

Supplementary Figures

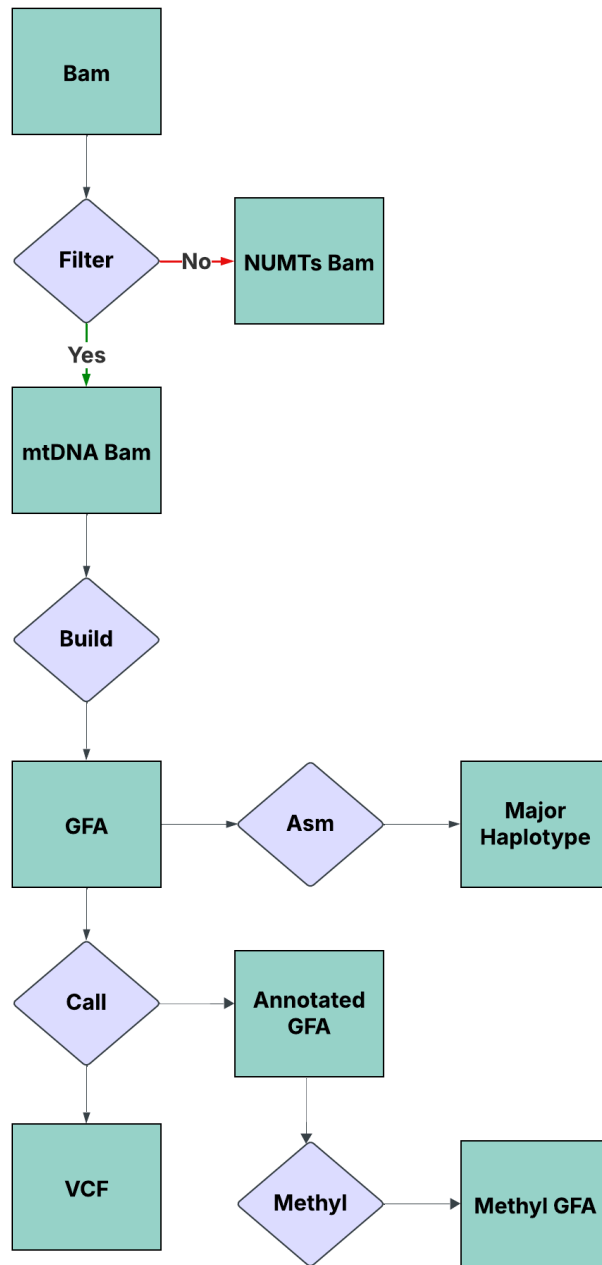
Table of contents for Supplementary Figures

