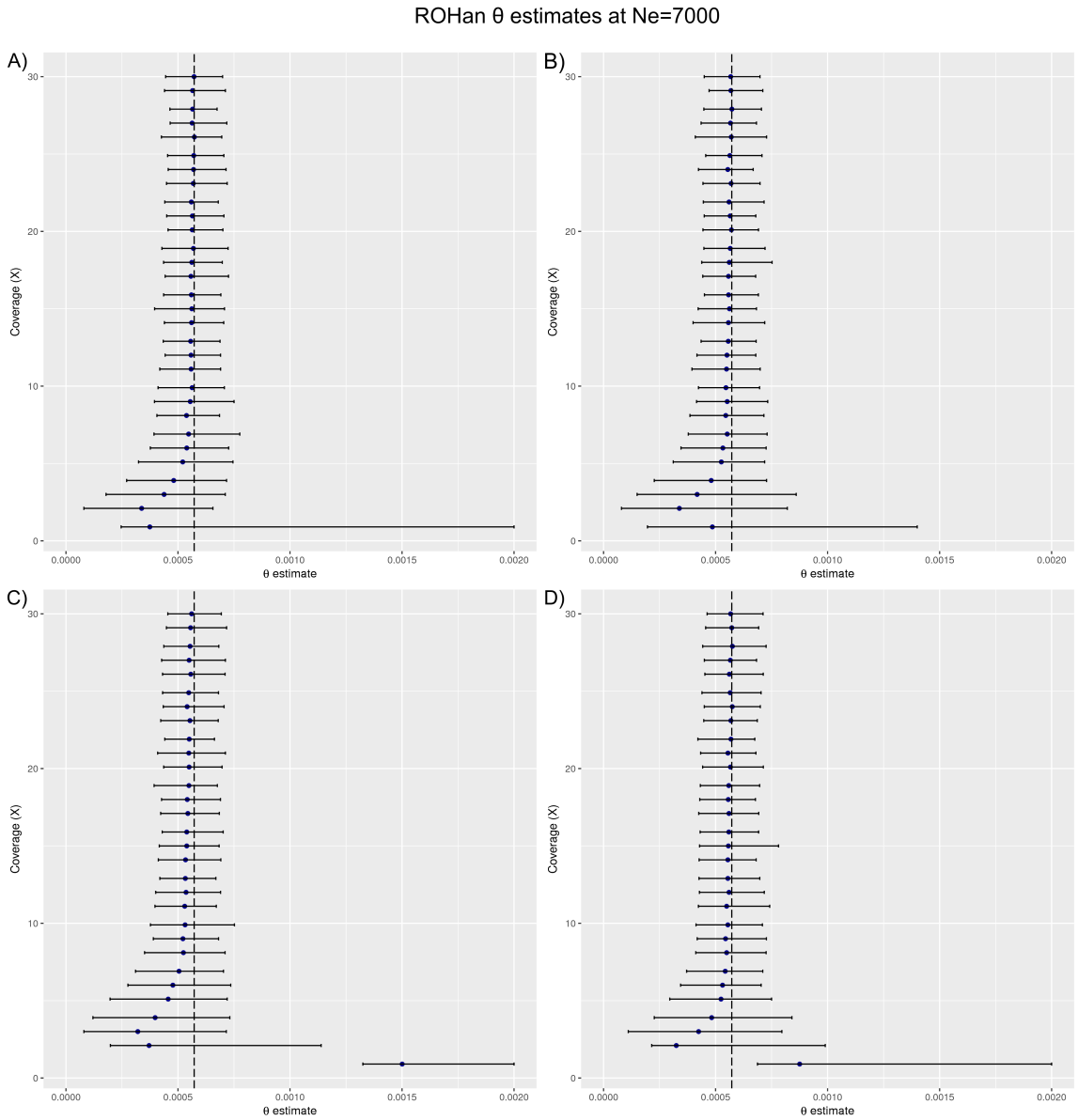

Figure 9: Simulated versus predicted genome-wide θ by ROHan for a simulated chromosome of 15Mbp and an effective population size of 7000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

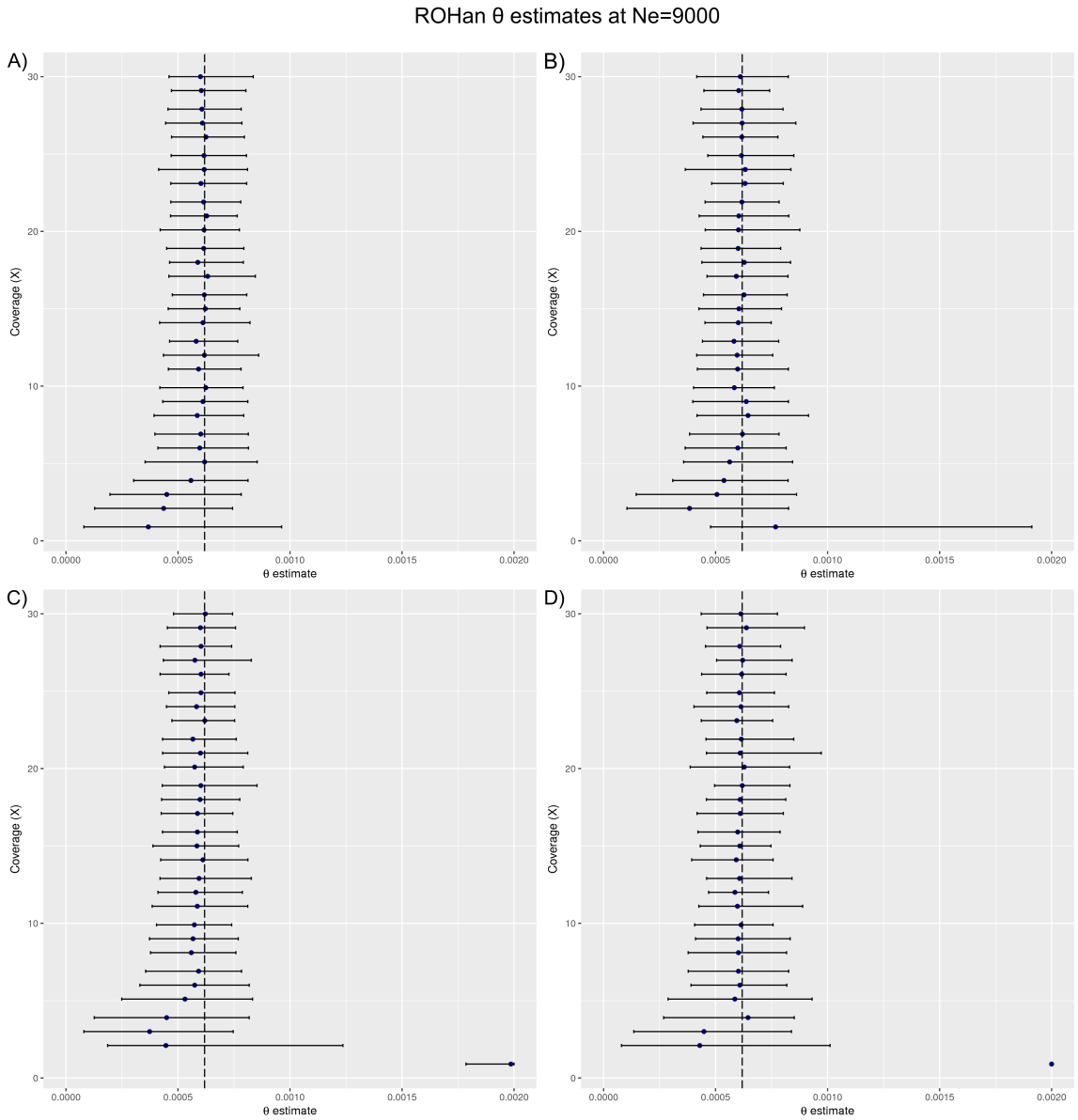

Figure 10: Simulated versus predicted genome-wide θ by ROHan for a simulated chromosome of 15Mbp and an effective population size of 9000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

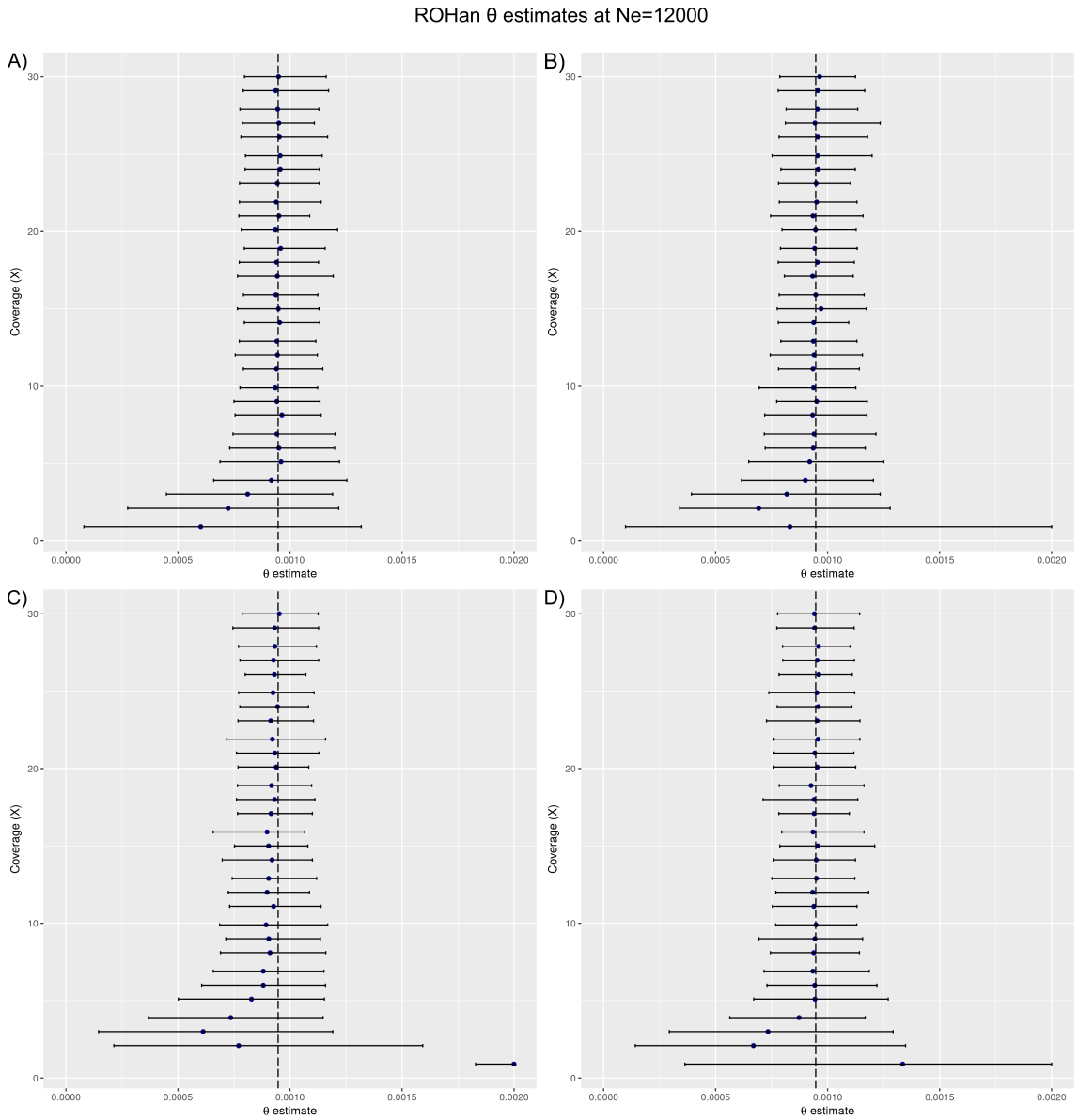

Figure 11: Simulated versus predicted genome-wide θ by ROHan for a simulated chromosome of 15Mbp and an effective population size of 12000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

1.1.2.2 ATLAS

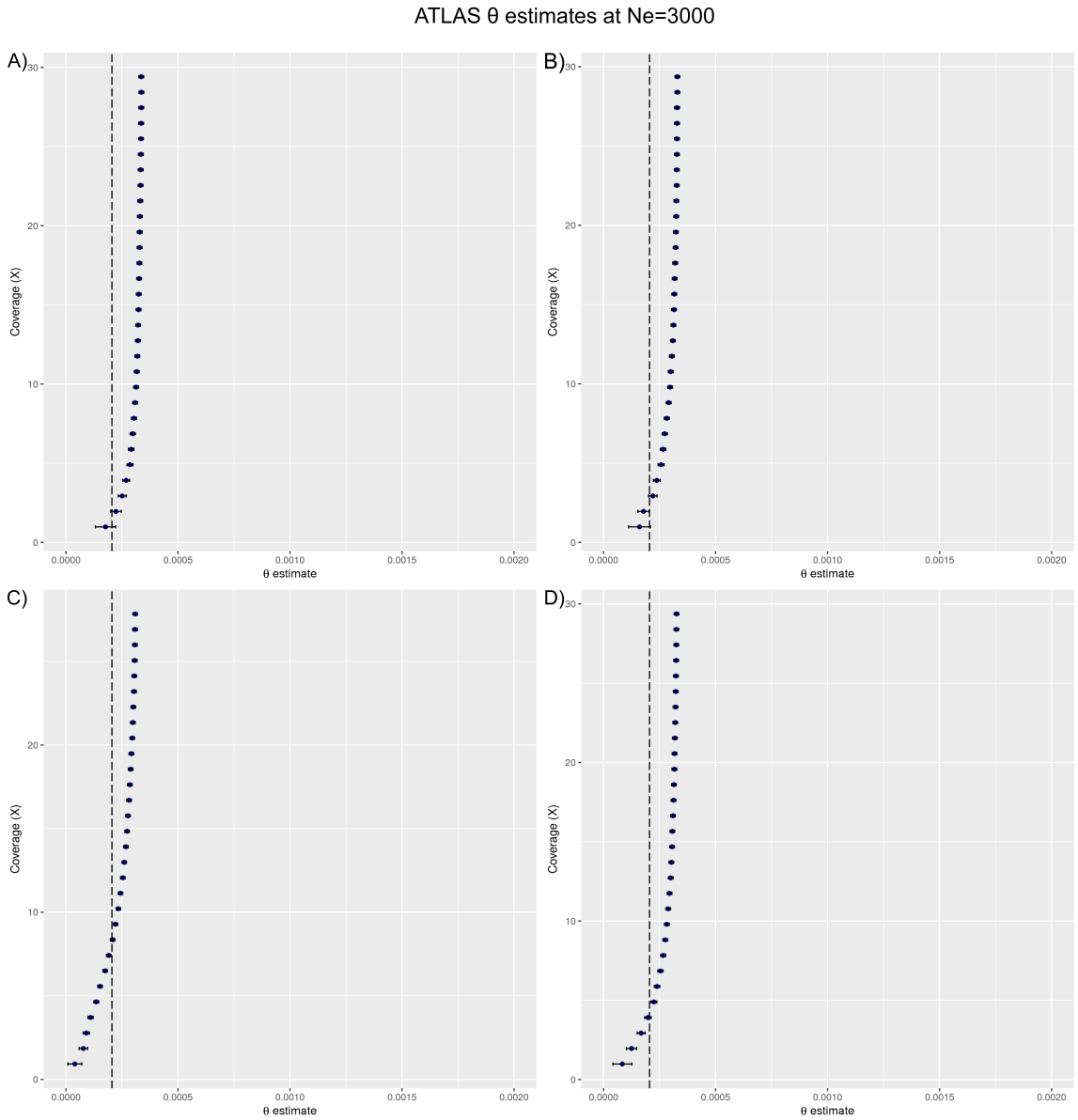

Figure 12: Simulated versus predicted genome-wide θ by ATLAS for a simulated chromosome of 15Mbp and an effective population size of 3000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

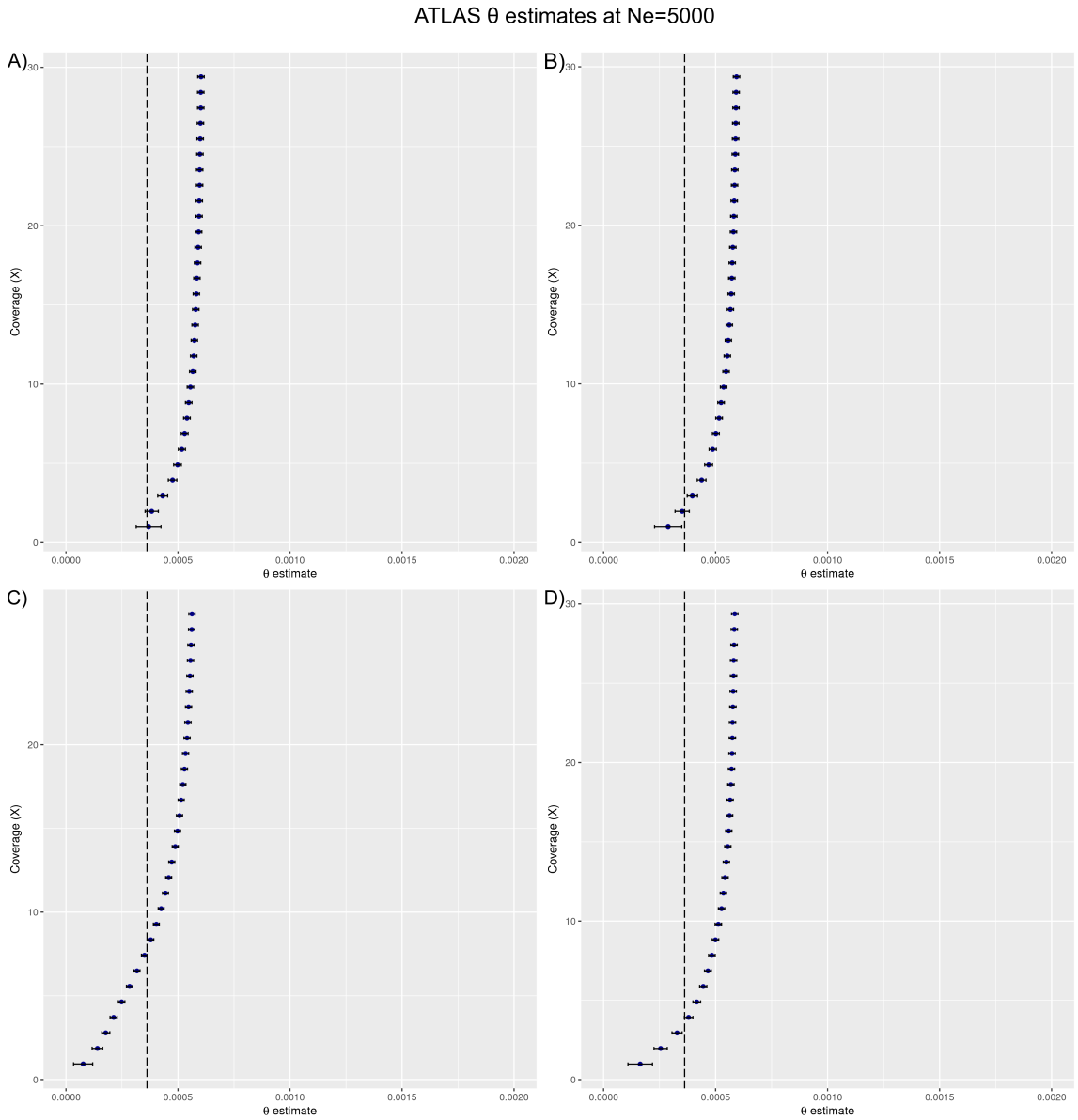

Figure 13: Simulated versus predicted genome-wide θ by ATLAS for a simulated chromosome of 15Mbp and an effective population size of 5000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

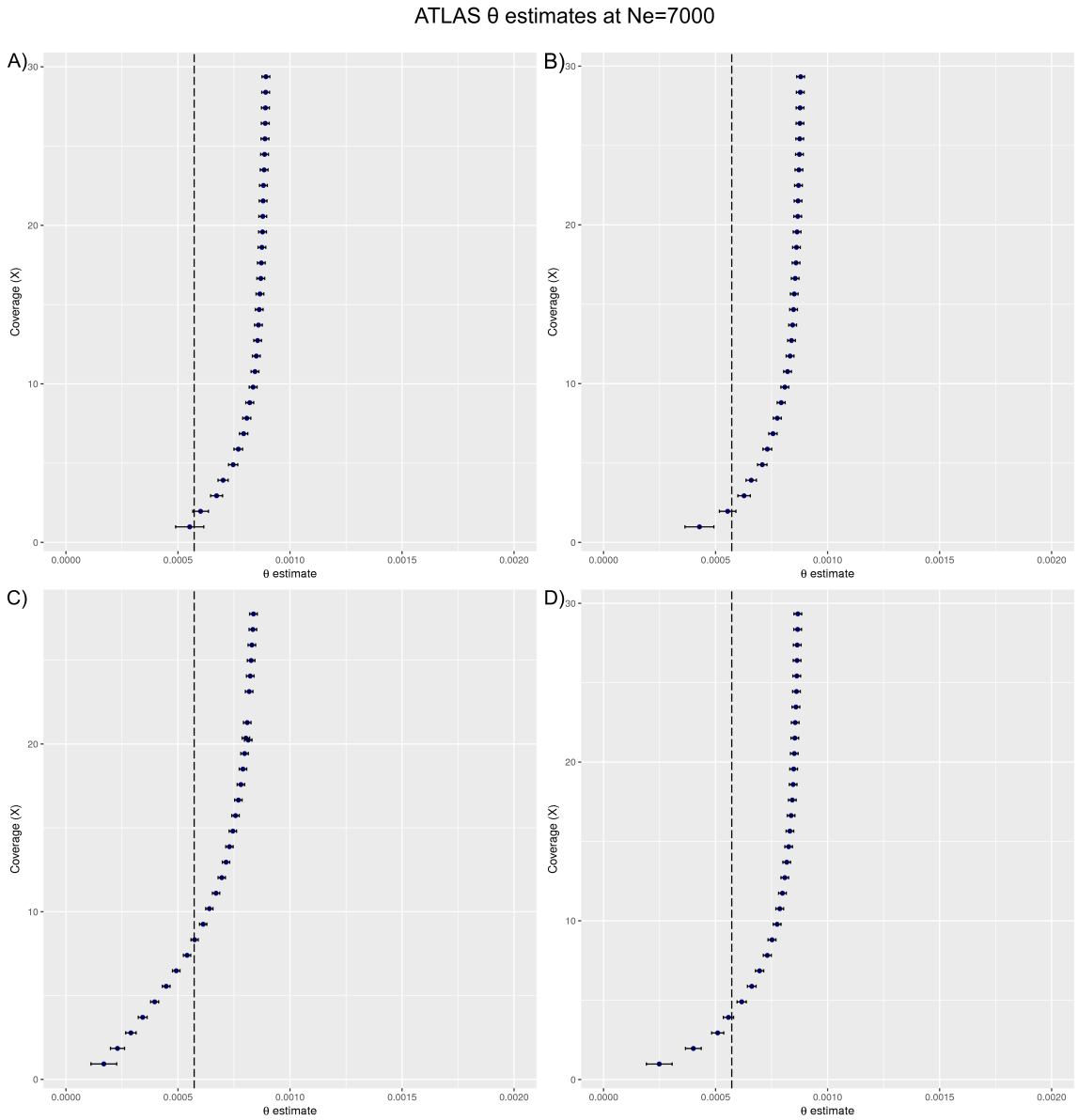

Figure 14: Simulated versus predicted genome-wide θ by ATLAS for a simulated chromosome of 15Mbp and an effective population size of 7000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