Supplementary Figures.....	1
Supplementary Figure 1 Workflow for Himito.....	3
Supplementary Figure 2 In silico mixture experiment for heteroplasmic calling evaluation.....	4
Supplementary Figure 3 Distinct methylation patterns in NUMTs and mtDNA sequences.....	5
Supplementary Figure 4 Homoplasmic calling evaluation.....	6
Supplementary Figure 5. Assembly benchmarking using HPRC PacBio and ONT (R9, R10) long read samples.....	7
Supplementary Figure 6 Variant calling benchmark against Mitosaw at default parameters (Mitosaw minimal-MAF = 0.1, Himito vaf-threshold = 0.01).....	8
Supplementary Figure 7 Mitosaw and Himito comparison with the same detection limits.....	9
Supplementary Figure 8 Himito variant calling performance using ONT-R9 and ONT-R10 data.....	10
Supplementary Figure 9 HapMap mixture benchmarking using PacBio and ONT long read sequencing technology.....	11
Supplementary Figure 10 The impact of anchor size selection and reference genome choice on graph topology, assembly accuracy, and variant calling performance.....	12
Supplementary Figure 11. Benchmarking of methylation filtering threshold with the variant calling performance.....	13
Supplementary Figure 12 Runtime and peak memory usage.....	14
Supplementary File 13 The impact of sample coverage to runtime and peak memory usage.....	15
Supplementary Figure 14, Benchmark of Himito Methyl module against PbCpG-count Mode.....	16
Supplementary Figure 15 Hierarchical clustering of the graph vertices presence-absence matrix in each assembly.....	18
Supplementary Figure 16, PCA plot for CpG modification rate in 41 paired PacBio and ONT samples in AoU v8 cohort.....	19
Supplementary Figure 17, The methylation modification score distribution in 1490 All of us v8 PacBio samples and 324 ONT samples.....	19
Supplementary Figure 18, Recurrent hypermethylated CpG sites across PacBio (modification rate threshold = 0.417) and ONT samples (modification rate threshold = 0.125).....	20
Supplementary Figure 20. The number of different variant types in Age groups (sliding window of 10).....	20
Supplementary Figure 21. Number of length heteroplasmic variants per individual and the joint alleles distribution.....	21
Supplementary Figure 22. Fisher's Exact test for the association for length heteroplasmic alleles across 229985 reads.....	22
Supplementary Figure 23 Co-occurrence of length heteroplasmies.....	23
Supplementary Figure 24 Haplotype Number of chrM:302-316 in Age groups (sliding window of 10).....	24
Supplementary Figure 25 Kmer composition of Himito, MitoHiFi and HiMT assemblies for Lycopodium japonicum mitochondrial genome (SRA accession: SRX20558281).....	26
Supplementary Figure 26 Graph topology and sequence variations in chrM: 302-315 in HG005.....	27
Supplementary Figure 27 Variant length distribution in All of Us v8 PacBio samples.....	28

Supplementary Tables

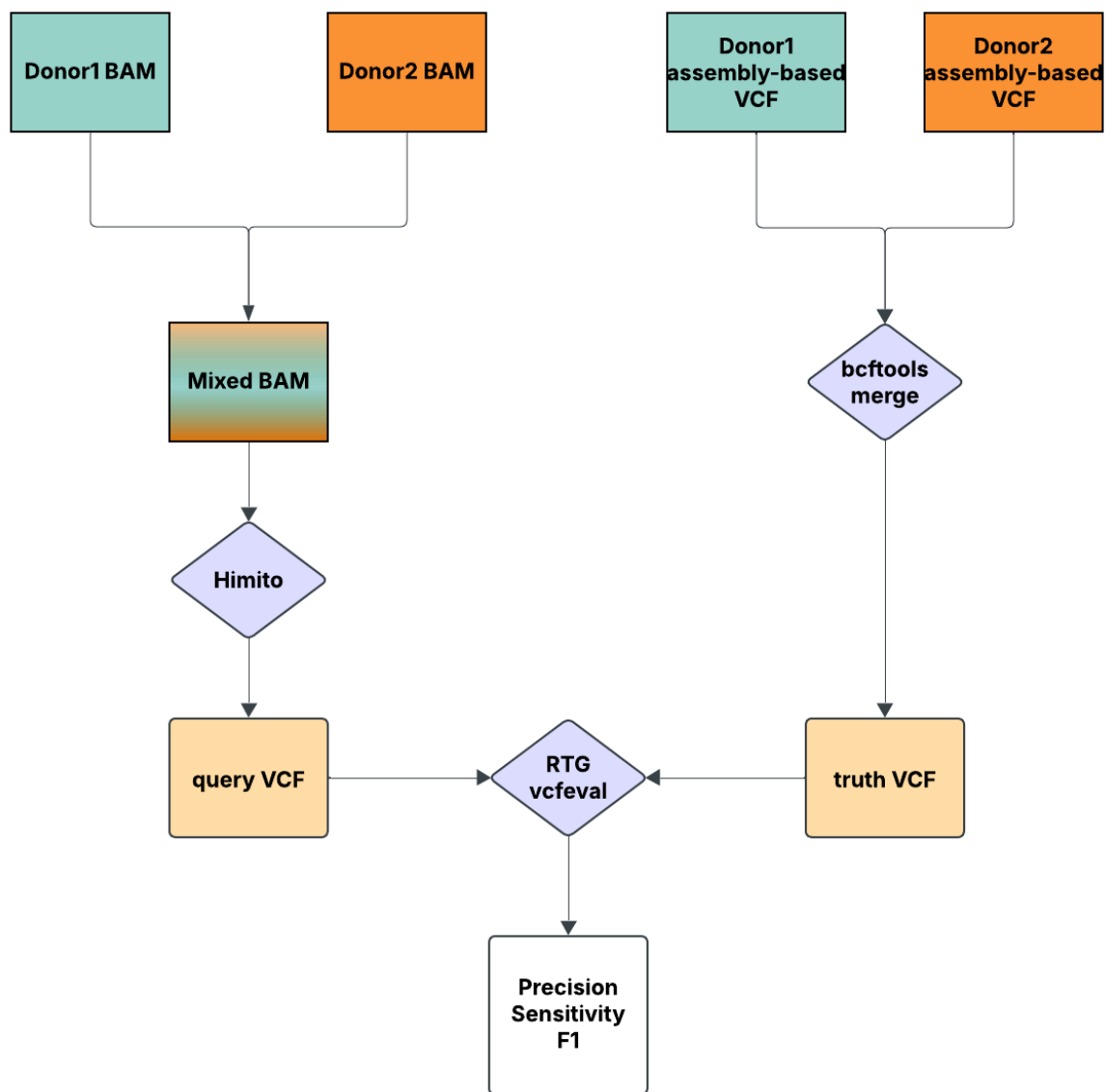
Supplementary Table of Content:

Supplementary Table 1	In silico mixture experiment design (PacBio samples).....	29
Supplementary Table 2	In silico mixture experiment design (ONT-R10).....	30
Supplementary Table 3	In silico mixture experiment design.....	31



Supplementary Figure 1 Workflow for Himito.

Himito comprises 5 key modules for comprehensive mitochondrial genome analysis. Filter takes IrWGS as input to classify reads as mtDNA-derived and NUMTs-derived reads. It aims to remove NUMTs-derived reads from the downstream analysis. Build constructs a sequence graph to represent the diversity of mtDNAs. Asm takes the sequence graph as the input and extracts primary haplotypes from the graph. Call performs pair-wise alignment between parallel paths and aims to call both homoplasmic and heteroplasmic variants from the graph. Methyl annotates the methylation signals to the graph and aggregates the methylation signals along the major haplotype.

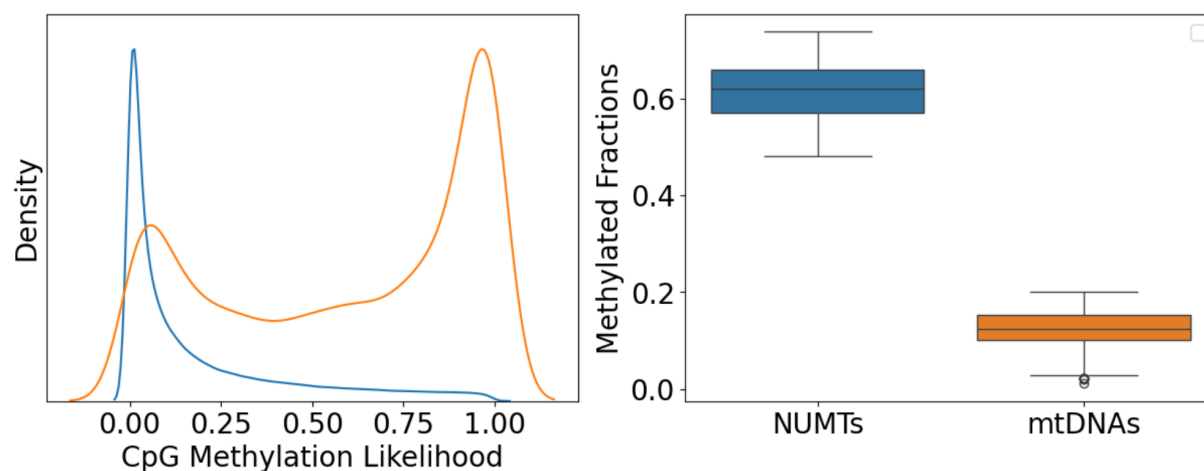


Supplementary Figure 2 In silico mixture experiment for heteroplasmic calling evaluation.

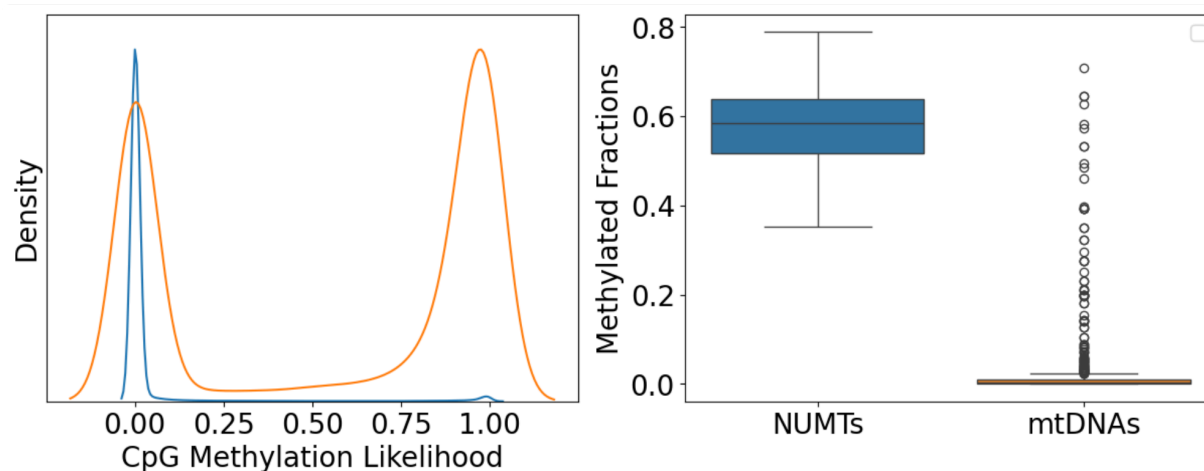
In silico mixture experiments were designed and performed to evaluate the heteroplasmic variant calling performance at predefined frequencies. The performance of the variant caller on the mixed sample was assessed by comparing the called variants against the truth set using the RTG vcfeval tool, allowing for the quantification of precision and recall at the simulated heteroplasmic frequency.