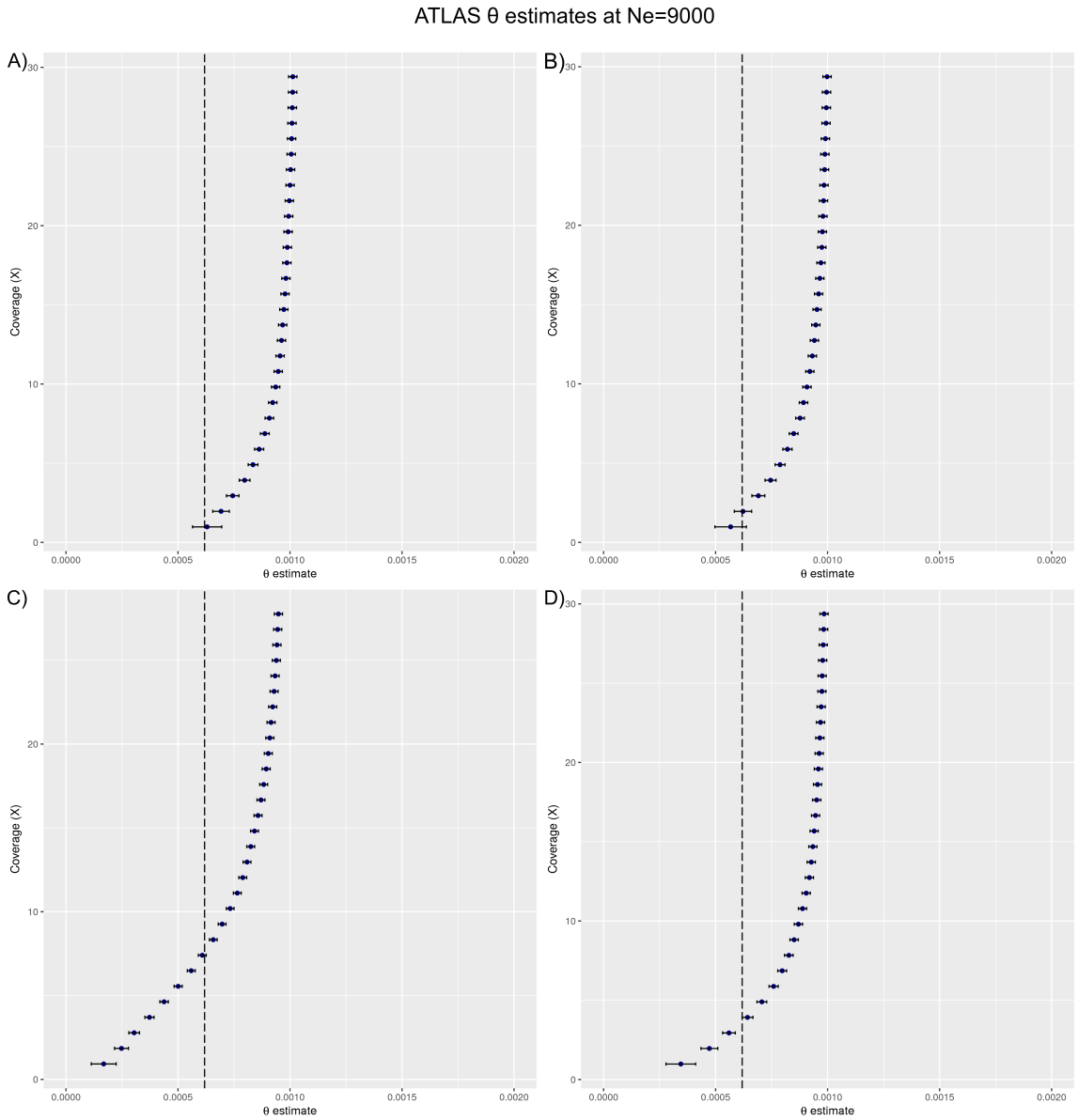

Figure 15: Simulated versus predicted genome-wide θ by ATLAS for a simulated chromosome of 15Mbp and an effective population size of 9000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

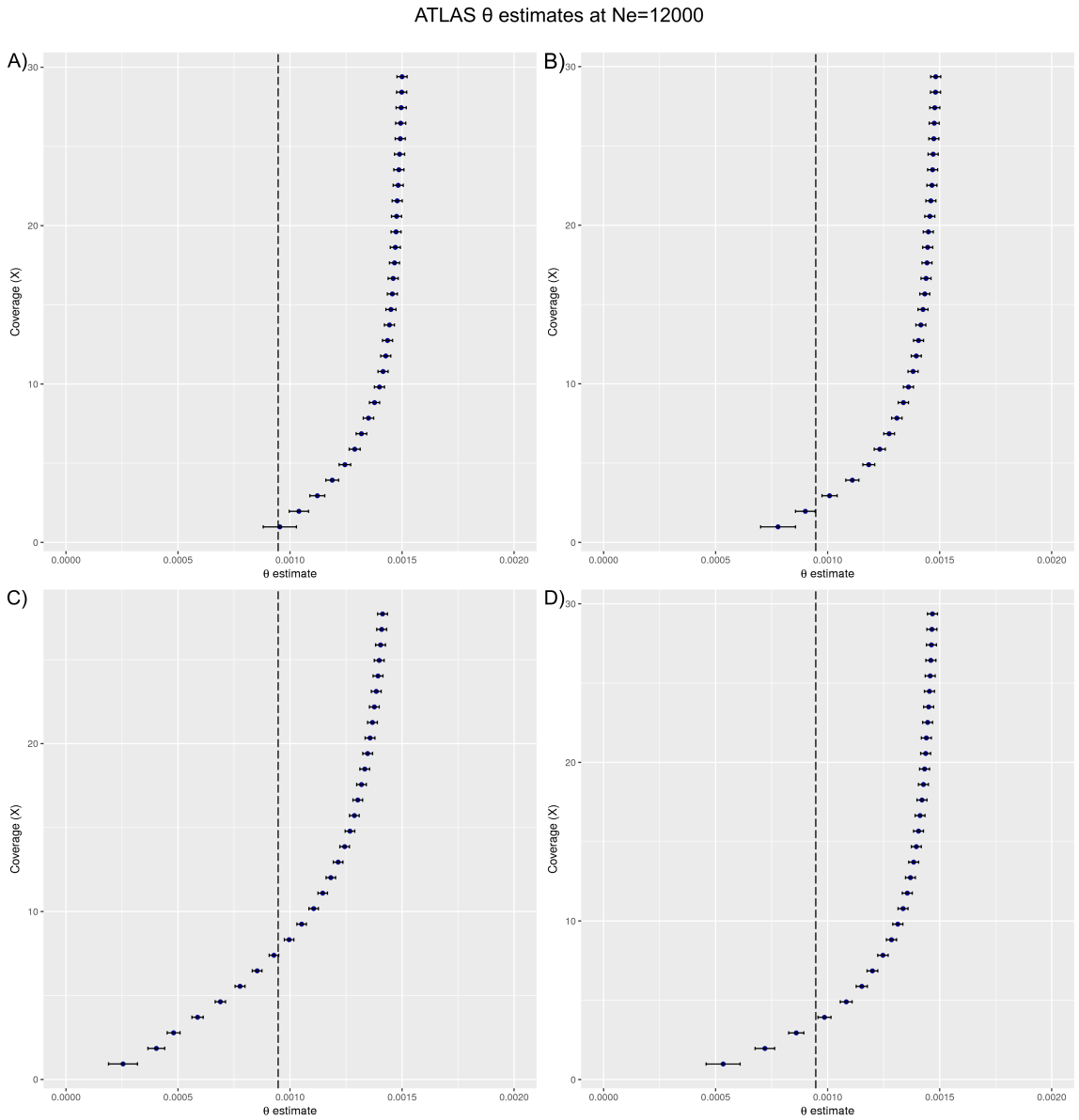

Figure 16: Simulated versus predicted genome-wide θ by ATLAS for a simulated chromosome of 15Mbp and an effective population size of 12000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

1.1.2.3 ANGSD

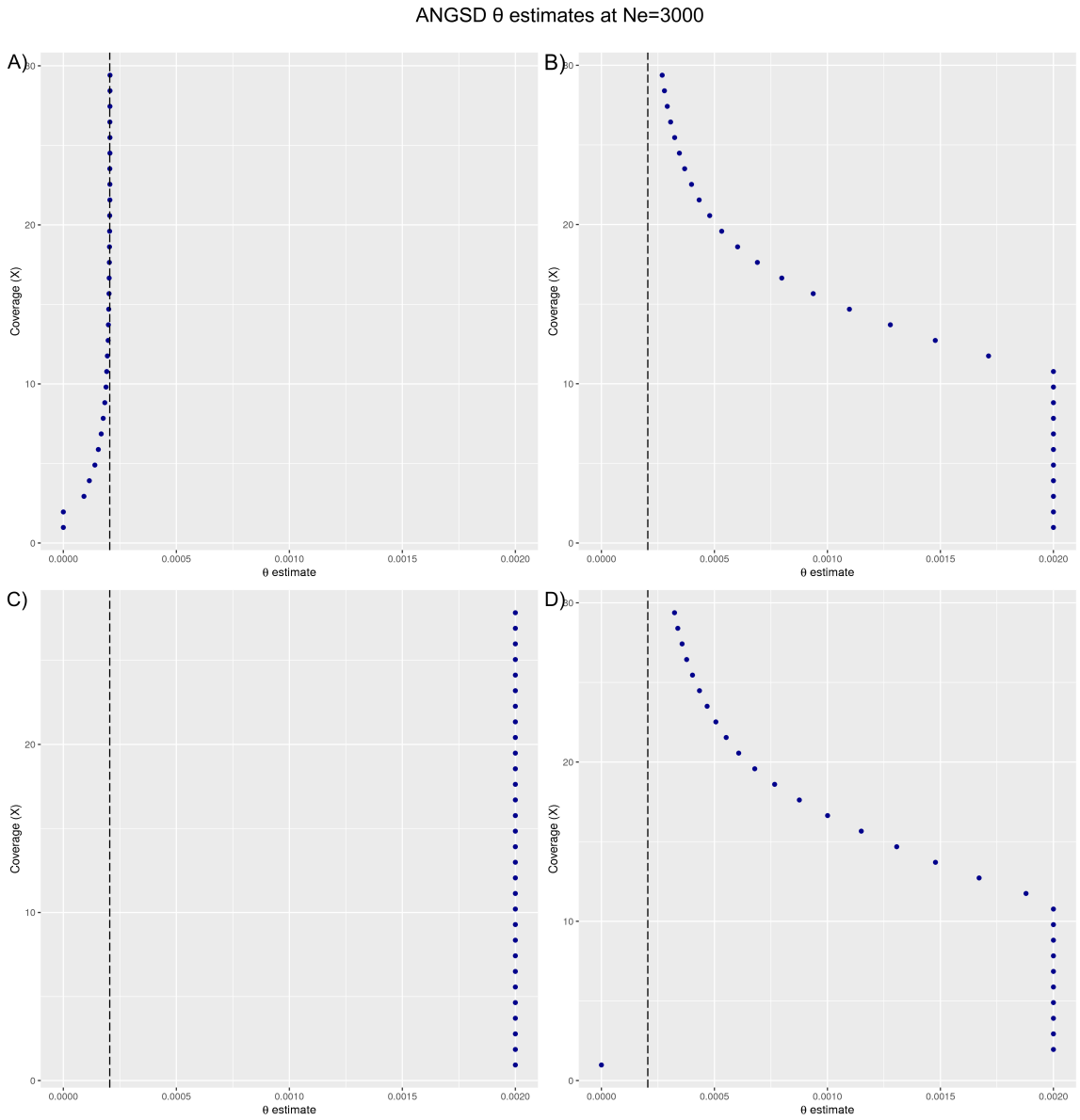

Figure 17: Simulated versus predicted genome-wide θ by ANGSD for a simulated chromosome of 15Mbp and an effective population size of 3000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

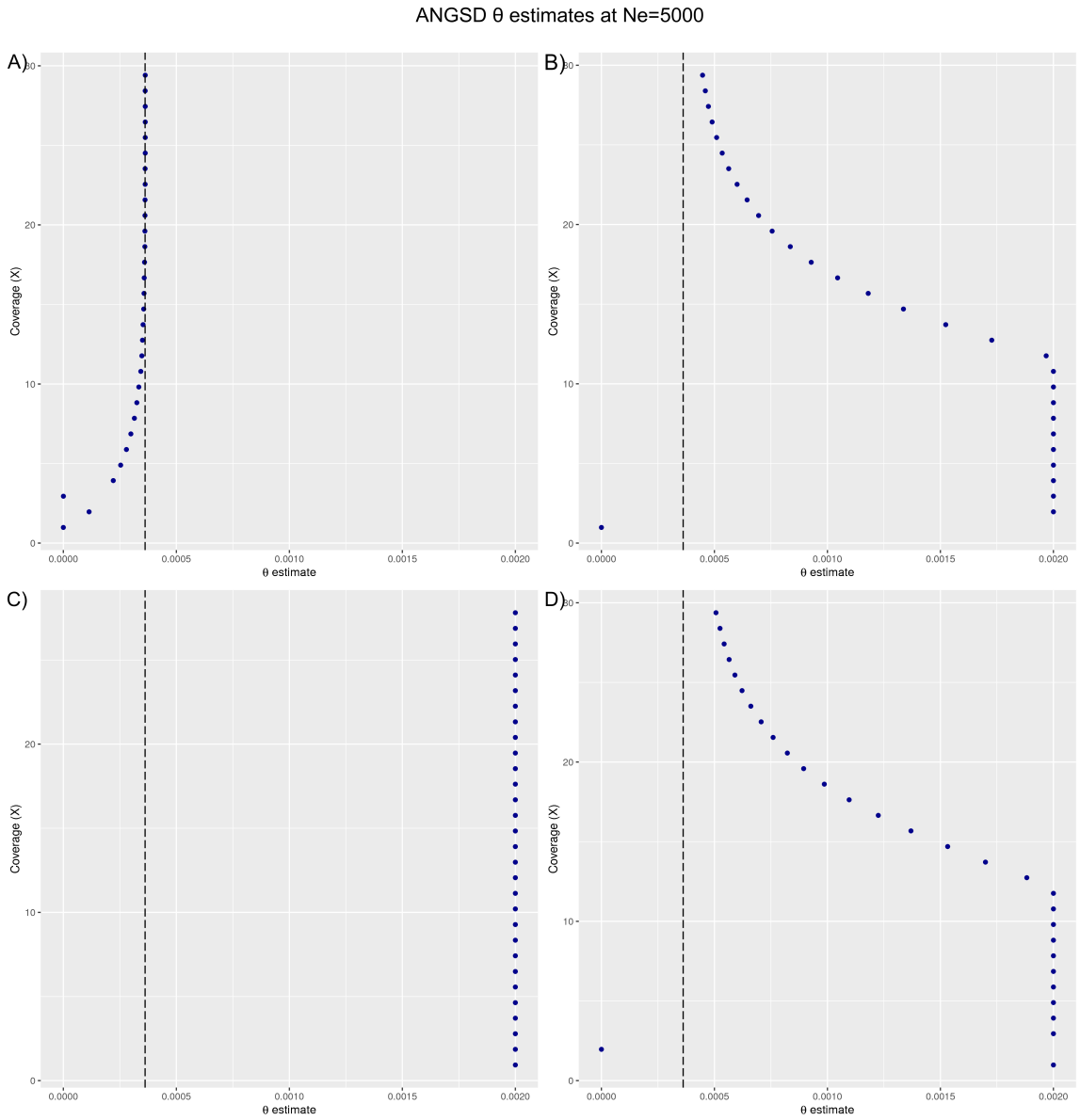

Figure 18: Simulated versus predicted genome-wide θ by ANGSD for a simulated chromosome of 15Mbp and an effective population size of 5000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

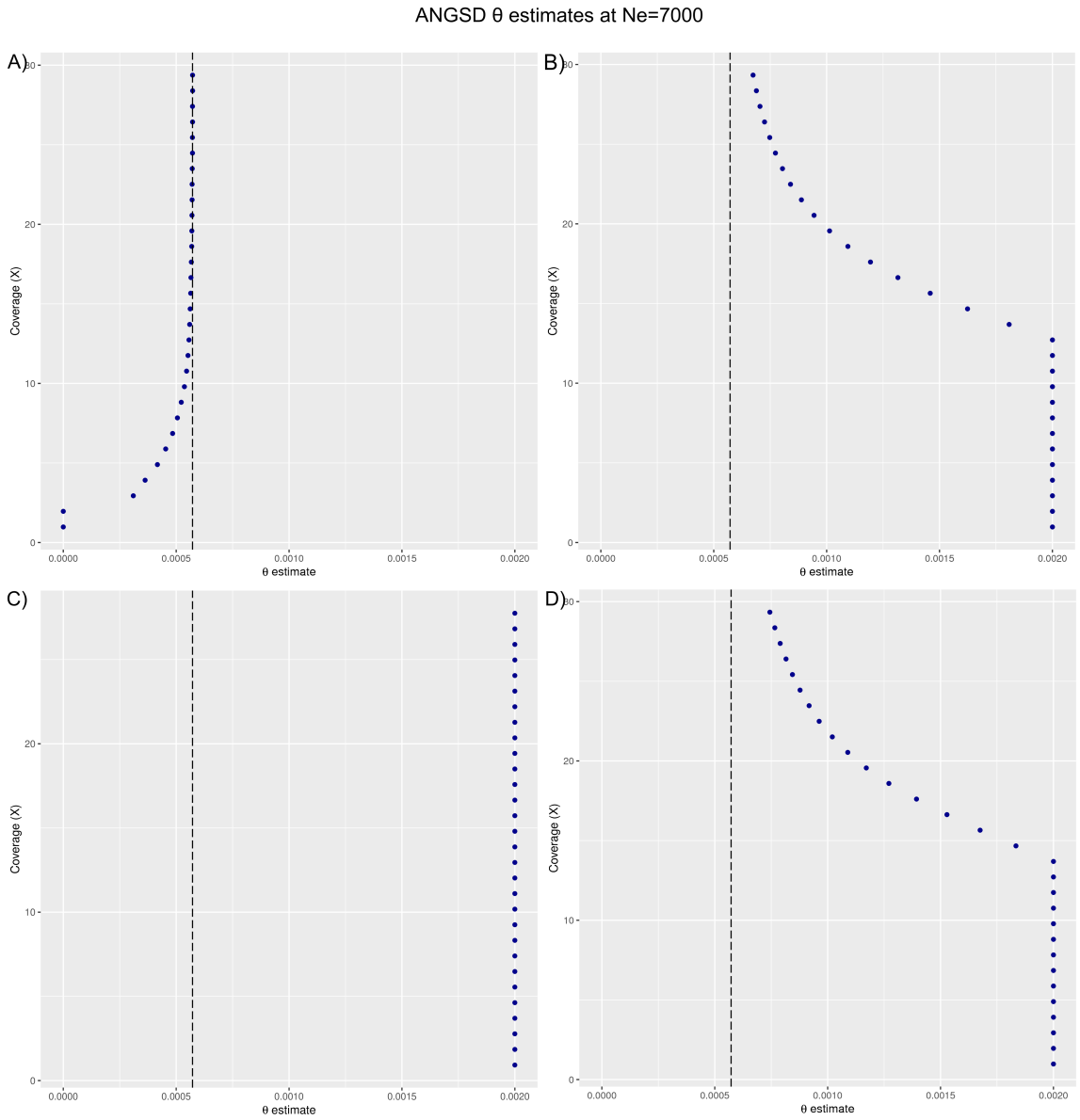

Figure 19: Simulated versus predicted genome-wide θ by ANGSD for a simulated chromosome of 15Mbp and an effective population size of 7000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

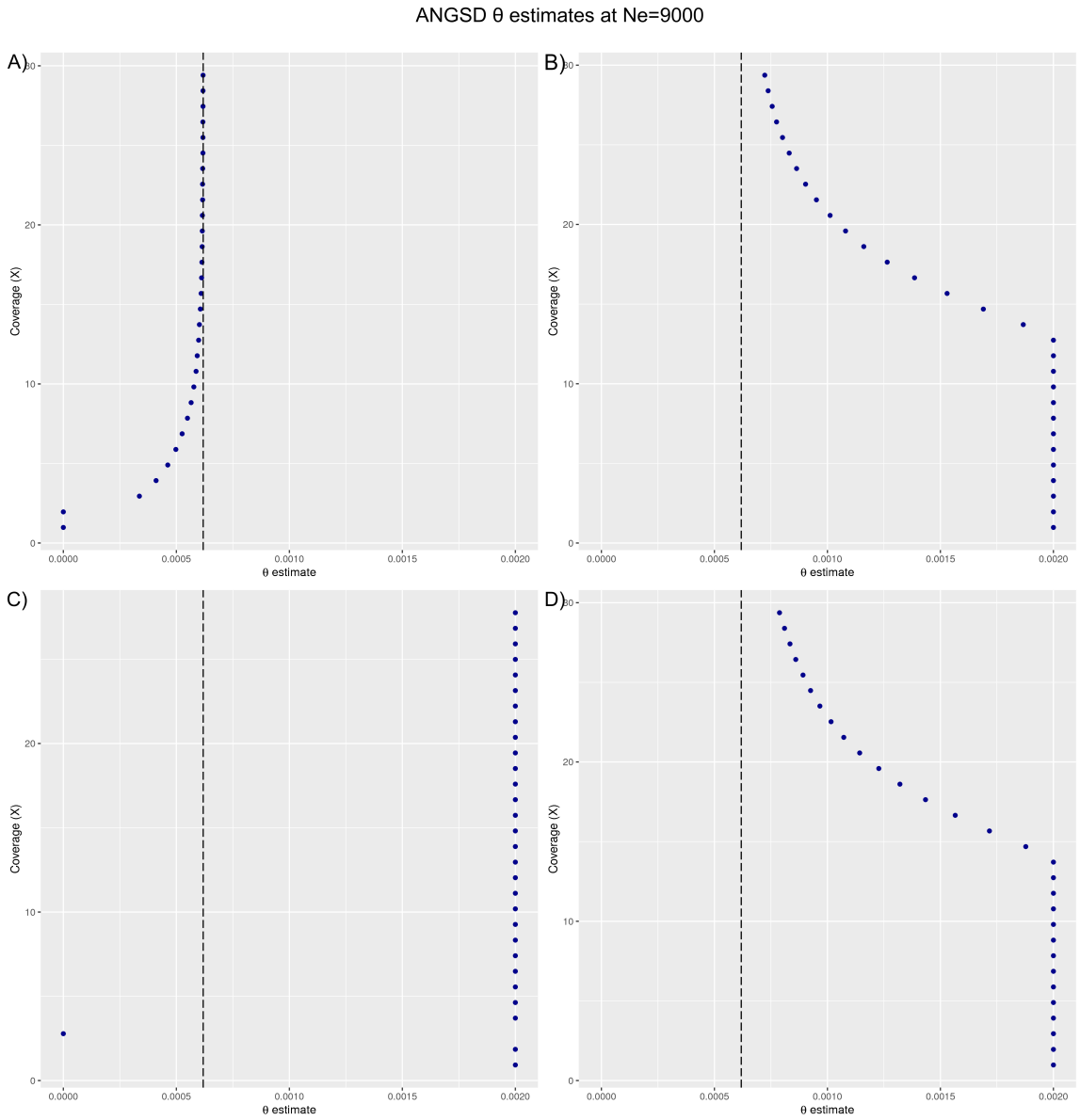

Figure 20: Simulated versus predicted genome-wide θ by ANGSD for a simulated chromosome of 15Mbp and an effective population size of 9000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

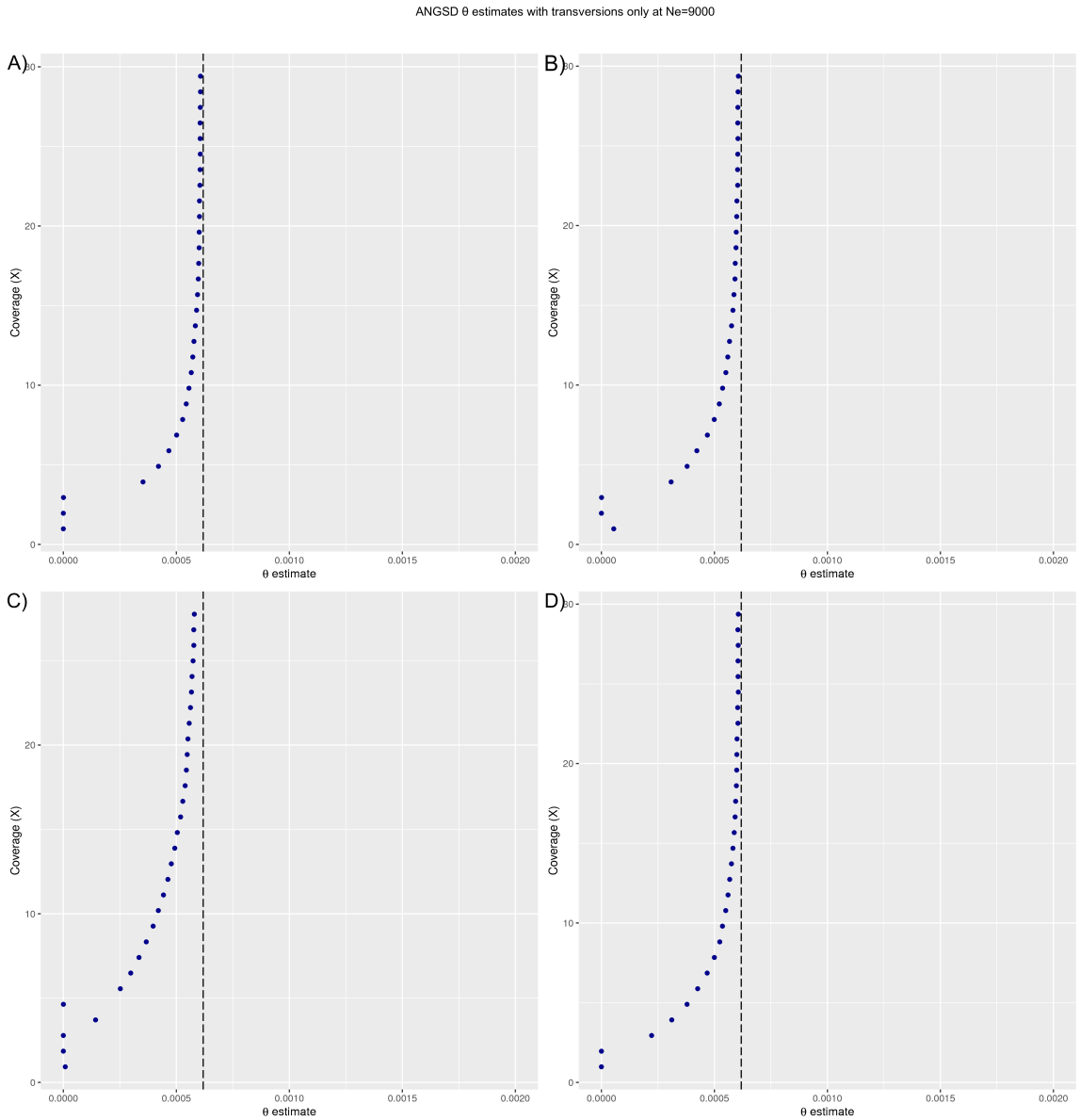

Figure 21: Simulated versus predicted genome-wide θ by ANGSD using transversions only for a simulated chromosome of 15Mbp and an effective population size of 9000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

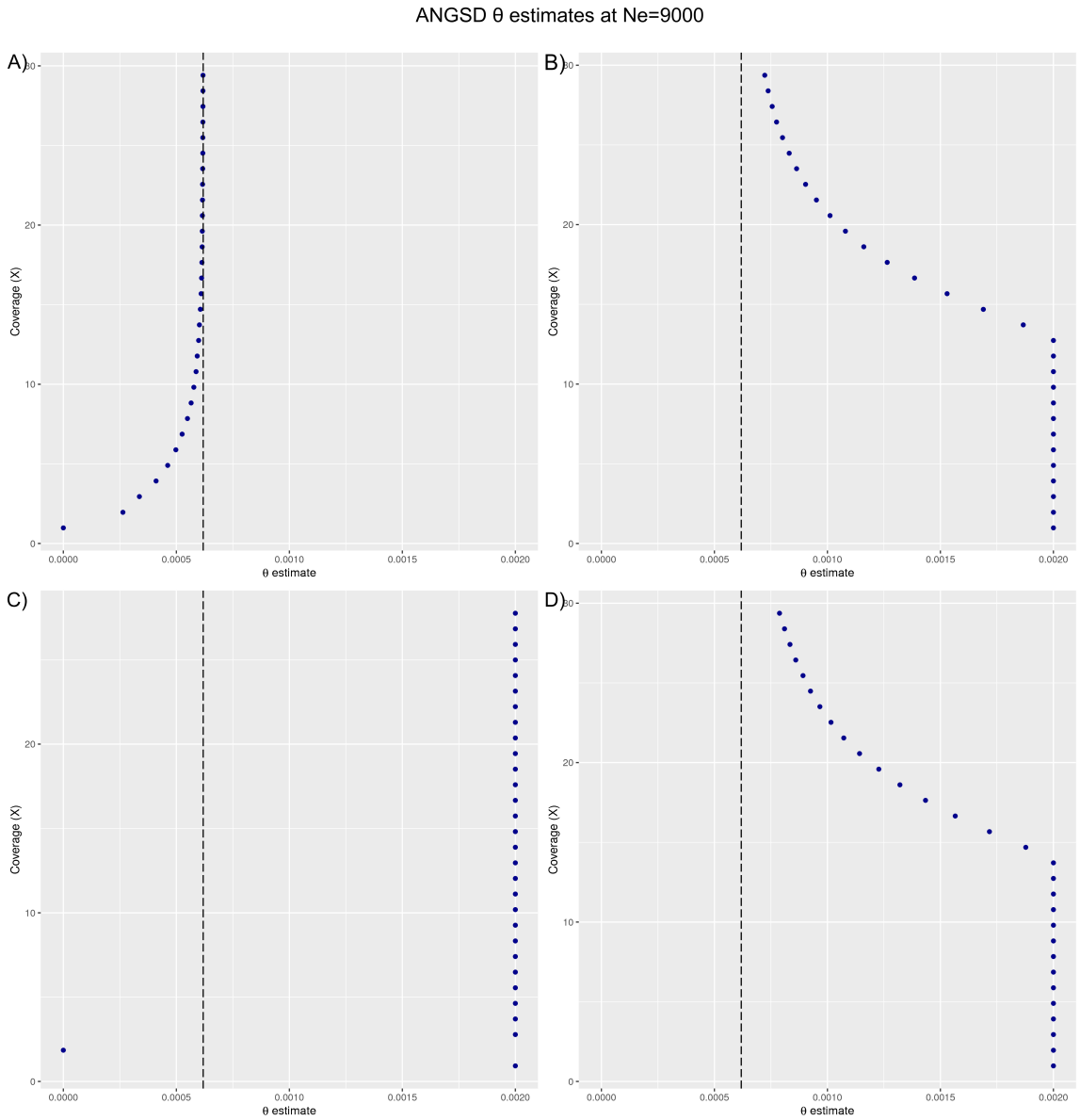

Figure 22: Simulated versus predicted genome-wide θ by ANGSD using options "-tole 10e-12 -maxIter 200" for a simulated chromosome of 15Mbp and an effective population size of 9000 . The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

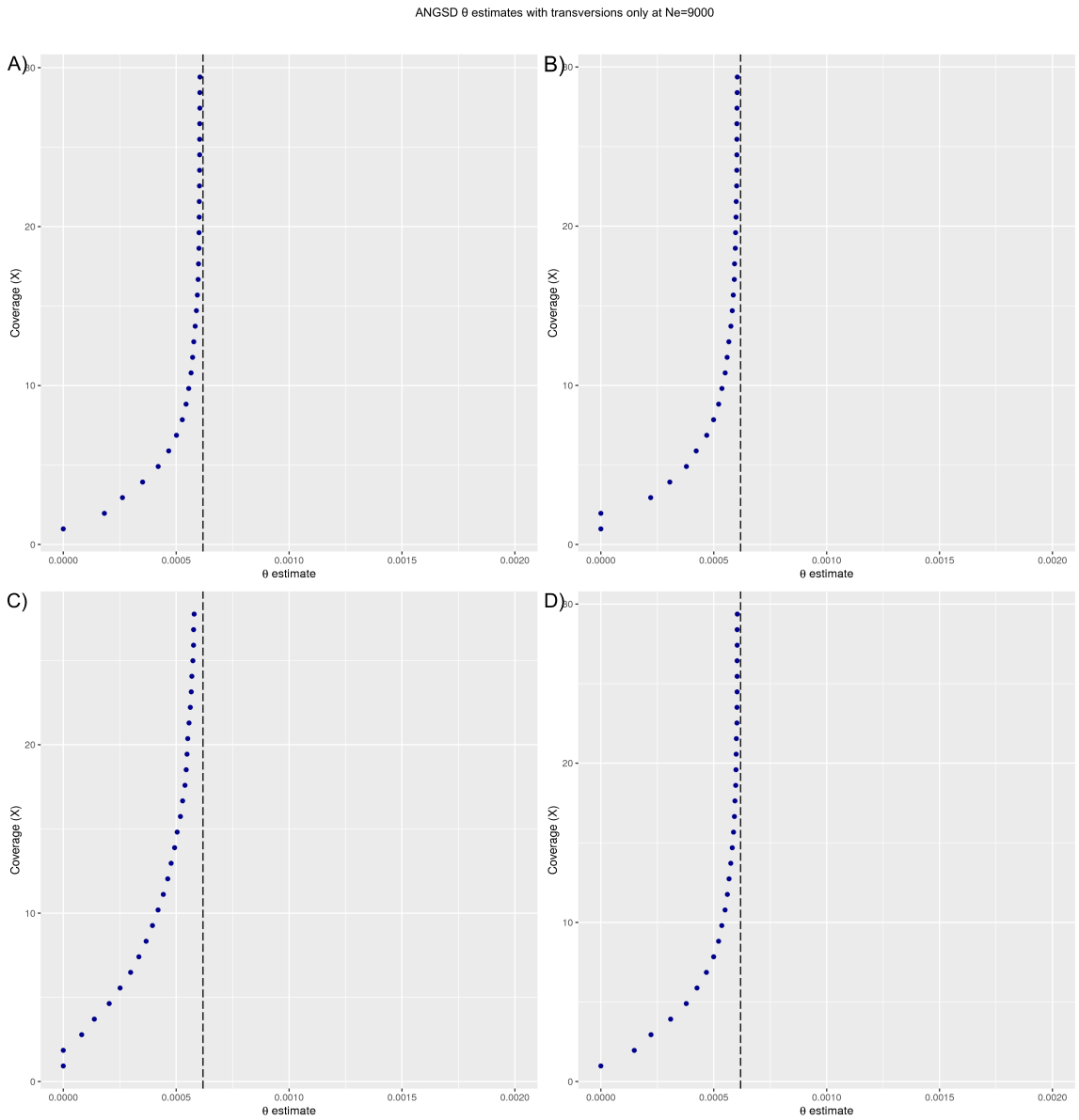

Figure 23: Simulated versus predicted genome-wide θ by ANGSD using transversions and options "-tole 10e-12 -maxIter 200" only for a simulated chromosome of 15Mbp and an effective population size of 9000. The dotted line represented the simulated rate of heterozygosity. The different subpanels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

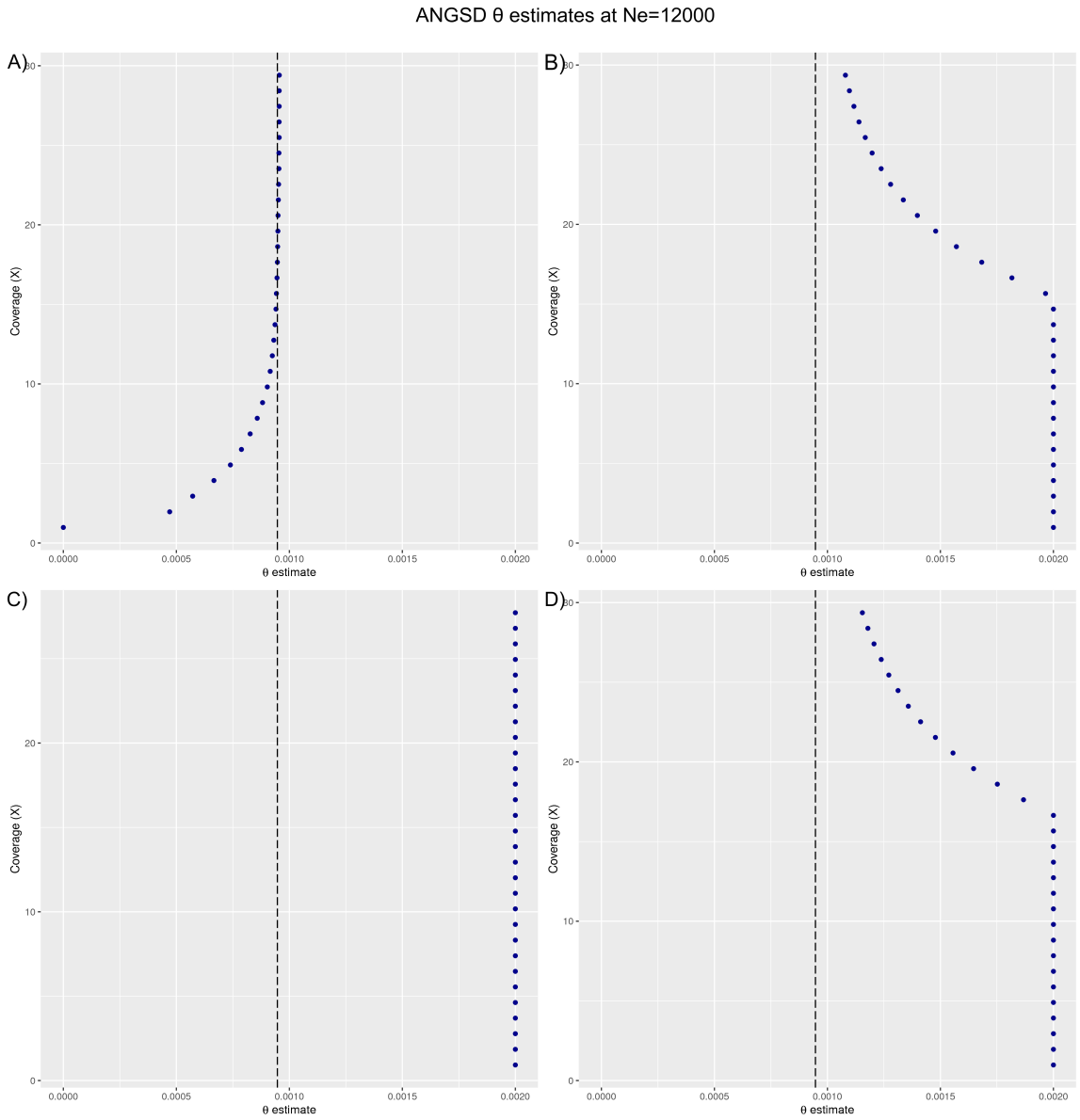

Figure 24: Simulated versus predicted genome-wide θ by ANGSD for a simulated chromosome of 15Mbp and an effective population size of 12000. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

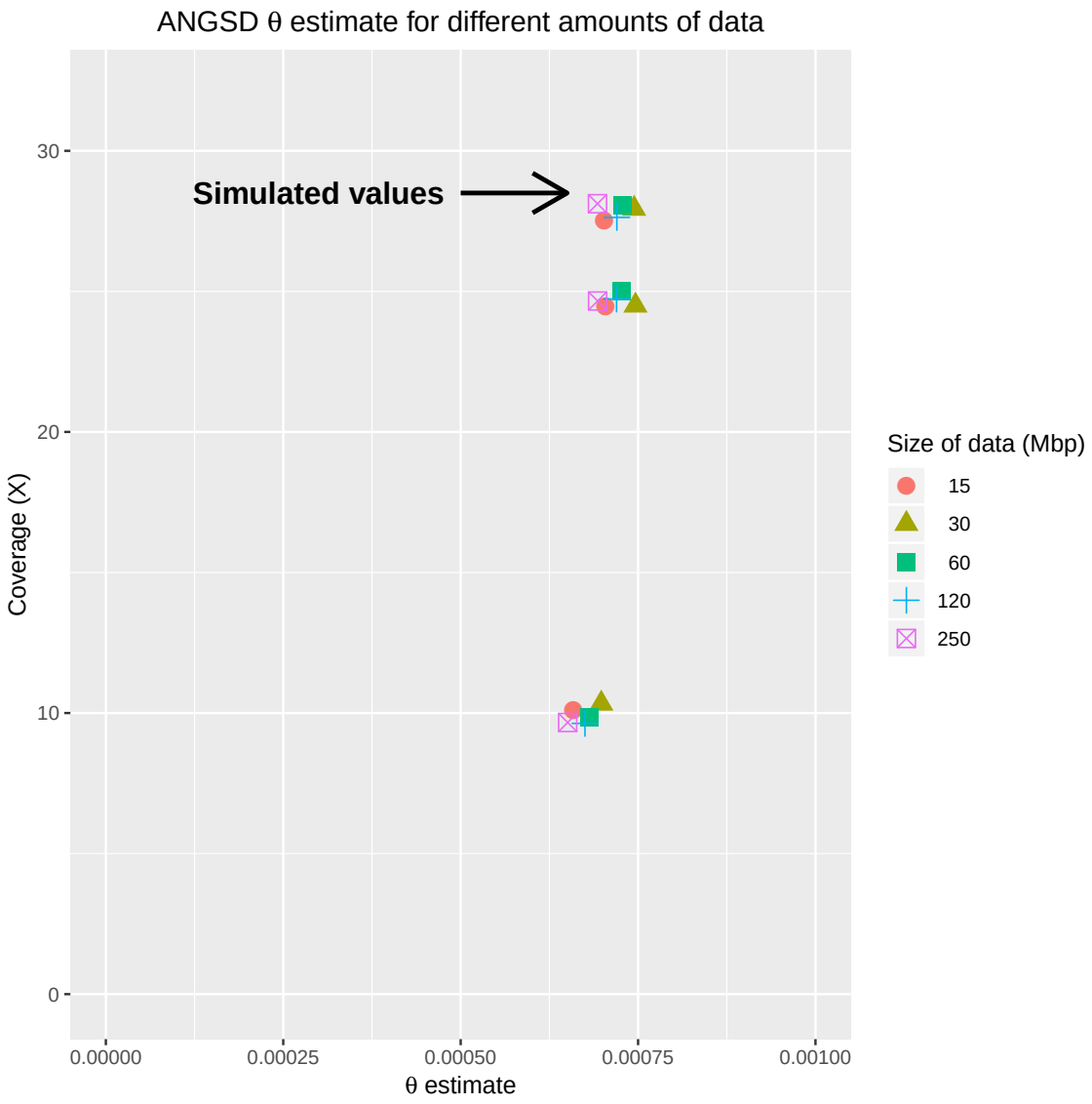

Figure 25: Simulated versus predicted genome-wide θ by ANGSD when a certain subsampling of the data is performed. The length of the chromosome was subsampled as well as the coverage. The simulated values for 15M, 30M, 60M, 120M and 250M are found in the upper portion followed by a subsampling of the coverage at 25X and finally, 10X.

1.1.3 Ignoring deamination from the computation

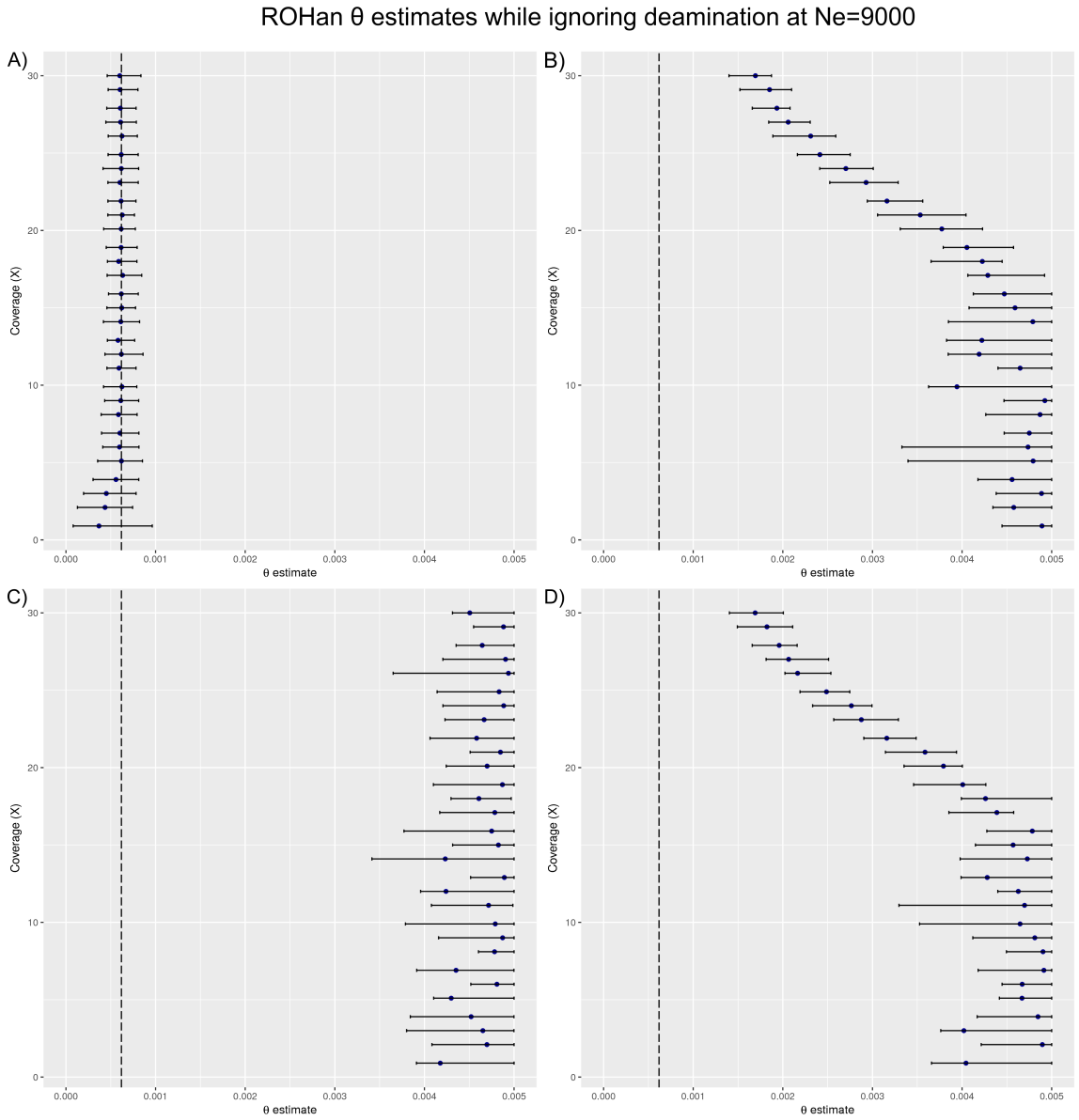

Figure 26: Simulated versus predicted genome-wide θ by ROHan while ignoring deamination in the computation for a simulated chromosome of 15Mbp and an effective population size of 9000. This was evaluated to verify whether ignoring the rates of deamination would have a significant impact. The dotted line represents the measured simulated rate of heterozygosity. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different rates of simulated ancient DNA damage. A) no aDNA damage B) low rates of single-stranded damage from Ust'-Ishim C) high rates of double-stranded damage patterns from ATP2 D) medium rates of double-stranded damage from LaBraña.

1.1.4 Incorrectly inferring deamination rates

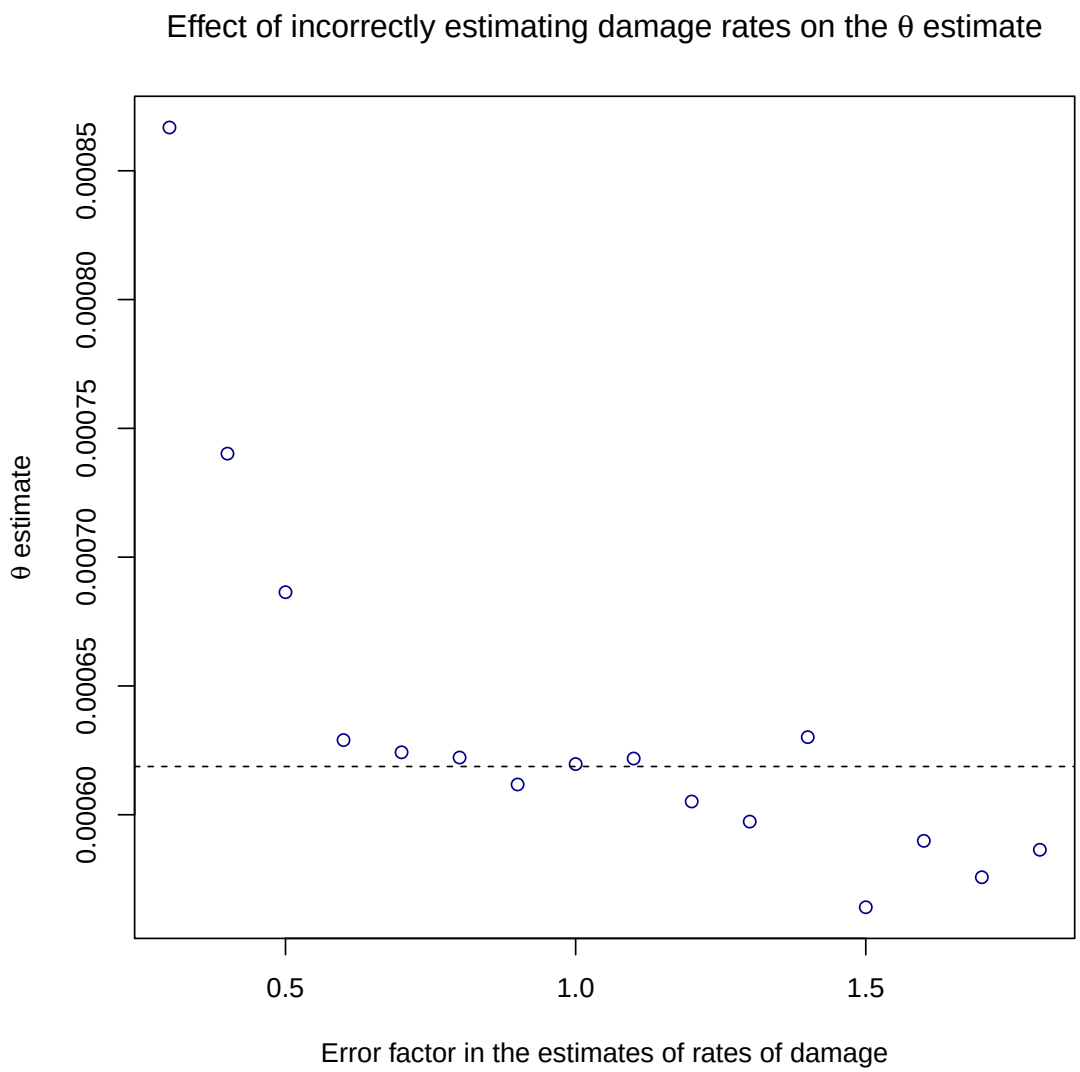

Figure 27: Simulated versus predicted genome-wide θ by ROHan if the incorrect deamination rates are supplied. On a dataset of 15Mbp, a effective population of 9000 was used and the high damage rates from the ATP2 sample were applied. The measured rates of damage were multiplied by a factor (ranging from 0.3 to 1.8) and ROHan was supplied these incorrect rates of damage. The dotted line corresponds to the simulated rate of heterozygosity. As expected, our results show that an overestimate of deamination rates (factor >1.0) causes an underestimate of θ and an underestimate of deamination rates (factor <1.0) causes an overestimate of θ. However, our results show that underestimates ranging from 80% to 120% do not cause a significant error in the estimation of θ. While there is a certain robustness to incorrect estimates of damage, care should be taken while estimating those rates and programs to do so are provided with the software package. Namely, script to mask potentially polymorphic positions is provided and is evaluated on simulated data on page 1.1.5.

1.1.5 Error in inferring deamination rates

coverage	position from the 5' end											position from the 3' end										
	1	2	3	4	5	6	7	8	9	10	RMSD	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	RMSD
1	0.57	0.52	0.54	0.51	0.51	0.50	0.49	0.51	0.46	0.50	0.07	0.54	0.54	0.52	0.50	0.50	0.52	0.51	0.50	0.49	0.51	0.07
2	0.74	0.70	0.71	0.69	0.68	0.67	0.68	0.68	0.66	0.67	0.04	0.73	0.73	0.72	0.71	0.71	0.72	0.72	0.71	0.69	0.70	0.04
3	0.81	0.78	0.78	0.77	0.76	0.76	0.76	0.76	0.74	0.75	0.03	0.81	0.82	0.81	0.80	0.80	0.80	0.80	0.81	0.79	0.79	0.03
4	0.85	0.82	0.81	0.80	0.80	0.79	0.80	0.80	0.78	0.79	0.03	0.84	0.86	0.85	0.84	0.84	0.85	0.84	0.85	0.83	0.84	0.02
5	0.86	0.84	0.83	0.82	0.81	0.81	0.82	0.82	0.80	0.81	0.02	0.86	0.88	0.86	0.86	0.86	0.86	0.86	0.87	0.85	0.86	0.02
6	0.87	0.84	0.84	0.83	0.83	0.82	0.82	0.82	0.81	0.82	0.02	0.87	0.89	0.88	0.87	0.87	0.88	0.87	0.88	0.87	0.88	0.02
7	0.87	0.84	0.84	0.83	0.83	0.83	0.83	0.82	0.81	0.82	0.02	0.88	0.89	0.88	0.87	0.87	0.88	0.88	0.88	0.87	0.88	0.02
8	0.87	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.81	0.82	0.02	0.88	0.89	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.87	0.02
9	0.87	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.81	0.81	0.02	0.88	0.89	0.88	0.87	0.88	0.88	0.87	0.88	0.88	0.87	0.02
10	0.87	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.81	0.81	0.02	0.88	0.89	0.88	0.88	0.88	0.88	0.87	0.88	0.88	0.87	0.02
11	0.87	0.84	0.84	0.83	0.83	0.83	0.82	0.82	0.82	0.82	0.02	0.89	0.89	0.88	0.87	0.88	0.88	0.87	0.88	0.88	0.87	0.02
12	0.87	0.84	0.84	0.83	0.83	0.83	0.82	0.83	0.82	0.82	0.02	0.89	0.89	0.88	0.87	0.88	0.88	0.87	0.87	0.88	0.87	0.02
13	0.87	0.84	0.84	0.83	0.83	0.83	0.81	0.83	0.81	0.82	0.02	0.88	0.89	0.88	0.87	0.88	0.88	0.87	0.87	0.88	0.87	0.02
14	0.87	0.84	0.84	0.83	0.83	0.83	0.81	0.82	0.81	0.81	0.02	0.88	0.89	0.88	0.87	0.88	0.88	0.87	0.88	0.88	0.87	0.02
15	0.86	0.84	0.84	0.82	0.82	0.83	0.81	0.82	0.81	0.81	0.02	0.88	0.89	0.88	0.86	0.88	0.87	0.87	0.88	0.87	0.87	0.02
16	0.86	0.84	0.83	0.82	0.82	0.82	0.80	0.82	0.81	0.81	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.88	0.86	0.02
17	0.86	0.84	0.83	0.82	0.82	0.82	0.80	0.82	0.81	0.81	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.88	0.86	0.02
18	0.86	0.84	0.83	0.82	0.82	0.82	0.80	0.81	0.81	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.86	0.87	0.87	0.86	0.02
19	0.86	0.84	0.83	0.82	0.82	0.81	0.80	0.81	0.81	0.80	0.03	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.87	0.86	0.02
20	0.86	0.84	0.83	0.82	0.82	0.81	0.80	0.81	0.81	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.86	0.87	0.87	0.87	0.02
21	0.86	0.84	0.84	0.82	0.82	0.82	0.80	0.80	0.80	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.02
22	0.86	0.84	0.83	0.82	0.82	0.82	0.80	0.81	0.80	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.02
23	0.86	0.84	0.83	0.82	0.82	0.82	0.80	0.80	0.80	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.02
24	0.86	0.84	0.84	0.82	0.82	0.82	0.80	0.81	0.80	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.02
25	0.86	0.84	0.84	0.82	0.83	0.82	0.80	0.80	0.80	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.02
26	0.86	0.84	0.83	0.82	0.83	0.82	0.80	0.80	0.79	0.79	0.02	0.88	0.89	0.87	0.86	0.87	0.86	0.87	0.87	0.87	0.87	0.02
27	0.86	0.84	0.83	0.82	0.82	0.82	0.81	0.81	0.79	0.80	0.02	0.88	0.89	0.87	0.86	0.87	0.86	0.87	0.88	0.87	0.88	0.02

Table 1: Error made at different rate of subsampling in the estimate by masking polymorphic positions of C to T substitution at the 5' end and the G to A substitutions for the simulated data using the high rates of misincorporations from the ATP2 sample.The estimate of substitutions was computed by masking potentially polymorphic and was performed using a script provided with the software package. The number reported is the ratio of the deamination rate found at that position to the one simulated. The consistent underestimate is likely due to mapping issues of the heavily deaminated aDNA fragments. RMSD stands for root-mean-square deviation.

coverage	position from the 5' end											position from the 3' end										
	1	2	3	4	5	6	7	8	9	10	RMSD	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	RMSD
1	0.66	0.63	0.64	0.62	0.62	0.61	0.61	0.63	0.59	0.62	0.05	0.61	0.61	0.59	0.58	0.57	0.61	0.59	0.57	0.56	0.58	0.05
2	0.86	0.84	0.85	0.84	0.83	0.83	0.84	0.85	0.83	0.84	0.02	0.83	0.83	0.82	0.83	0.81	0.84	0.83	0.81	0.80	0.80	0.02
3	0.95	0.93	0.93	0.94	0.92	0.93	0.94	0.94	0.93	0.94	0.01	0.92	0.92	0.93	0.93	0.92	0.93	0.93	0.92	0.92	0.91	0.01
4	0.99	0.98	0.97	0.98	0.97	0.97	0.98	0.99	0.99	0.99	0.00	0.96	0.97	0.97	0.98	0.96	0.98	0.97	0.97	0.96	0.95	0.00
5	1.01	1.00	0.99	1.00	0.99	0.99	1.02	1.01	1.01	1.02	0.00	0.98	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.98	0.98	0.00
6	1.02	1.01	1.00	1.01	1.01	1.01	1.02	1.02	1.02	1.02	0.00	1.00	1.00	1.01	1.01	1.00	1.02	1.00	1.00	1.00	1.00	0.00
7	1.02	1.01	1.00	1.01	1.01	1.02	1.02	1.02	1.02	1.02	0.00	1.00	1.00	1.01	1.01	1.00	1.02	1.01	1.00	1.01	1.00	0.00
8	1.02	1.01	1.00	1.01	1.01	1.02	1.02	1.02	1.02	1.03	0.00	1.01	1.00	1.02	1.01	1.01	1.02	1.01	1.00	1.01	0.99	0.00
9	1.02	1.01	1.00	1.01	1.01	1.02	1.02	1.02	1.03	1.02	0.00	1.01	1.00	1.01	1.01	1.01	1.02	1.01	1.00	1.02	1.00	0.00
10	1.02	1.00	1.01	1.01	1.01	1.02	1.01	1.02	1.03	1.02	0.00	1.01	1.00	1.01	1.01	1.01	1.02	1.01	1.00	1.01	1.00	0.00
11	1.01	1.01	1.01	1.01	1.01	1.02	1.01	1.02	1.03	1.02	0.00	1.01	1.01	1.01	1.01	1.01	1.02	1.01	1.00	1.02	0.99	0.00
12	1.01	1.01	1.01	1.01	1.01	1.02	1.01	1.03	1.03	1.02	0.00	1.01	1.01	1.01	1.01	1.01	1.02	1.01	1.00	1.02	0.99	0.00
13	1.01	1.01	1.00	1.01	1.01	1.02	1.01	1.03	1.03	1.03	0.00	1.01	1.01	1.01	1.00	1.01	1.02	1.01	1.00	1.02	0.99	0.00
14	1.01	1.01	1.00	1.00	1.00	1.02	1.01	1.02	1.02	1.02	0.00	1.01	1.00	1.01	1.00	1.01	1.02	1.01	1.00	1.01	0.99	0.00
15	1.01	1.00	1.00	1.00	1.00	1.01	1.00	1.02	1.02	1.02	0.00	1.01	1.00	1.01	1.00	1.01	1.01	1.00	1.00	1.01	0.99	0.00
16	1.01	1.00	1.00	1.00	1.00	1.01	0.99	1.01	1.02	1.02	0.00	1.00	1.01	1.00	1.00	1.00	1.01	1.00	1.00	1.01	0.99	0.00
17	1.00	1.00	1.00	0.99	1.00	1.01	0.99	1.01	1.03	1.02	0.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	0.99	1.01	0.98	0.00
18	1.00	1.00	1.00	0.99	1.00	1.01	0.99	1.01	1.03	1.01	0.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	1.00	1.00	0.98	0.00
19	1.00	1.00	1.00	0.99	0.99	1.00	0.99	1.00	1.02	1.00	0.00	1.00	1.00	1.00	1.00	0.99	1.01	1.00	0.99	1.00	0.99	0.00
20	1.00	1.00	1.00	0.99	0.99	1.00	0.99	1.00	1.02	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	0.99	1.00	0.99	0.00
21	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.01	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.00
22	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.01	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	0.99	1.00	0.99	0.00
23	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.01	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	0.99	1.00	0.99	0.00
24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.00
25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.00
26	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Table 2: Error made at different rate of subsampling in the estimate of C to T substitution at the 5' end and the G to A substitutions for the simulated data using the high rates of misincorporations from the ATP2 sample. The estimate of substitutions was computed by masking potentially polymorphic and was performed using a script provided with the software package. The number reported is the ratio of the deamination rate found at that position to the one found at 27X at the same position. RMSD stands for root-mean-square deviation.

coverage	$\theta \times 10^4$	$\theta_{low} \times 10^4$	$\theta_{high} \times 10^4$
1	47.65	41.20	50.00
2	32.08	20.12	39.67
3	8.75	2.65	16.97
4	6.13	2.46	10.69
5	6.38	2.90	10.10
6	6.06	3.43	9.37
6	5.99	4.05	8.66
7	6.14	4.30	8.24
8	5.90	4.43	8.06
9	5.92	4.31	8.18
10	6.06	4.41	8.43
11	6.15	4.13	7.94
12	5.78	4.39	8.54
13	6.00	4.21	8.62
14	6.53	4.11	8.55
15	5.91	3.75	8.19
16	5.91	4.41	8.05
17	6.13	4.28	7.87
18	5.93	4.54	7.45
18	6.04	4.50	8.51
19	5.81	4.66	7.73
20	6.13	4.44	8.11
21	6.07	4.38	8.47
22	6.21	4.54	7.98
23	5.98	4.36	7.65
24	6.20	4.40	7.92
25	6.19	4.41	7.76
26	6.15	4.58	8.13
27	6.08	4.21	8.26

Table 3: Predicted θ using ROHan on simulated sample of 15M using an effective population size of 9000. The aDNA damage was simulated using the high rates of misincorporations from the ATP2 sample. The simulated θ for this dataset was of 6.19 segregating sites per 10^4. Damage patterns were evaluated using a script provided with the software package which masks potentially polymorphic sites. The underestimate in estimating aDNA damage seen at coverage 1X-3X (see Supplementary Table 3) causes overestimates. Currently, our method cannot estimate substitutions due to aDNA damage highly deaminated samples at 1X-3X while masking potentially polymorphic positions.

1.1.6 Simulating multiple libraries with different damage rates

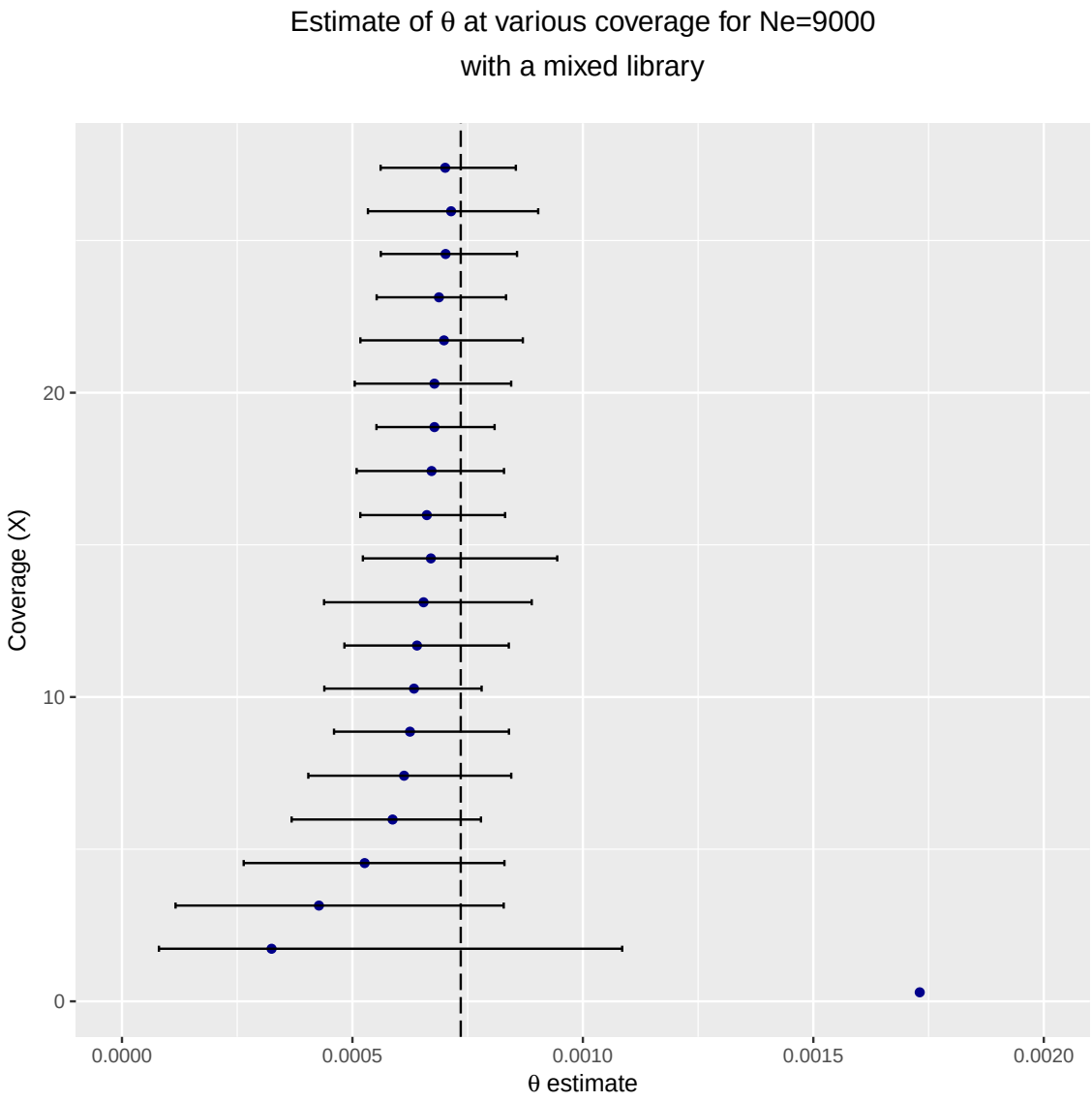

Figure 28: Simulated versus predicted genome-wide θ by ROHan for a simulated dataset which was composed of a 50%/50% blend of a highly deaminated library from the ATP2 sample and a non-deaminated one. Damage rates were evaluated on this new dataset and were intermediate between the damage rates of the 2 original datasets.

1.1.7 Different window sizes

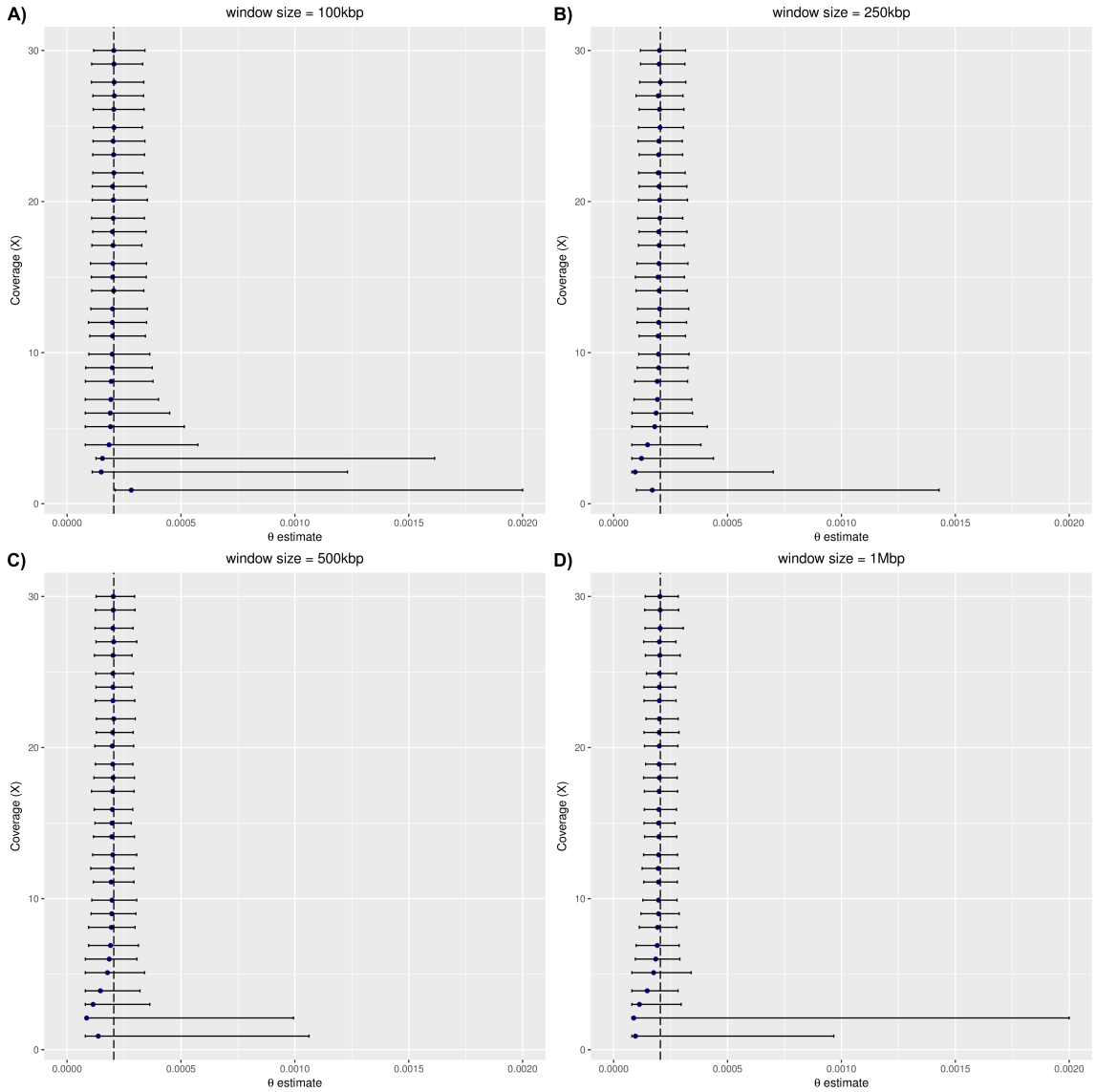

Figure 29: Effect of using different windows for the estimation of local heterozygosity on the estimate for the genome-wide estimate of θ. No aDNA damage was added and an effective population size of 3000 was used. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different window sizes A) 1kbp B) 2.5kbp C) 5kbp D) 1Mbp.

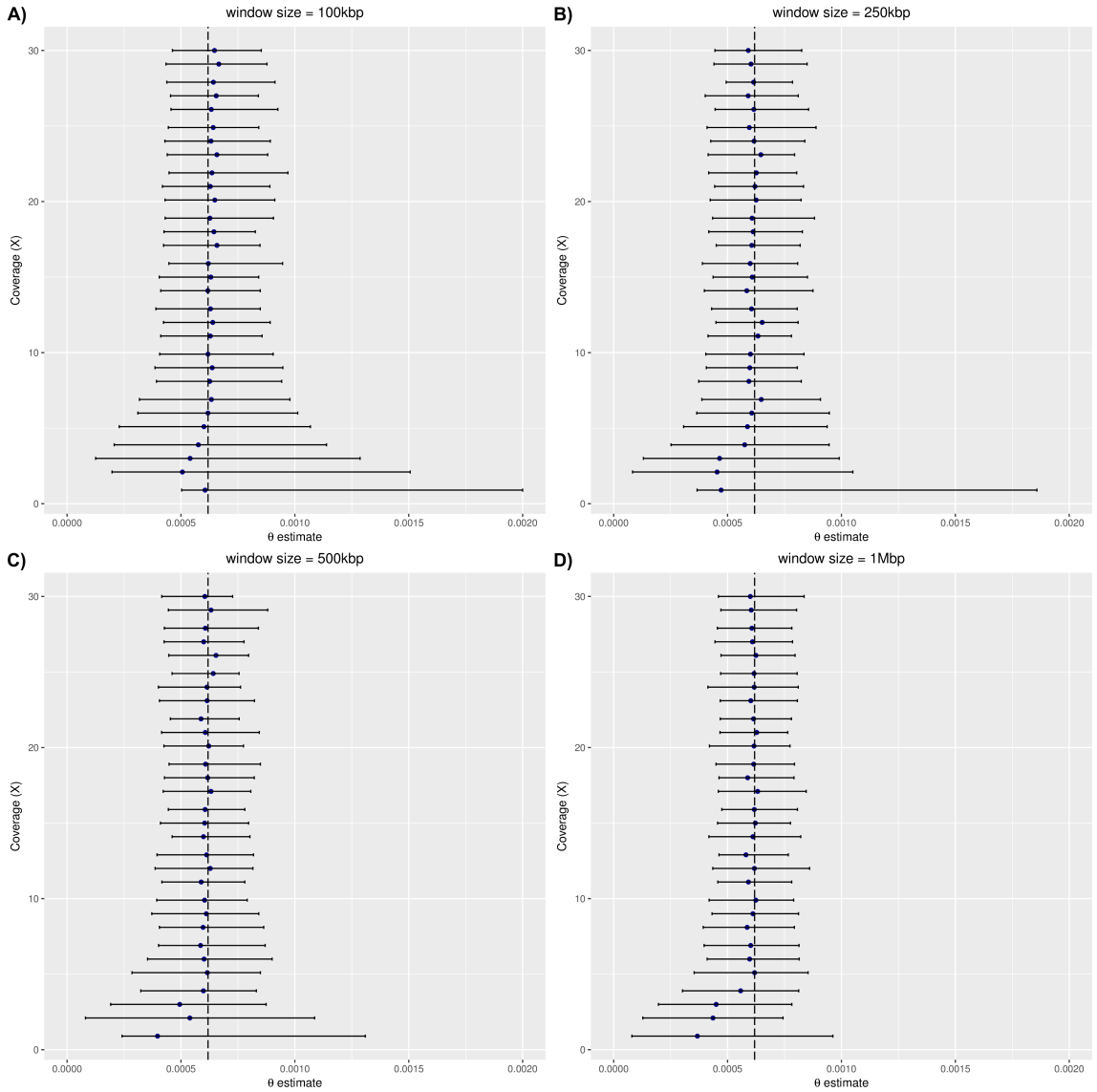

Figure 30: Effect of using different windows for the estimation of local heterozygosity on the estimate for the genome-wide estimate of θ. No aDNA damage was added and an effective population size of 9000 was used. The dotted line represented the simulated rate of heterozygosity. The different sub-panels represent different window sizes A) 1kbp B) 2.5kbp C) 5kbp D) 1Mbp.

1.1.8 High sequencing error rate

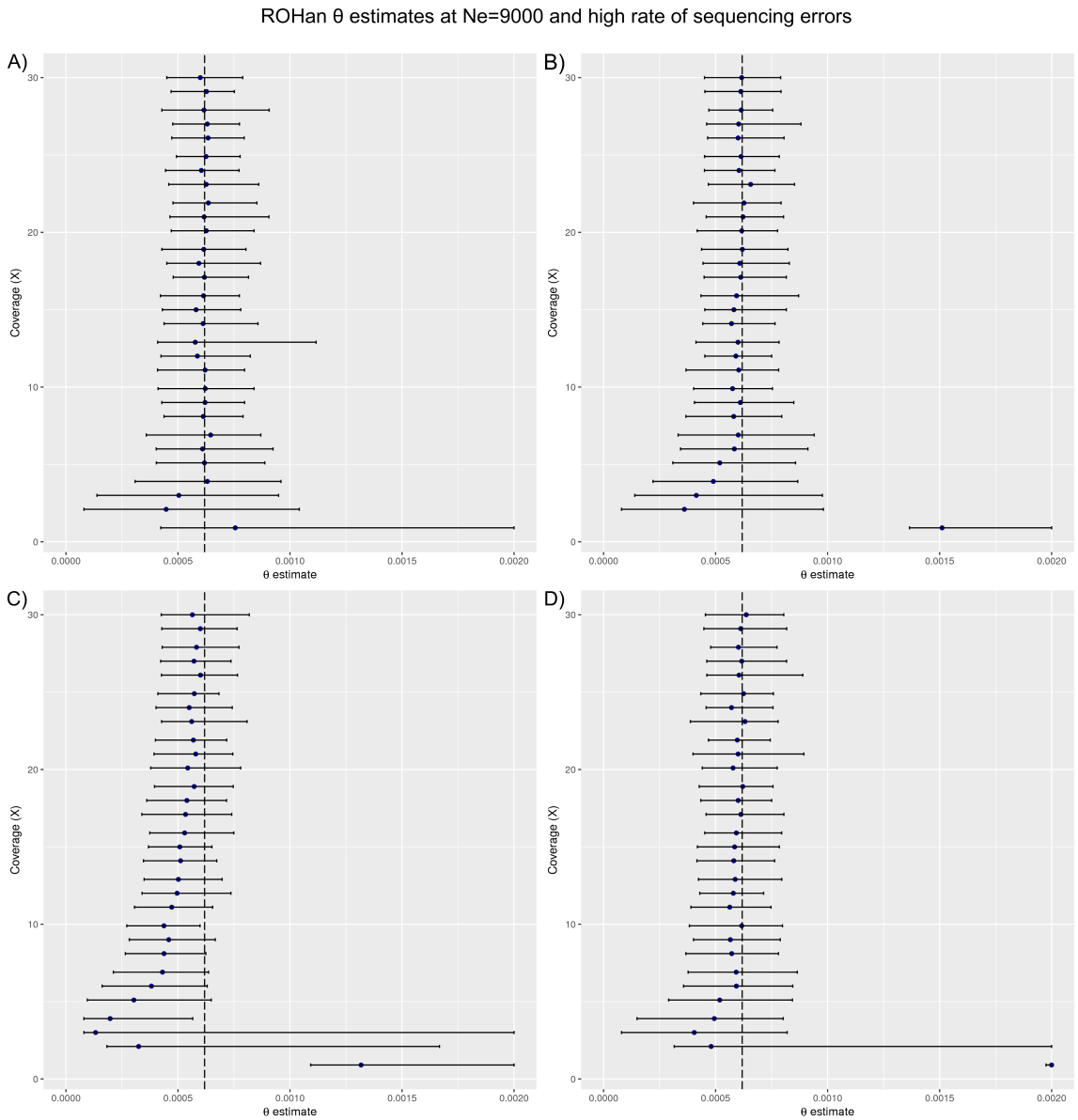

Figure 31: Robustness of our methodology for inferring heterozygosity rates to a substantial increase in sequencing errors. We increased the amount of simulated sequencing errors 10-fold to reach a probability of error of 1.6% (please refer to Appendix Table 1). The amount of ancient DNA damage was the same as in previous sections: A) no simulated damage due to deamination B) damage levels from the Ust'-Ishim sample, which contains a low rate of misincorporations and followed patterns corresponding to a single-stranded library building protocol C) damage levels from the APT2 sample, which contains a high rate of misincorporations and followed patterns corresponding to a double-stranded library building protocol D) damage levels from the LaBraña sample, which contains a medium rate of misincorporations and followed patterns corresponding to a double-stranded library building protocol.

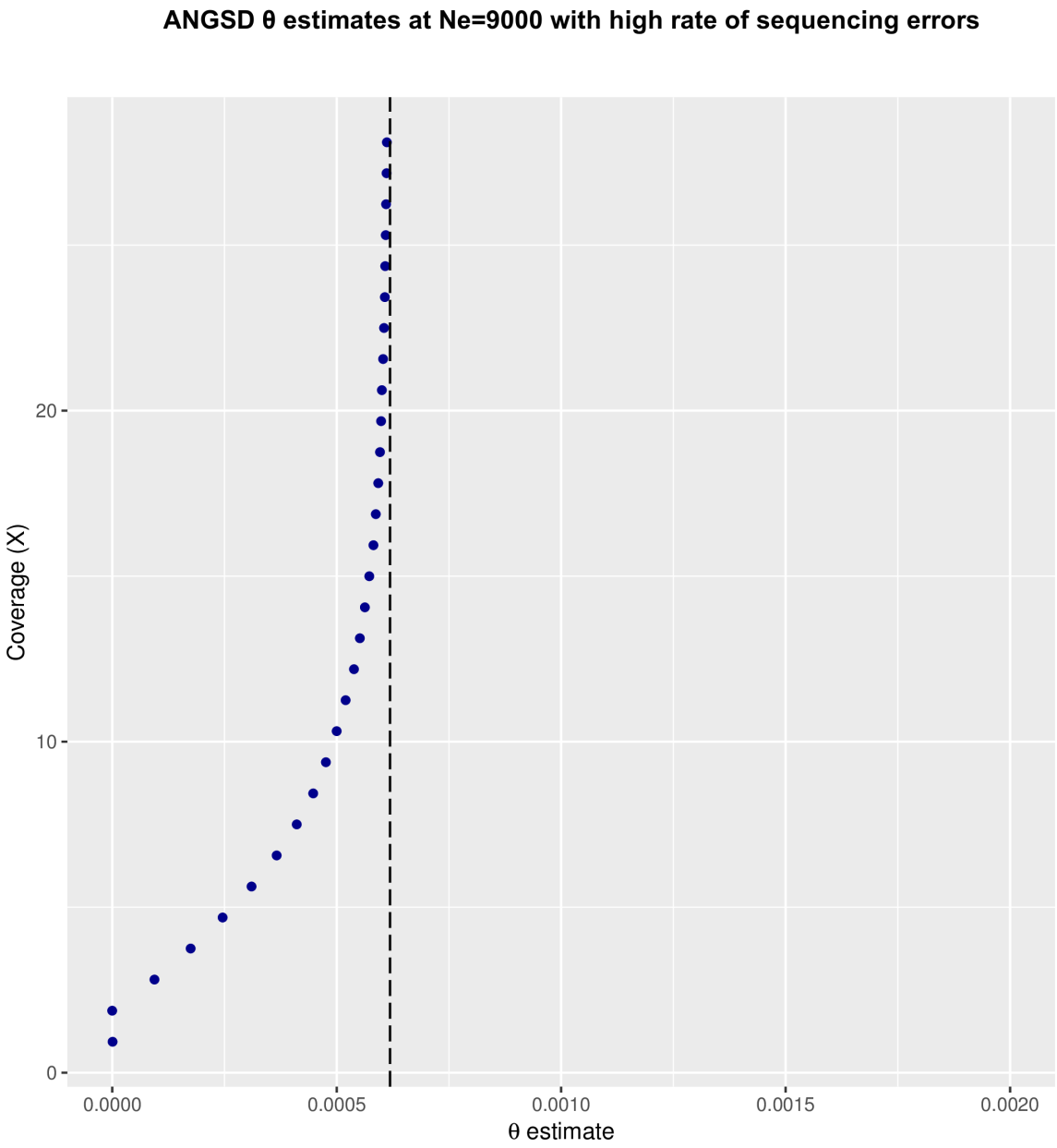

Figure 32: Robustness of ANGSD θ estimate to a substantial increase in sequencing errors without any additional simulated deamination. We increased the amount of simulated sequencing errors 10-fold to reach a probability of error of 1.6% (please refer to Appendix Table 1). The results for this dataset without additional sequencing errors is found in Supplementary Figure 20A).

1.1.9 Identifying runs of homozygosity

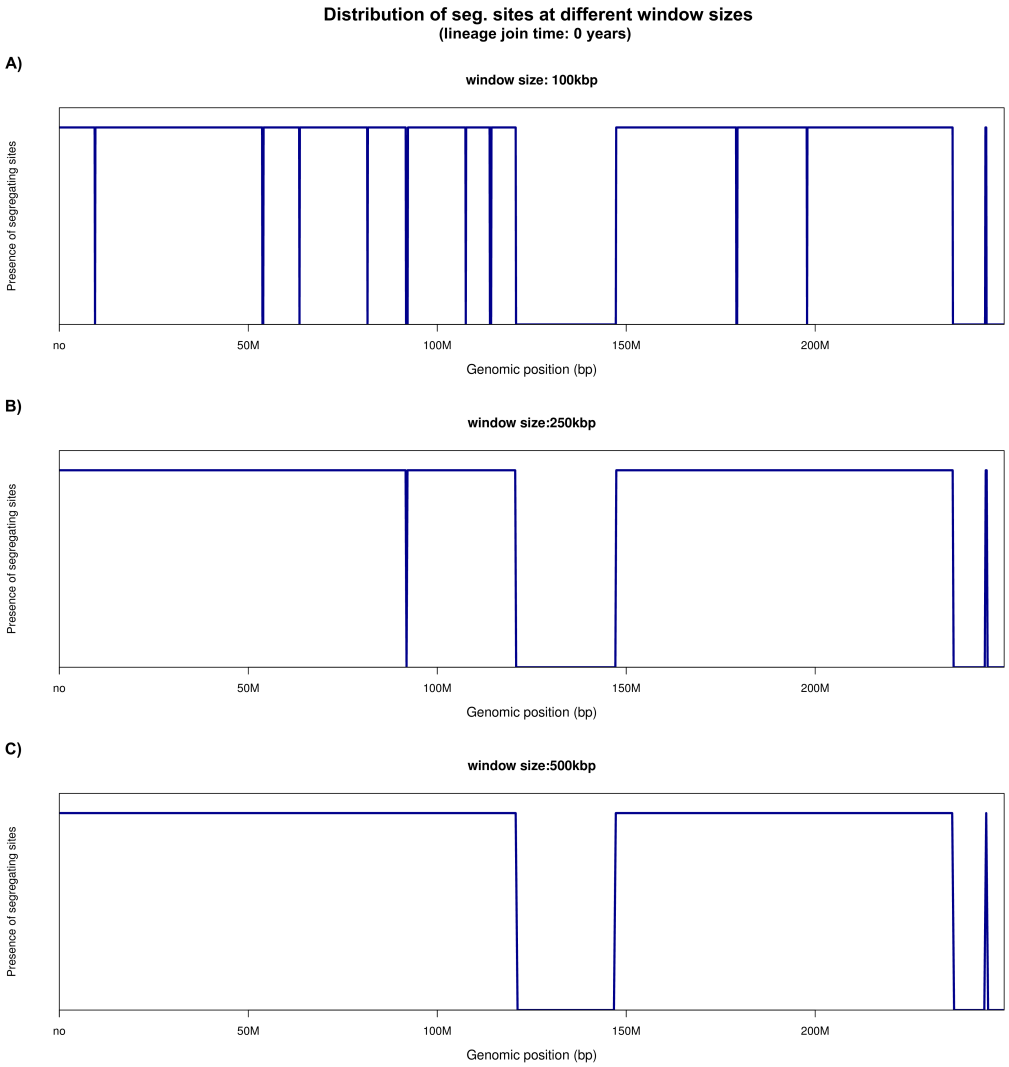

Figure 33: Presence or absence of segregating sites on the simulated chromosomes using windows of A) 1kbp B) 2.5kbp C) 5kbp using the inbreeding scenario 1 (inbreeding between siblings). For the evaluation of BCFtools/RoH, the lineages between the 16 chromosomes to form the grand-parents chromosomes and the 1000 chromosomes which provide allele frequencies is at 0 years.

1.1.9.1 ROHan

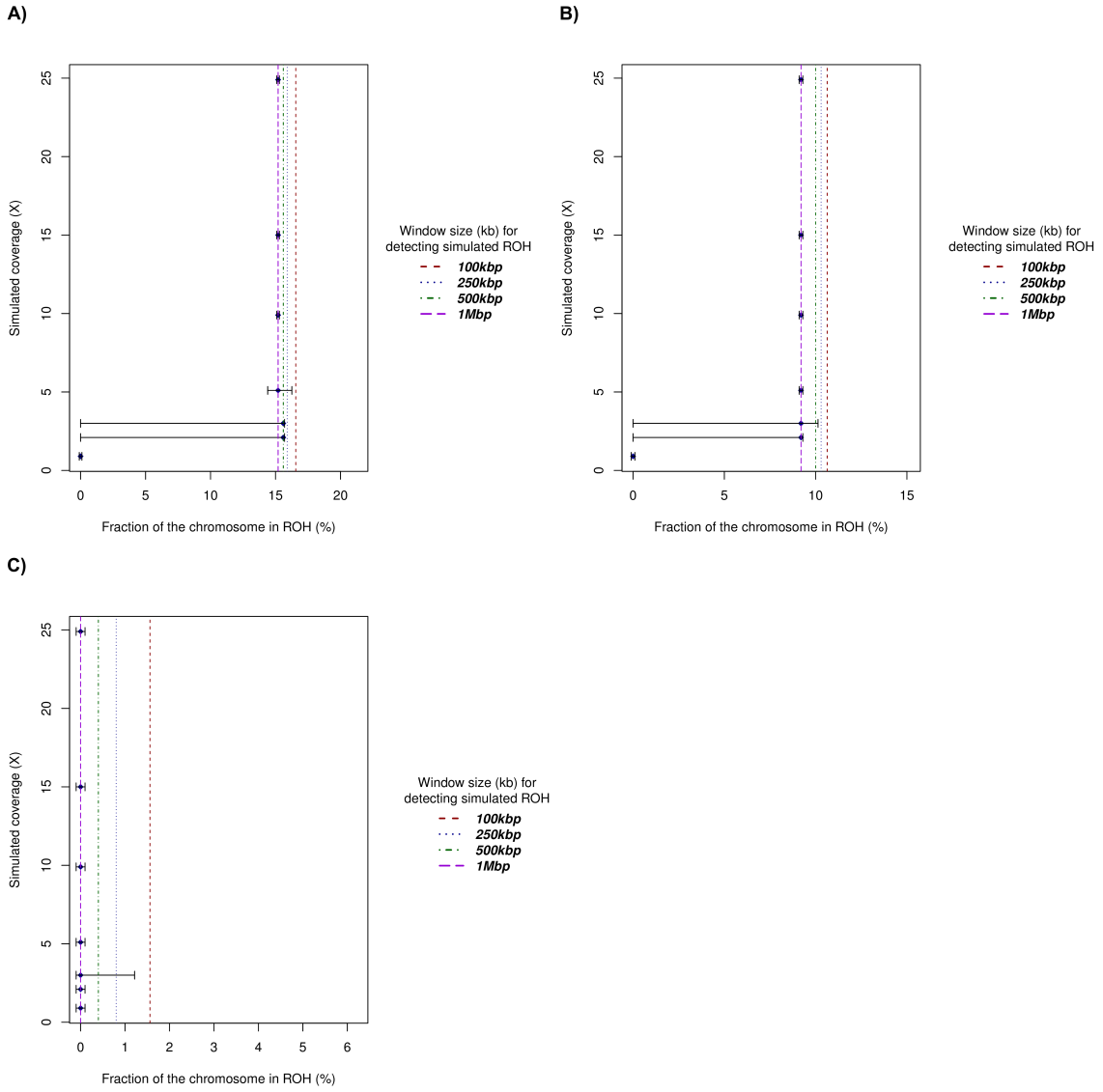

Figure 34: ROHan's accuracy in predicting the percentage of genomic regions in an ROH for a chromosome of 250Mbp using A) inbreeding scenario 1 (inbreeding between siblings) B) inbreeding scenario 2 (inbreeding between a grandparent and a grandchild) C) using inbreeding scenario 3 (inbreeding between first cousins). As coverage increases, the greater the accuracy in predicting ROHs. ROHan was used with a window of 1Mbp for the local heterozygosity estimates. The different dotted lines represent the measured percentage of genomic windows in an ROH at different genomic window sizes. The blue dots represent the maximum-likelihood point estimate, the black whiskers represent the 95% confidence interval.

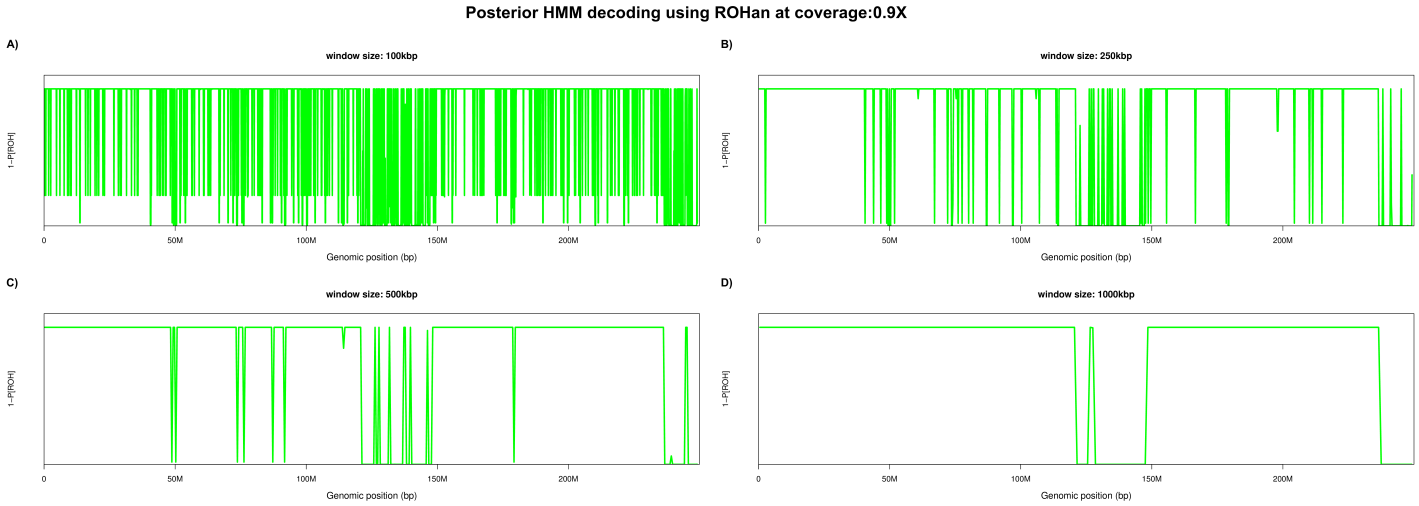

Figure 35: Posterior decoding using ROHan at different window sizes for the computation of local heterozygosity. The average simulated coverage was of 0.9X. The window sizes were A) 100kbp B) 250kbp C) 500 kbp D) 1Mbp. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome.

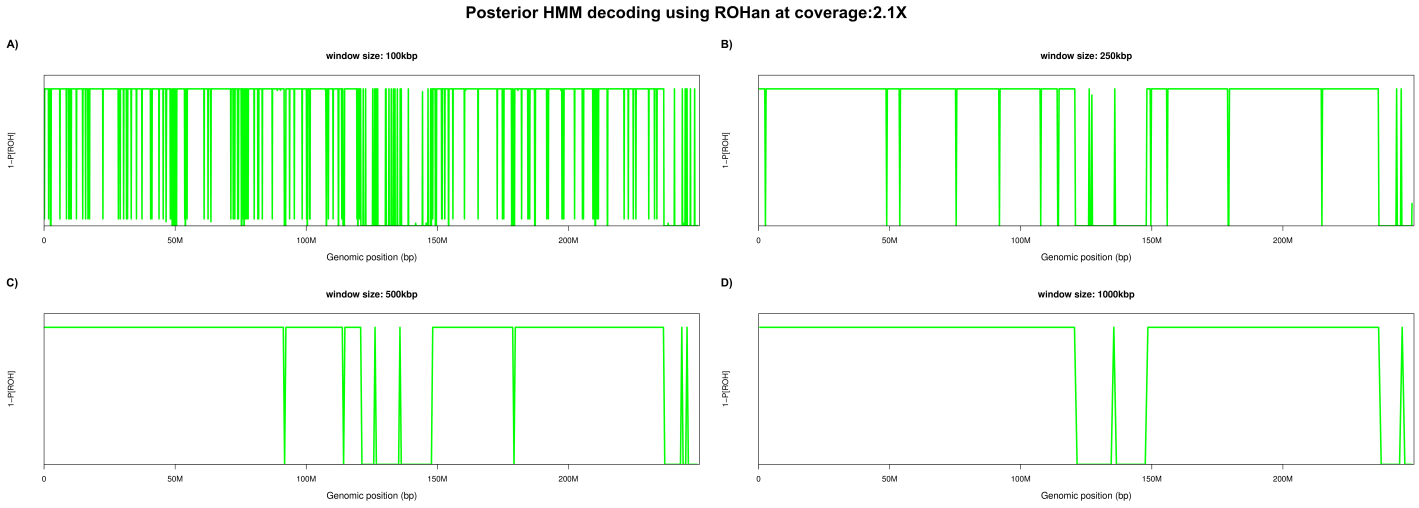

Figure 36: Posterior decoding using ROHan at different window sizes for the computation of local heterozygosity. The average simulated coverage was of 2.1X. The window sizes were A) 100kbp B) 250kbp C) 500 kbp D) 1Mbp. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome.

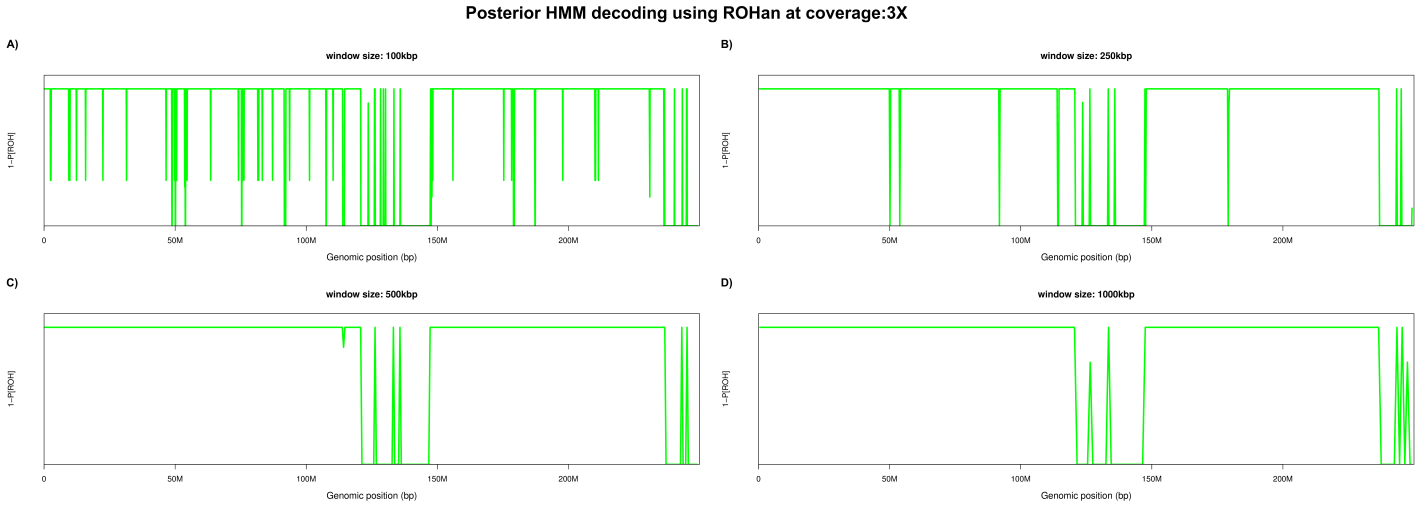

Figure 37: Posterior decoding using ROHan at different window sizes for the computation of local heterozygosity. The average simulated coverage was of 3.0X. The window sizes were A) 100kbp B) 250kbp C) 500 kbp D) 1Mbp. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome.

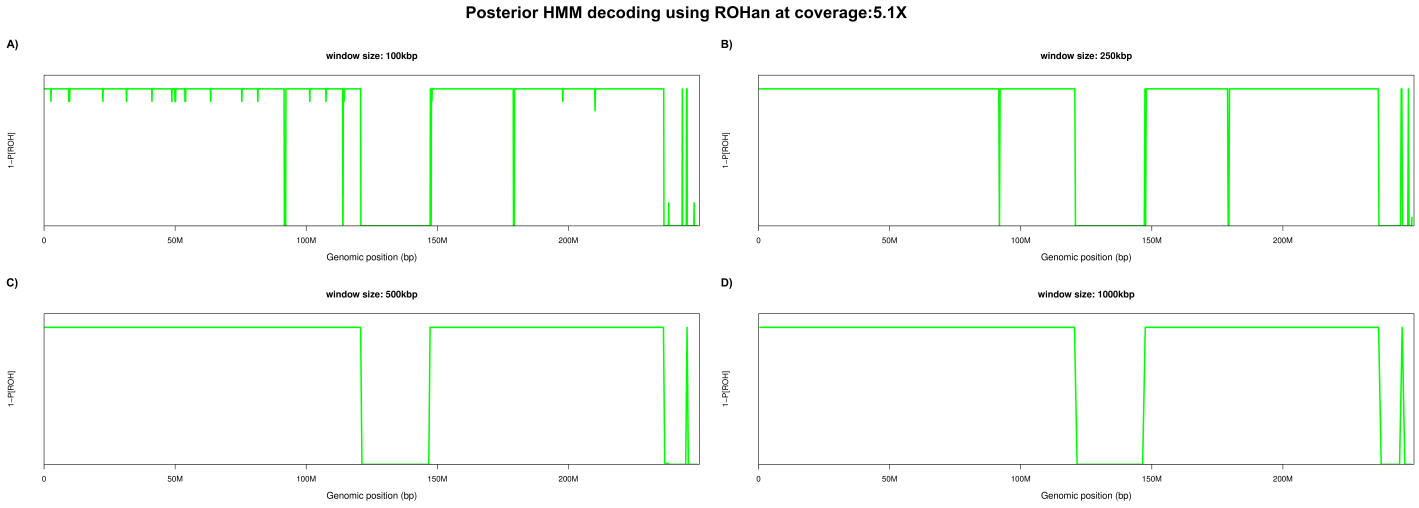

Figure 38: Posterior decoding using ROHan at different window sizes for the computation of local heterozygosity. The average simulated coverage was of 5.1X. The window sizes were A) 100kbp B) 250kbp C) 500 kbp D) 1Mbp. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome.

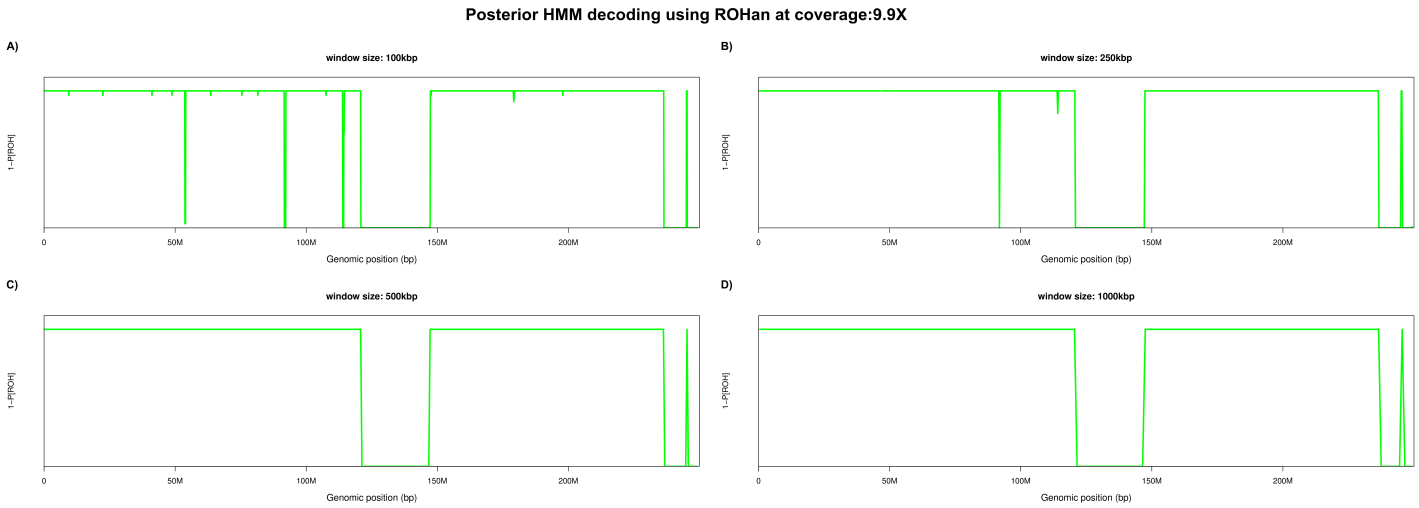

Figure 39: Posterior decoding using ROHan at different window sizes for the computation of local heterozygosity. The average simulated coverage was of 9.9X. The window sizes were A) 100kbp B) 250kbp C) 500 kbp D) 1Mbp. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome.

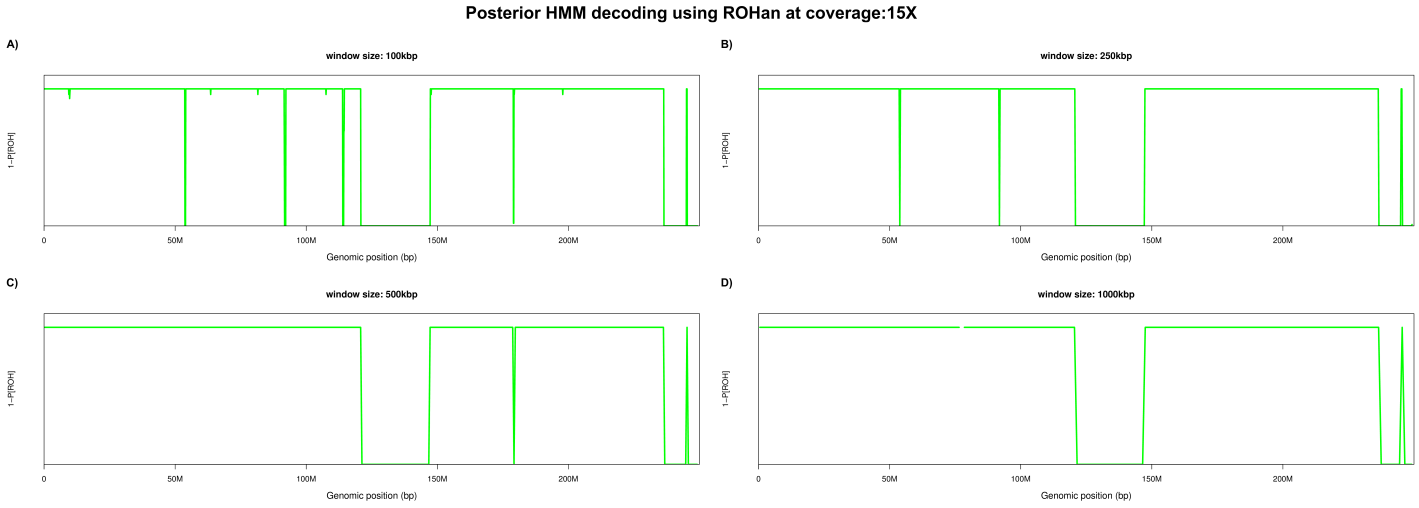

Figure 40: Posterior decoding using ROHan at different window sizes for the computation of local heterozygosity. The average simulated coverage was of 15X. The window sizes were A) 100kbp B) 250kbp C) 500 kbp D) 1Mbp. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome.

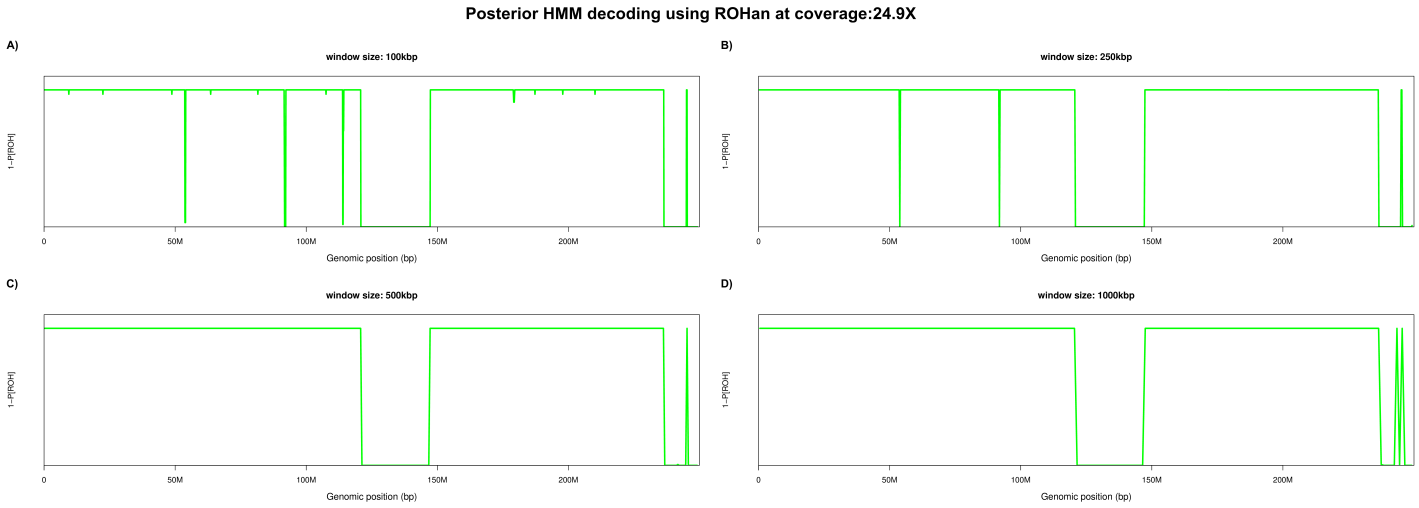

Figure 41: Posterior decoding using ROHan at different window sizes for the computation of local heterozygosity. The average simulated coverage was of 24.3X. The window sizes were A) 100kbp B) 250kbp C) 500 kbp D) 1Mbp. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome.

1.1.9.2 PLINK

1.1.9.3 BCFtools/RoH

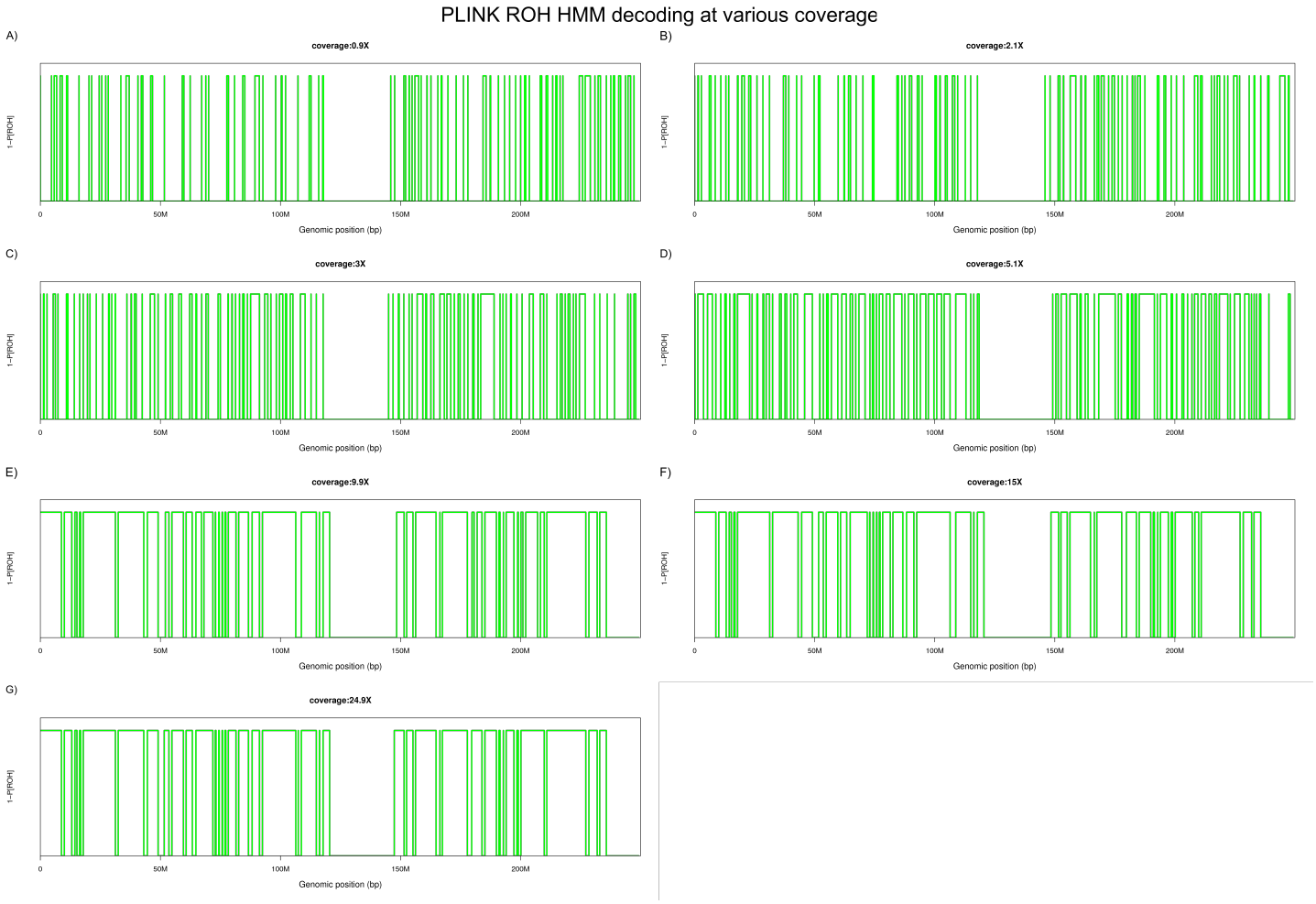

Figure 42: Posterior decoding using PLINK at different levels of simulated coverage namely: A) 0.9X B) 2.1X C) 3.0X D) 5.1X E) 9.9X F) 15X G) 24.3X. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome. The lineages of the 16 chromosomes to form the grand-parents' chromosomes and the 1000 chromosomes which provide the allele frequency were jointed at a time of 0 years.

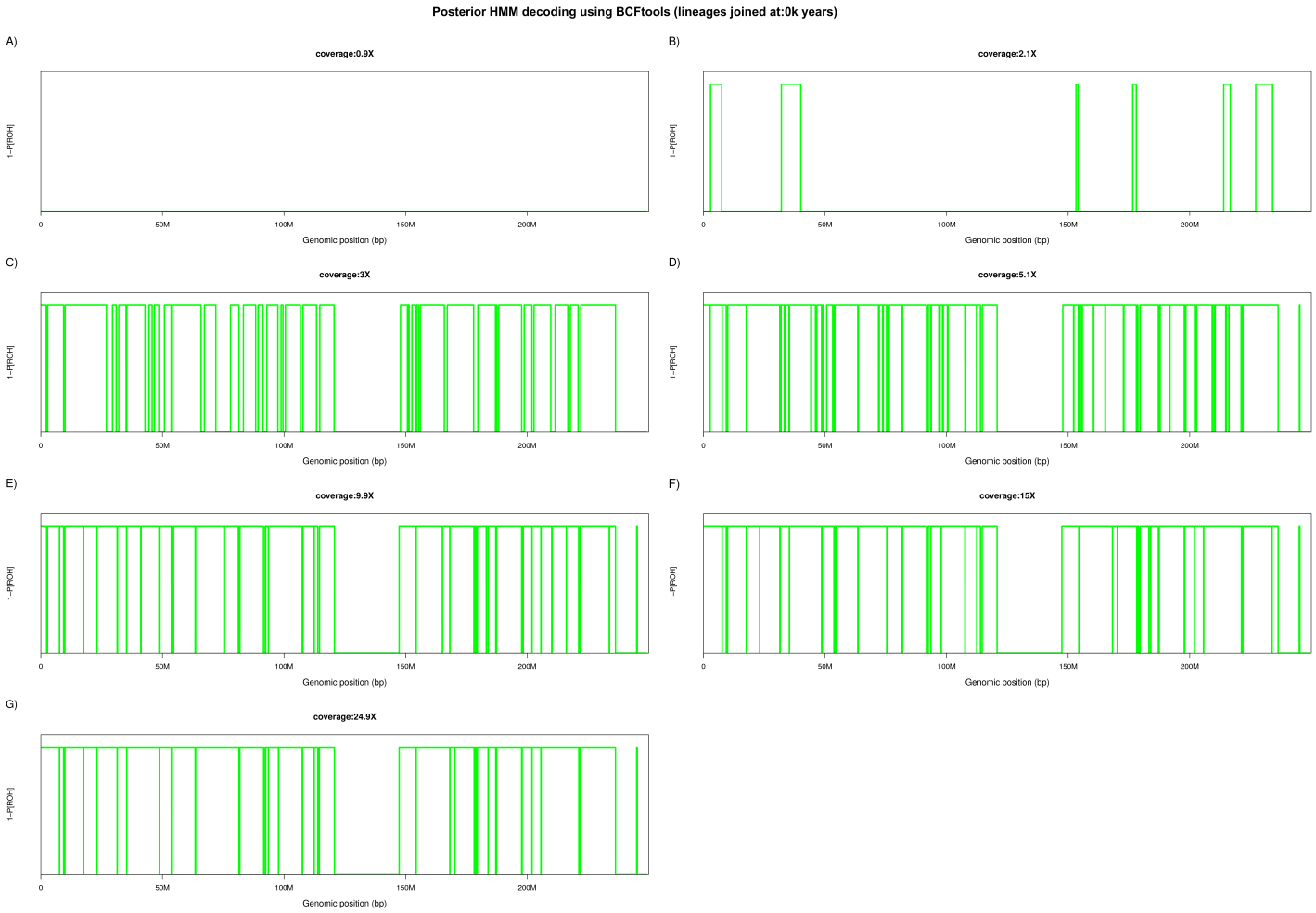

Figure 43: Posterior decoding using BCFtools/RoH at different levels of simulated coverage namely: A) 0.9X B) 2.1X C) 3.0X D) 5.1X E) 9.9X F) 15X G) 24.3X. Please refer to Supplementary Figure 33 for the distribution of the segregating sites on the chromosome. The lineages of the 16 chromosomes to form the grand-parents' chromosomes and the 1000 chromosomes which provide the allele frequency were jointed at a time of 0 years.

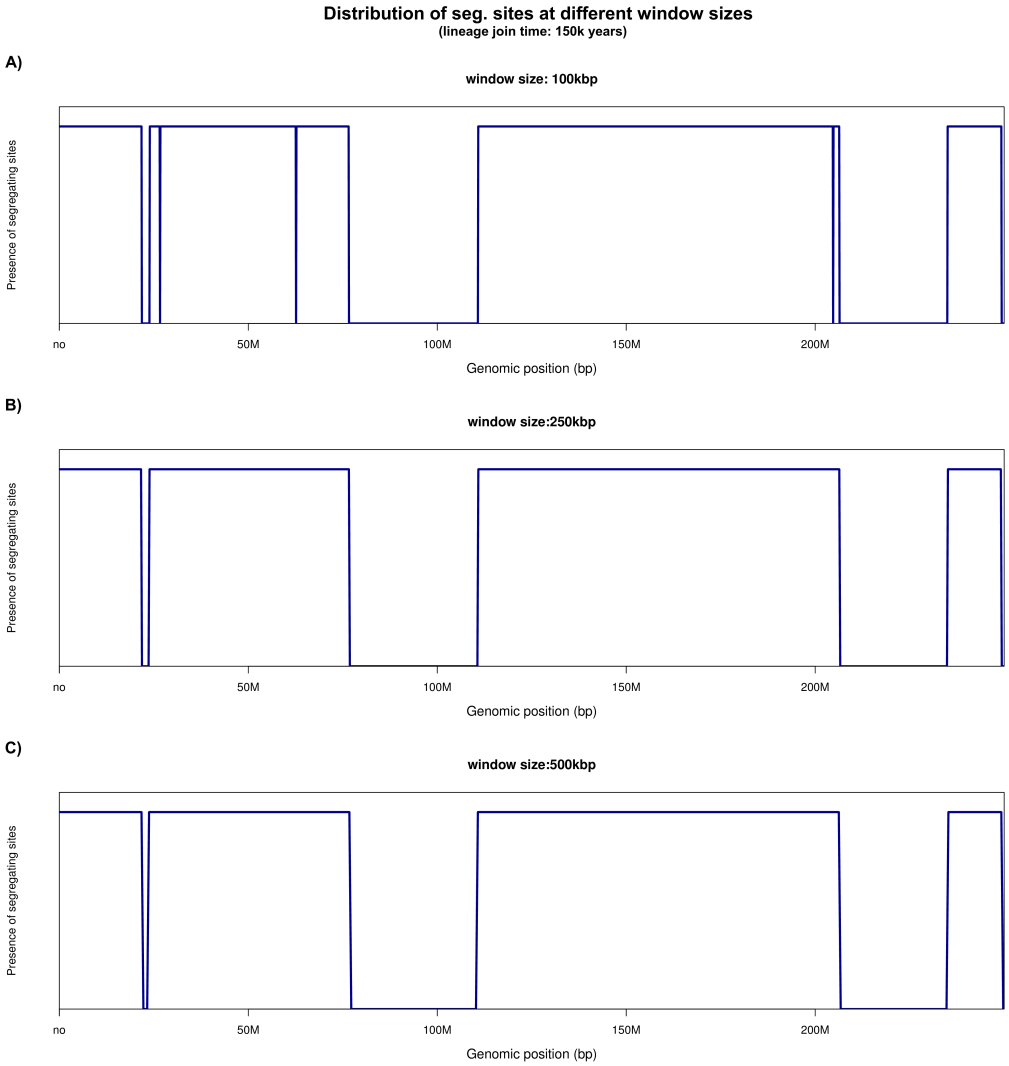

Figure 44: Presence or absence of segregating sites on the simulated chromosomes using windows of A) 1kbp B) 2.5kbp C) 5kbp using the inbreeding scenario 1 (inbreeding between siblings). For the evaluation of BCFtools/RoH, the lineages of the 16 chromosomes to form the grand-parents' chromosomes and the 1000 chromosomes which provide the allele frequency were jointed at a time of 150k years.

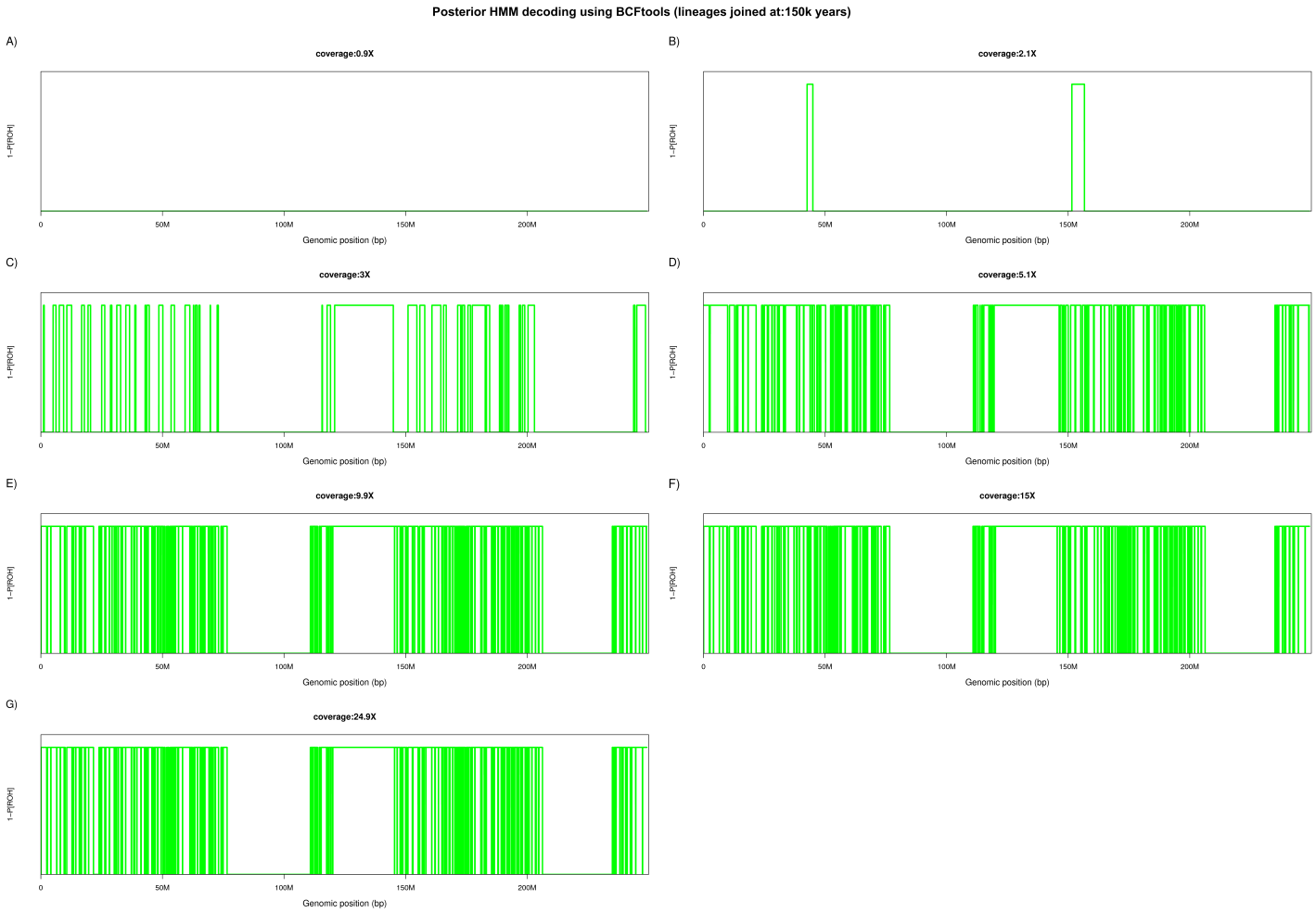

Figure 45: Posterior decoding using BCFtools/RoH at different levels of simulated coverage namely: A) 0.9X B) 2.1X C) 3.0X D) 5.1X E) 9.9X F) 15X G) 24.3X. Please refer to Supplementary Figure 44 for the distribution of the segregating sites on the chromosome. The lineages of the 16 chromosomes to form the grand-parents' chromosomes and the 1000 chromosomes which provide the allele frequency were jointed at a time of 150k years.

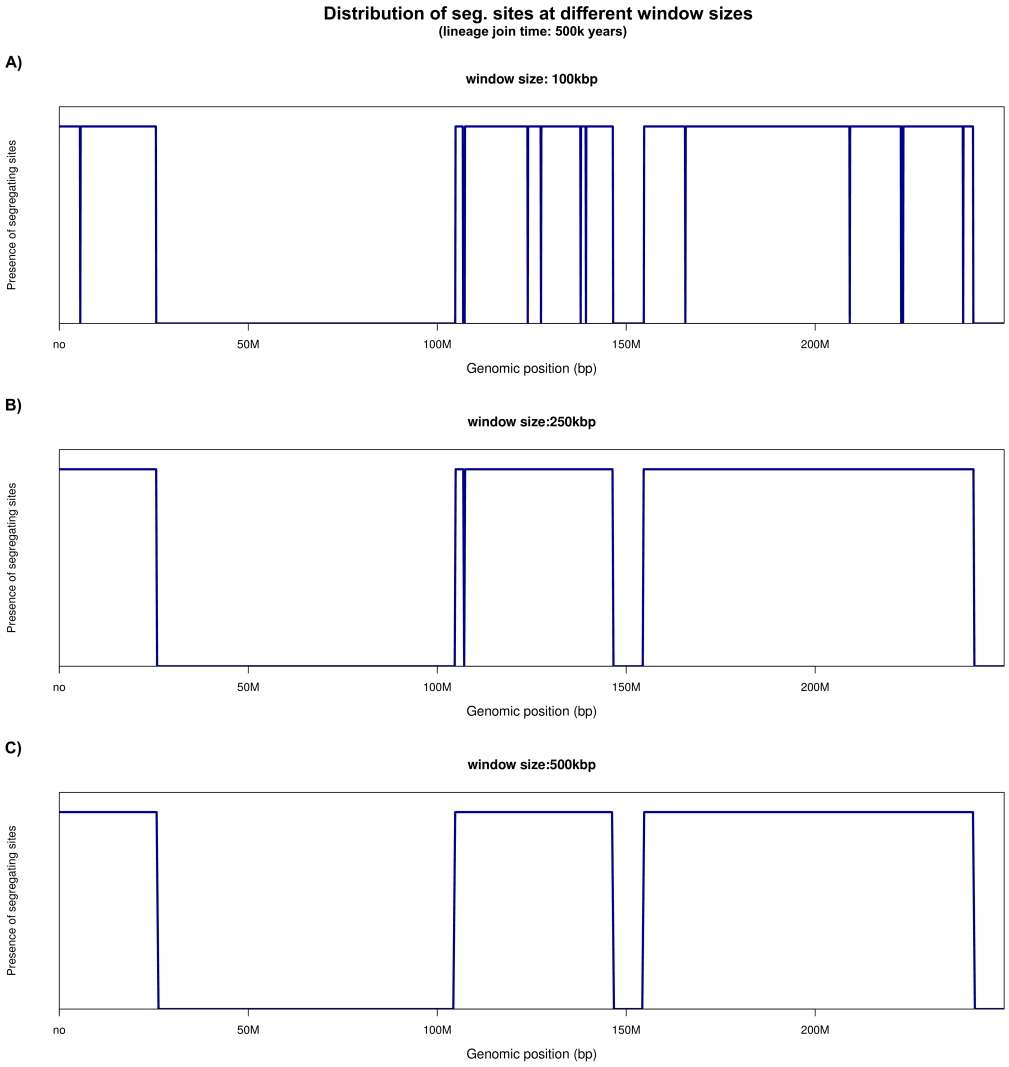

Figure 46: Presence or absence of segregating sites on the simulated chromosomes using windows of A) 1kbp B) 2.5kbp C) 5kbp using the inbreeding scenario 1 (inbreeding between siblings). For the evaluation of BCFtools/RoH, the lineages of the 16 chromosomes to form the grand-parents' chromosomes and the 1000 chromosomes which provide the allele frequency were jointed at a time of 500k years.

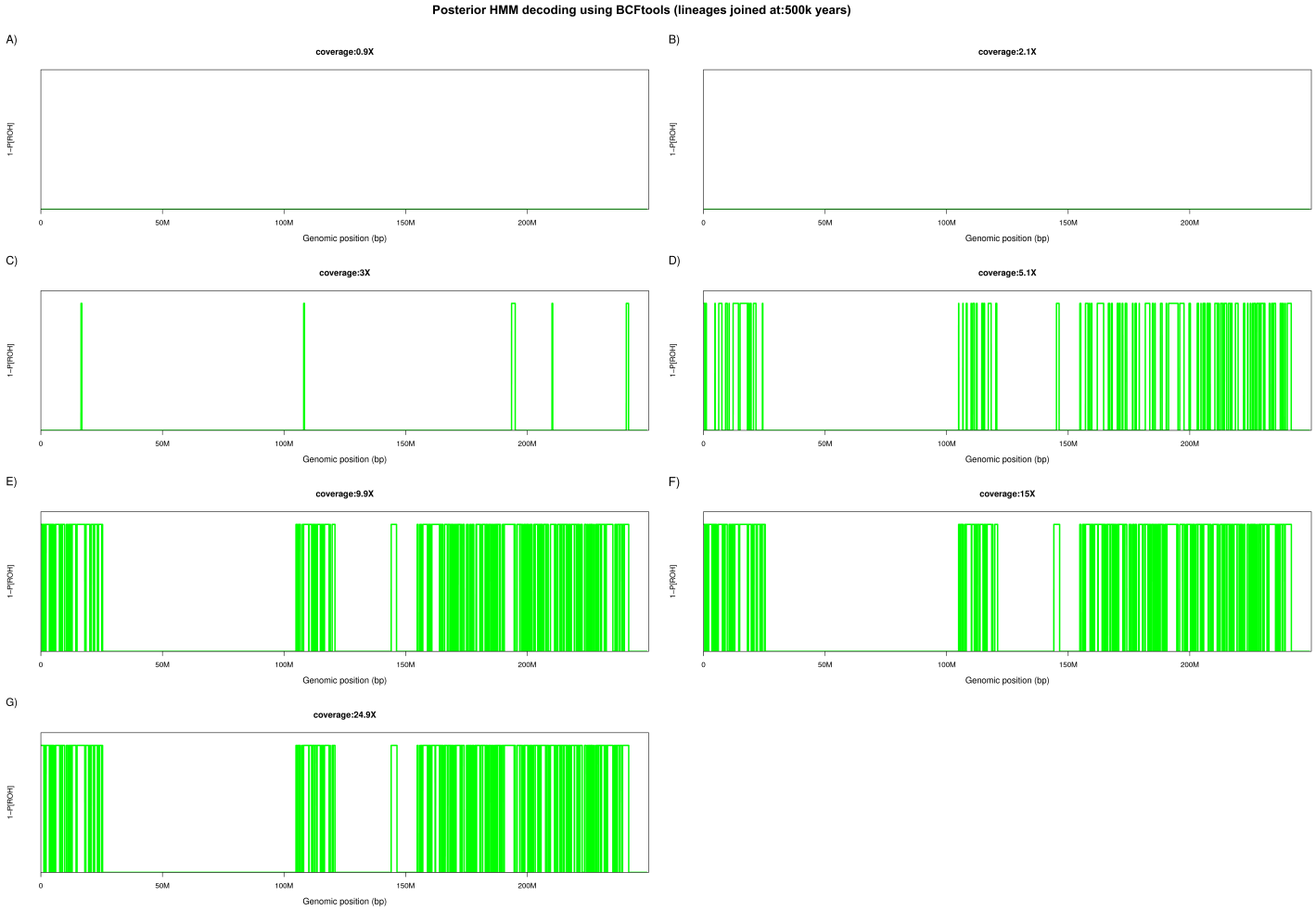

Figure 47: Posterior decoding using BCFtools/RoH at different levels of simulated coverage namely: A) 0.9X B) 2.1X C) 3.0X D) 5.1X E) 9.9X F) 15X G) 24.3X. Please refer to Supplementary Figure 46 for the distribution of the segregating sites on the chromosome. The lineages of the 16 chromosomes to form the grand-parents' chromosomes and the 1000 chromosomes which provide the allele frequency were jointed at a time of 500k years.

1.2 Empirical data

1.2.1 Humans

ID	population	coverage	θ	θ_{low}	θ_{high}	ROH (%)	SGDP θ
HG00707	Southern Han Chinese (CHS)	4.5	0.000713603	0.000545385	0.000924186	0	0.000792535
HG00708		4.9	0.000759869	0.000607272	0.000915707	0	
HG02364	Chinese Dai in Xishuangbanna, China (CDX)	7.1	0.000810391	0.00069312	0.000932023	0	0.000793714-0.00082176
HG02367		7.2	0.000805949	0.000687336	0.000930138	0.138074	
HG02085	Kinh in Ho Chi Minh City, Vietnam (KHV)	8.2	0.000840848	0.000732537	0.00094835	0	0.000806996-0.000828772
HG02086		7.9	0.000859125	0.000751194	0.000977007	0	
NA19068	Japanese in Tokyo, Japan (JPT)	9.2	0.000867054	0.000750137	0.000990337	0	0.000801412-0.000826611
NA19070		6.3	0.000961556	0.000819523	0.00110828	0	
HG01974	Peruvians from Lima, Peru (PEL)	11.0	0.000883585	0.000779036	0.000984885	0	0.000706983-0.000756678
HG01976		11.0	0.000970756	0.000862864	0.00107667	0	
HG03708	Punjabi from Lahore, Pakistan (PJL)	7.1	0.000896297	0.000768666	0.00102456	0	0.000863818-0.000895647
HG03709		7.6	0.00091677	0.000805633	0.00103292	0	
NA21137	Gujarati Indian from Houston, Texas (GIH)	12.7	0.00101854	0.000920601	0.0011136	0	0.000863594[1]
NA21141		7.8	0.000909886	0.000799777	0.00102094	0.0690369	
HG04225	Indian Telugu from the UK (ITU)	6.7	0.000910065	0.000782637	0.00104459	0	0.000826589-0.000887523[2]
HG04222		8.2	0.000924757	0.000818621	0.00103226	0.172592	
HG04171	Bengali from Bangladesh (BEB)	8.1	0.000927666	0.000815712	0.00104155	0	0.00088649-0.000887424
HG04173		5.7	0.00114297	0.000979405	0.00131118	0	
HG04038	Sri Lankan Tamil from the UK (STU)	5.4	0.00118858	0.000993661	0.00137724	0	NA
HG04039		6.0	0.00117458	0.00101575	0.00133913	0	
HG03136	Esan in Nigeria (ESN)	8.3	0.00123733	0.00111082	0.00135873	0	0.00109226-0.00114413
HG03139		7.3	0.00114978	0.00101661	0.00128663	0.0345185	
HG02891	Gambian in Western Divisions in the Gambia (GWD)	6.9	0.00117041	0.00102119	0.001323	0	0.00109896-0.00114846
HG02895		5.9	0.00118227	0.00102562	0.00133952	0	
HG02537	African Caribbeans in Barbados (ACB)	16.7	0.00123678	0.00113072	0.00134187	0	NA
HG02536		5.1	0.00133456	0.00113154	0.00153834	0	

Table 4: Comparison between the genome-wide θ obtained by ROHan on lower coverage samples from the 1000 Genomes project Phase III [Genomes Project Consortium et al., 2015] data and the heterozygosity estimates obtained by the Simons Genome Diversity Project [Mallick et al., 2016]

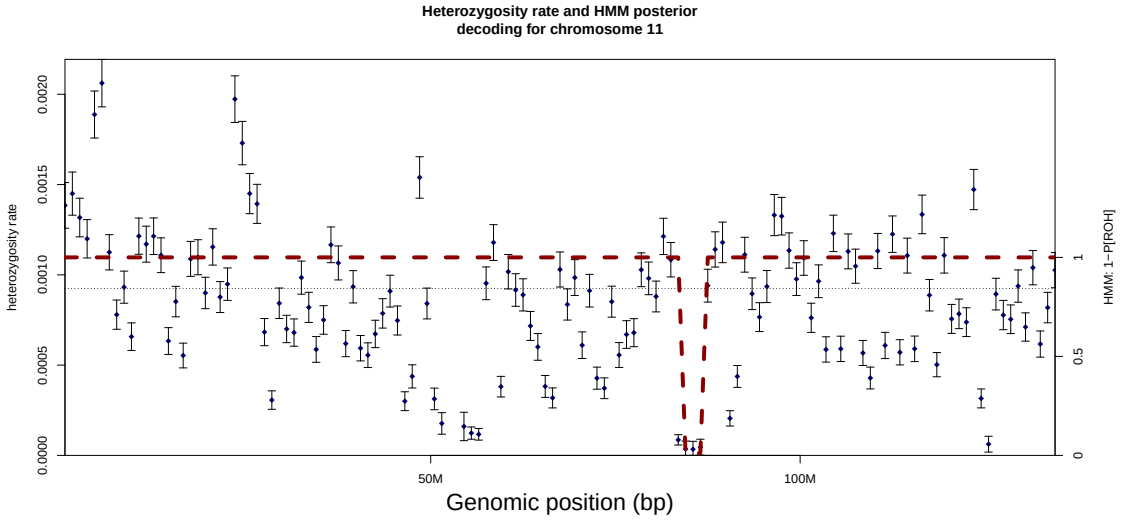

Figure 48: Local estimate of heterozygosity and HMM posterior decoding for chromosomes 11 for the HG04222 individual.

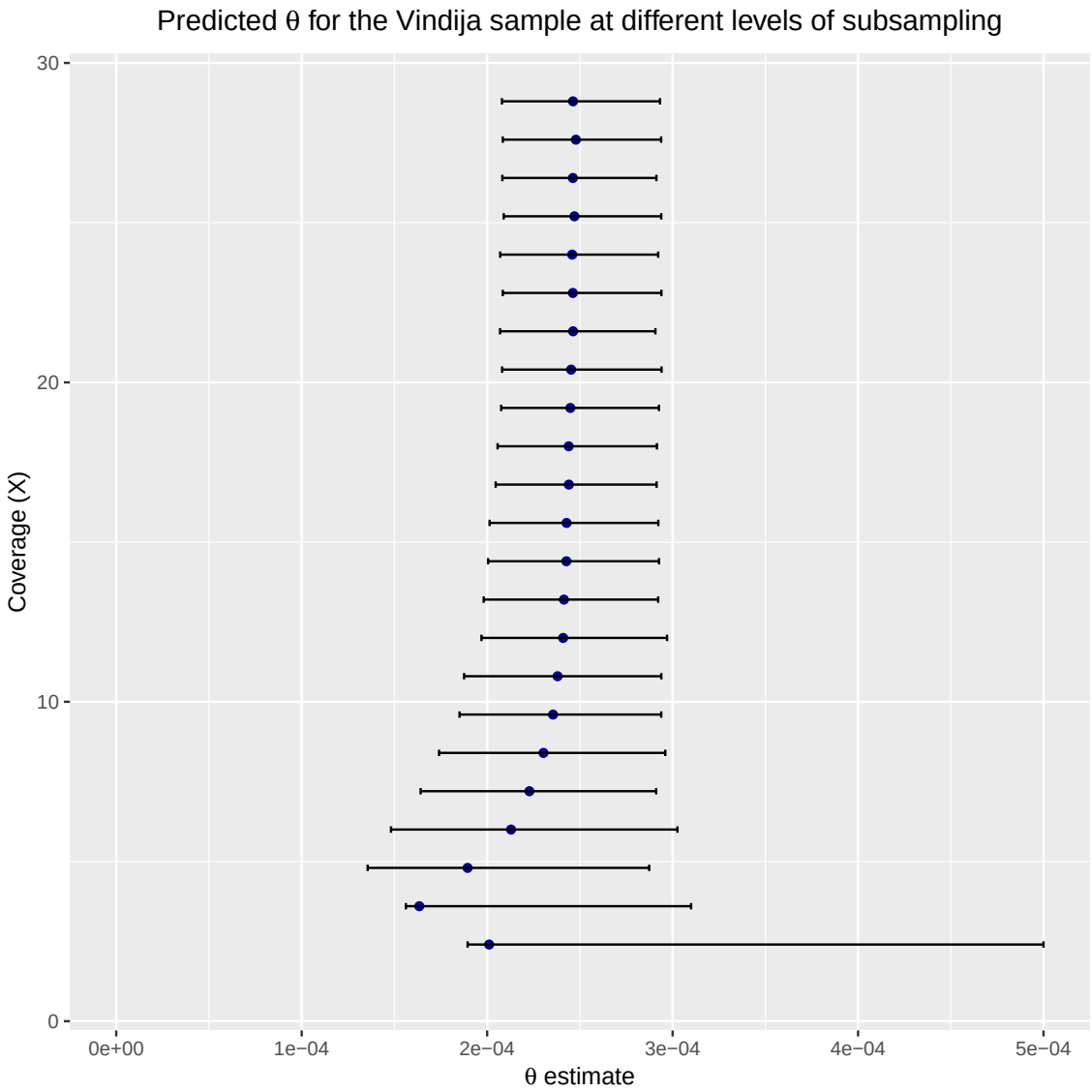

Figure 49: Global estimate of θ for the Vindija 33.19 sample at different rate of subsampling . The deamination rates were evalutated using the script provided with the software where potentially polymorphic positions are masked.

coverage	position from the 5' end											position from the 3' end										
	1	2	3	4	5	6	7	8	9	10	RMSD	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	RMSD
1	0.77	0.73	0.73	0.76	0.76	0.74	0.75	0.75	0.70	0.71	0.03	0.81	0.75	0.74	0.74	0.75	0.73	0.72	0.72	0.71	0.69	0.03
2	0.91	0.90	0.90	0.93	0.93	0.93	0.94	0.92	0.89	0.89	0.01	0.94	0.92	0.92	0.91	0.92	0.90	0.90	0.89	0.91	0.87	0.01
3	0.96	0.96	0.97	0.99	1.01	1.00	0.99	0.98	0.96	0.95	0.01	0.97	0.98	0.98	0.98	0.98	0.99	0.96	0.97	0.98	0.95	0.00
4	0.98	0.99	1.00	1.02	1.03	1.02	1.01	1.00	0.99	0.97	0.00	0.99	1.00	1.01	1.00	1.01	1.04	0.99	1.01	1.01	0.98	0.00
5	0.98	1.00	1.01	1.03	1.04	1.03	1.02	1.02	1.00	1.00	0.00	1.00	1.00	1.01	1.01	1.03	1.04	1.00	1.02	1.03	1.01	0.00
6	0.99	1.01	1.02	1.03	1.05	1.03	1.02	1.02	1.00	1.00	0.00	1.00	1.01	1.01	1.02	1.03	1.04	1.00	1.02	1.03	1.00	0.00
7	0.99	1.01	1.03	1.03	1.05	1.03	1.02	1.01	1.01	0.99	0.00	1.00	1.01	1.01	1.02	1.03	1.04	1.00	1.04	1.04	1.01	0.00
8	0.99	1.01	1.02	1.03	1.04	1.02	1.01	1.01	1.01	1.01	0.00	1.00	1.01	1.01	1.02	1.03	1.03	1.00	1.03	1.03	1.01	0.00
9	0.99	1.01	1.02	1.03	1.05	1.03	1.01	1.01	1.00	1.00	0.00	1.01	1.01	1.01	1.02	1.03	1.02	1.00	1.02	1.03	1.00	0.00
10	1.00	1.01	1.02	1.02	1.04	1.03	1.02	1.01	1.00	1.00	0.00	1.00	1.01	1.01	1.02	1.02	1.02	1.01	1.02	1.03	1.00	0.00
11	0.99	1.01	1.02	1.02	1.04	1.03	1.02	1.01	1.00	1.00	0.00	1.01	1.01	1.01	1.02	1.02	1.02	1.00	1.04	1.02	0.98	0.00
12	0.99	1.01	1.02	1.03	1.04	1.03	1.01	1.02	1.00	1.00	0.00	1.01	1.01	1.01	1.02	1.02	1.02	1.00	1.03	1.02	0.99	0.00
13	0.99	1.00	1.02	1.02	1.04	1.03	1.02	1.01	1.00	1.00	0.00	1.01	1.01	1.01	1.02	1.01	1.01	1.00	1.03	1.02	0.99	0.00
14	0.99	1.00	1.01	1.02	1.04	1.03	1.02	1.01	1.00	1.00	0.00	1.00	1.01	1.01	1.02	1.02	1.01	1.00	1.02	1.02	1.00	0.00
15	0.99	1.00	1.01	1.01	1.03	1.02	1.02	1.01	1.01	0.99	0.00	1.00	1.01	1.01	1.02	1.01	1.01	1.00	1.01	1.01	1.00	0.00
16	1.00	1.00	1.01	1.01	1.03	1.02	1.01	1.01	1.00	0.99	0.00	1.01	1.01	1.01	1.02	1.01	1.01	0.99	1.01	1.01	0.99	0.00
17	1.00	1.00	1.00	1.01	1.02	1.01	1.00	1.01	1.00	0.99	0.00	1.00	1.00	1.00	1.01	1.00	1.01	1.00	1.01	1.01	0.99	0.00
18	1.00	1.00	1.00	1.01	1.02	1.01	1.00	1.00	0.99	1.00	0.00	1.00	1.00	1.00	1.01	0.99	1.01	1.00	1.01	1.01	0.99	0.00
19	1.00	1.00	1.00	1.00	1.01	1.01	0.99	1.01	0.99	1.00	0.00	1.00	1.00	1.00	1.00	0.99	1.01	1.00	1.01	1.01	0.99	0.00
20	1.00	1.00	1.00	1.00	1.01	1.00	0.99	1.00	0.99	1.00	0.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.01	1.00	0.99	0.00
21	1.00	1.00	1.00	1.00	1.01	1.00	0.99	1.00	0.99	1.00	0.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.01	1.00	0.00
22	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.01	1.00	0.99	0.00
23	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.00

Table 5: Error made in the estimate of C to T substitution for the Vindija 33.19 sample at different rate of subsampling. The number reported is the ratio of the deamination rate found to the one found at 24X at the same position. The estimate of substitutions was computed by masking potentially polymorphic and was performed using a script provided with the software package. RMSD stands for root-mean-square deviation.

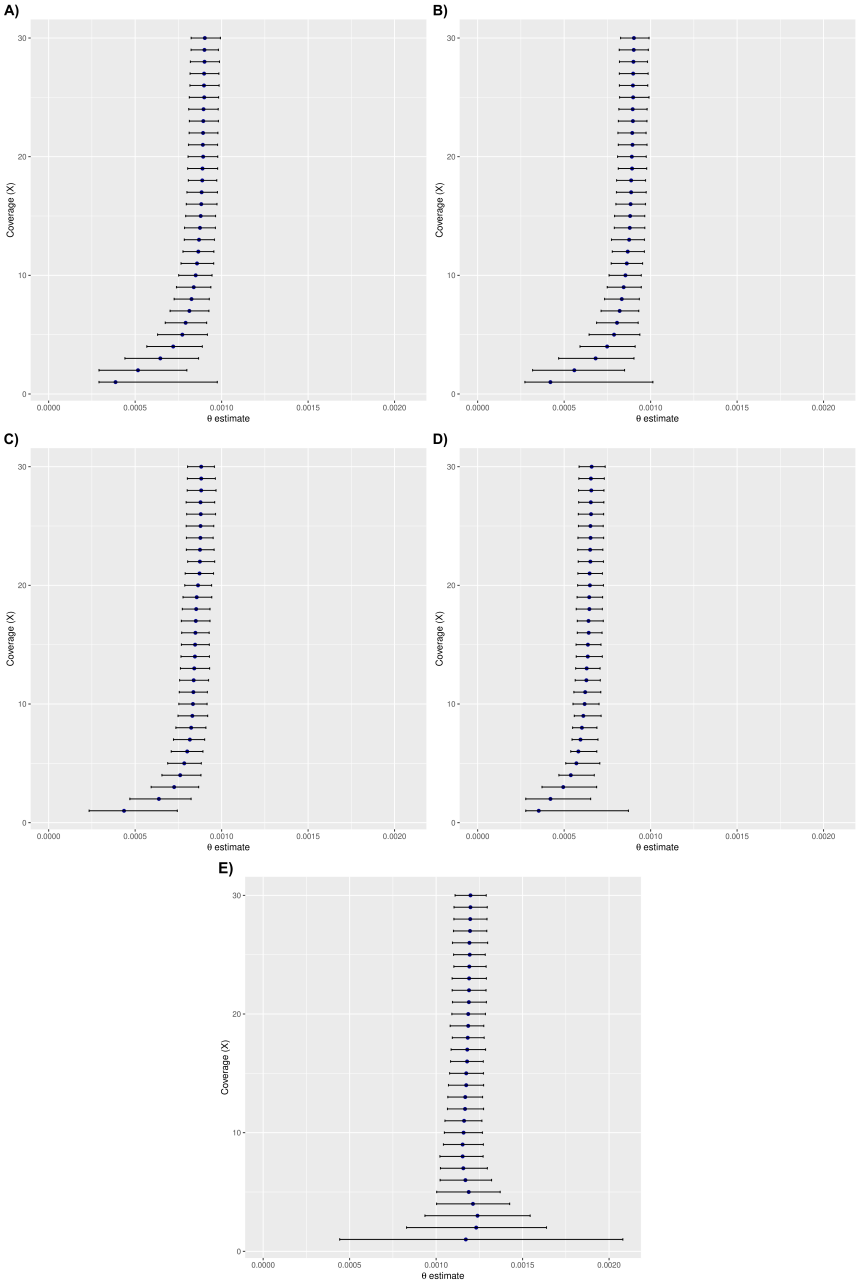

Figure 50: Global estimate of θ at a different rate of subsampling to simulate different depths of coverage for the following modern human individuals from the Simons Genome Diversity Project: A) Bergamo (LP6005441-DNA_B06) B) Czech (LP6005443-DNA_H05)C) Japanese (LP6005441-DNA_G06) D) Karitiana (LP6005441-DNA_G06) and E) Yoruba (LP6005442-DNA_A02). As expected, the rate of heterozygosity is highest in the Yoruba followed by the Czech, Bergamo, Japanese and finally the Karitiana. Our estimates are on par with those reported in the original publication [Mallick et al., 2016]. Our subsampling reveals that our estimates are robust to a depth of 3-4X for these data.

1.2.2 Horses

Used ID	Full sample name	population	age	publication of origin
Arab_0237A	Arab_0237A_SAMN02439777	Arabian	modern	[Metzger et al., 2014]
ARUS_0222A	ARUS_0222A_CGG101397	Yakutian	200 yrs	[Librado et al., 2015]
ARUS_0223A	ARUS_0223A_Batagai	Wild horse from Batagai	5.2k yrs	[Librado et al., 2015]
ARUS_0224A	ARUS_0224A_CGG10022	Wild horse from Taymyr	43k yrs	[Schubert et al., 2014]
ARUS_0225A	ARUS_0225A_CGG10023	Wild horse from Taymyr	16k yrs	[Schubert et al., 2014]
Borly4_PAVH11	Borly4_PAVH11_CGG_018171	Pavlodar site (Kazakhstan)	5k yrs	[Gaunitz et al., 2018]
Borly4_PAVH4	Borly4_PAVH4_CGG_018157	Pavlodar site (Kazakhstan)	5k yrs	[Gaunitz et al., 2018]
Borly4_PAVH8	Borly4_PAVH8_CGG_018165	Pavlodar site (Kazakhstan)	5k yrs	[Gaunitz et al., 2018]
Botai2	Botai2_CGG_1_018174	Botai Culture	5.5k yrs	[Gaunitz et al., 2018]
Botai5	Botai5_CGG_018177	Botai Culture	5.5k yrs	[Gaunitz et al., 2018]
Botai6	Botai6_CGG_018178	Botai Culture	5.5k yrs	[Gaunitz et al., 2018]
Icel_0247A	Icel_0247A_IS074	Icelandic	modern	[Jäderkvist et al., 2014]
Icel_0144A	Icel_0144A_P5782	Icelandic	modern	[Jäderkvist et al., 2014]
Jeju_0275A	Jeju_0275A_SAMN01057172	Jeju Pony	modern	[Kim et al., 2013]
Mong_0215A	Mong_0215A_TG1111D2628	Mongolian	modern	[Do et al., 2014]
Mong_0153A	Mong_0153A_KB7754	Mongolian	modern	[Der Sarkissian et al., 2015]
Prze_0150A	Prze_0150A_KB3879	Przewalski	modern	[Der Sarkissian et al., 2015]
Prze_0151A	Prze_0151A_KB7674	Przewalski	modern	[Der Sarkissian et al., 2015]
Prze_0157A	Prze_0157A_SB293	Przewalski	modern	[Der Sarkissian et al., 2015]
Prze_0158A	Prze_0158A_SB339	Przewalski	modern	[Der Sarkissian et al., 2015]
Prze_0159A	Prze_0159A_SB4329	Przewalski	modern	[Der Sarkissian et al., 2015]
Prze_0160A	Prze_0160A_SB533	Przewalski	modern	[Der Sarkissian et al., 2015]
SCYT_I_Ch118	I_Ch118_CGG_1_016176	Scythian kurgan	2.3k yrs	[Librado et al., 2017]
SCYT_E_Ch25	E_Ch25_CGG_1_016172	Scythian kurgan	2.3k yrs	[Librado et al., 2017]
SCYT_F_Ch26	F_Ch26_CGG_1_016173	Scythian kurgan	2.3k yrs	[Librado et al., 2017]
Shet_0249A	Shet_0249A_SPH020	Shetland Pony	modern	[Frischknecht et al., 2015]
Shet_0250A	Shet_0250A_SPH041	Shetland Pony	modern	[Frischknecht et al., 2015]
Stan_0081A	Stan_0081A_M5256	Standardbred	modern	[Der Sarkissian et al., 2015]
Thor_0290A	Thor_0290A_SAMN01047706	Thoroughbred	modern	[Do et al., 2014]
Thor_0145A	Thor_0145A_Twilight	Thoroughbred	modern	[Wade et al., 2009]
Yaku_0170A	Yaku_0170A_Yak8	Yakutian	modern	[Librado et al., 2015]
Yaku_0171A	Yaku_0171A_Yak9	Yakutian	modern	[Librado et al., 2015]
Yaku_0163A	Yaku_0163A_Yak1	Yakutian	modern	[Librado et al., 2015]

Table 6: Population of origin, coverage and inferred fraction of the genome to be an ROH for the different horse presented in the main manuscript.

Sample name	global θ estimate					
	accounting for deamination in modern			standard estimate		
	mid	low	high	mid	low	high
Prze_0150A_KB3879	0.00116396	0.00105745	0.00126795	0.00121625	0.00111703	0.00133975
Prze_0158A_SB339	0.00130449	0.00119473	0.0014413	0.00136348	0.00124967	0.00148758
Prze_0159A_SB4329	0.00137299	0.00123522	0.0015161	0.00151808	0.00137978	0.00165049
Prze_0160A_SB533	0.00108715	0.00097201	0.00120968	0.00123092	0.00111888	0.00134607
Icel_0144A_P5782	0.00132952	0.00110264	0.00158799	0.00167654	0.00154191	0.00181988
Thor_0145A_Twilight	0.00107117	0.000954924	0.0012049	0.00109072	0.000995605	0.00119876
Yaku_0163A_Yak1	0.00165856	0.00147891	0.00183537	0.0018521	0.00167155	0.00199765

Table 7: Effect of accounting for ancient DNA damage in modern samples. The θ was computed by disallowing ROHs to provide a global average.

References

[Der Sarkissian et al., 2015] Der Sarkissian, C., Ermini, L., Schubert, M., Yang, M. A., Librado, P., Fumagalli, M., Jónsson, H., Bar-Gal, G. K., Albrechtsen, A., Vieira, F. G., et al. (2015). Evolutionary genomics and conservation of the endangered Przewalski's horse. *Current Biology*, 25(19):2577–2583.

[Do et al., 2014] Do, K.-T., Kong, H.-S., Lee, J.-H., Lee, H.-K., Cho, B.-W., Kim, H.-S., Ahn, K., and Park, K.-D. (2014). Genomic characterization of the przewalski's horse inhabiting mongolian steppe by whole genome re-sequencing. *Livestock Science*, 167:86–91.

[Frischknecht et al., 2015] Frischknecht, M., Jagannathan, V., Plattet, P., Neuditschko, M., Signer-Hasler, H., Bachmann, I., Pacholewska, A., Drögemüller, C., Dietschi, E., Flury, C., et al. (2015). A non-synonymous HMGA2 variant decreases height in Shetland ponies and other small horses. *PLoS ONE*, 10(10):e0140749.

[Gaunitz et al., 2018] Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., Seguin-Orlando, A., Owens, I. J., Felkel, S., Bignon-Lau, O., et al. (2018). Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science*, 360(6384):111–114.

[Genomes Project Consortium et al., 2015] Genomes Project Consortium, . et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68.

[Jäderkvist et al., 2014] Jäderkvist, K., Andersson, L., Johansson, A., Árnason, T., Mikko, S., Eriksson, S., Andersson, L., and Lindgren, G. (2014). The DMRT3 'Gait keeper' mutation affects performance of Nordic and Standardbred trotters. *Journal of Animal Science*, 92(10):4279–4286.

[Kim et al., 2013] Kim, H., Lee, T., Park, W., Lee, J. W., Kim, J., Lee, B.-Y., Ahn, H., Moon, S., Cho, S., Do, K.-T., Kim, H.-S., Lee, H.-K., Lee, C.-K., Kong, H.-S., Yang, Y.-M., Park, J., Kim, H.-M., Kim, B. C., Hwang, S., Bhak, J., Burt, D., Park, K.-D., Cho, B.-W., and Kim, H. (2013). Peeling back the evolutionary layers of molecular mechanisms responsive to exercise-stress in the skeletal muscle of the racing horse. *DNA Research*, 20(3):287–298.

[Librado et al., 2015] Librado, P., Der Sarkissian, C., Ermini, L., Schubert, M., Jónsson, H., Albrechtsen, A., Fumagalli, M., Yang, M. A., Gamba, C., Seguin-Orlando, A., et al. (2015). Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proceedings of the National Academy of Sciences*, 112(50):E6889–E6897.

[Librado et al., 2017] Librado, P., Gamba, C., Gaunitz, C., Der Sarkissian, C., Pruvost, M., Albrechtsen, A., Fages, A., Khan, N., Schubert, M., Jagannathan, V., et al. (2017). Ancient genomic changes associated with domestication of the horse. *Science*, 356(6336):442–445.

[Mallick et al., 2016] Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

[Metzger et al., 2014] Metzger, J., Tonda, R., Beltran, S., Águeda, L., Gut, M., and Distl, O. (2014). Next generation sequencing gives an insight into the characteristics of highly selected breeds versus non-breed horses in the course of domestication. *BMC Genomics*, 15(1):562.

[Schubert et al., 2014] Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., Albrechtsen, A., Dupanloup, I., Foucal, A., Petersen, B., et al. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences*, 111(52):E5661–E5669.

[Wade et al., 2009] Wade, C., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T., Adelson, D., Bailey, E., Bellone, R., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326(5954):865–867.