Supplementary Table 1

PacBio Sample

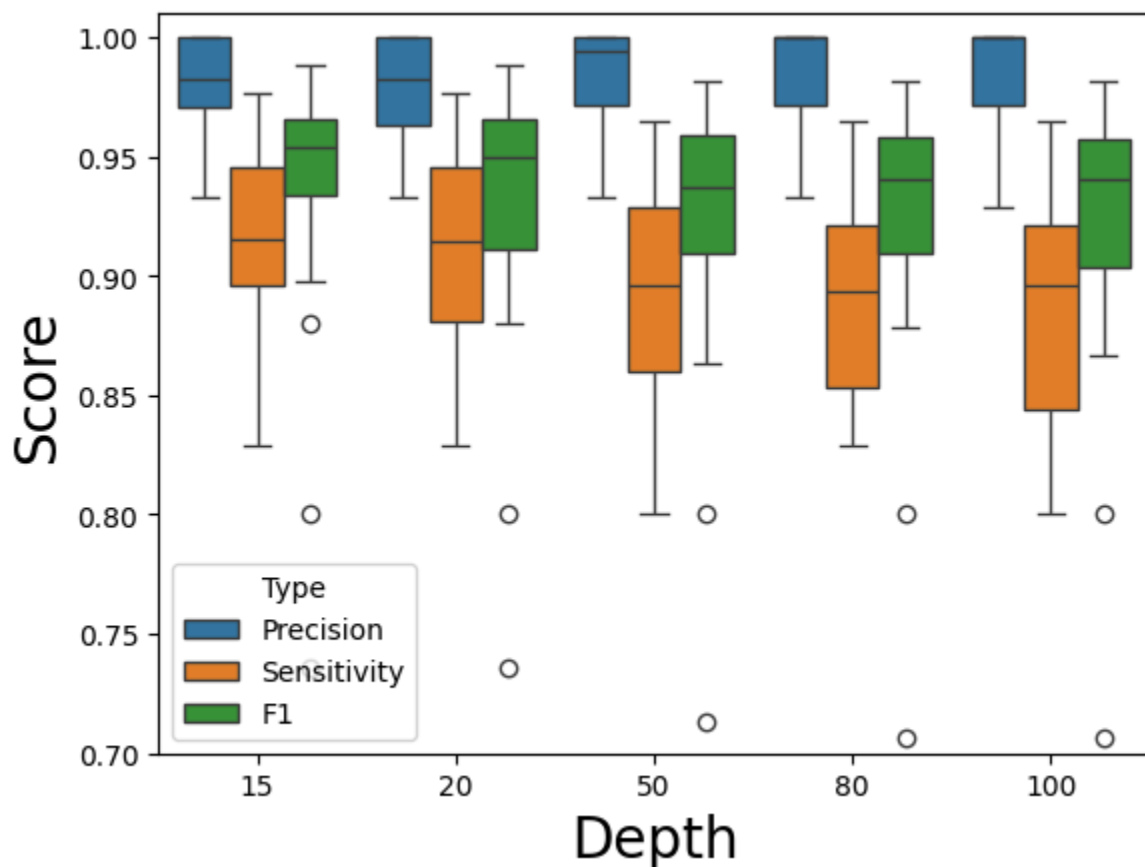


ONT Sample



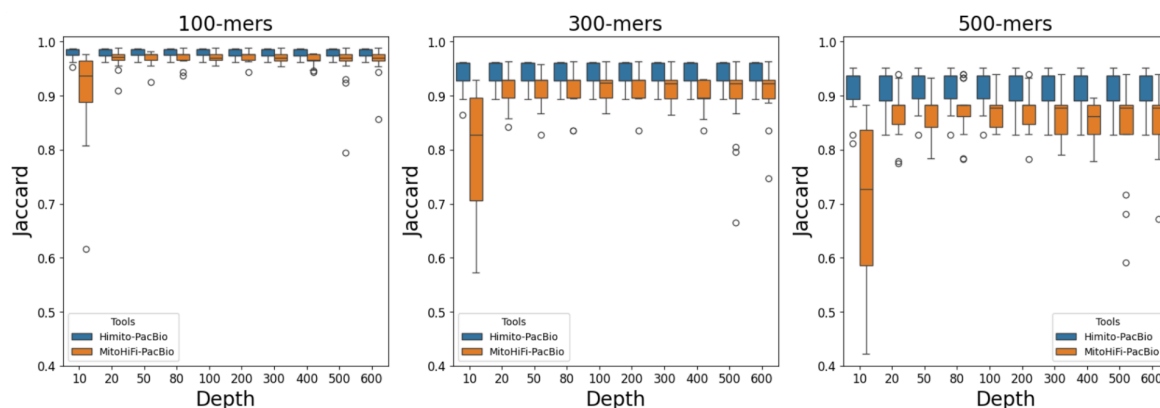
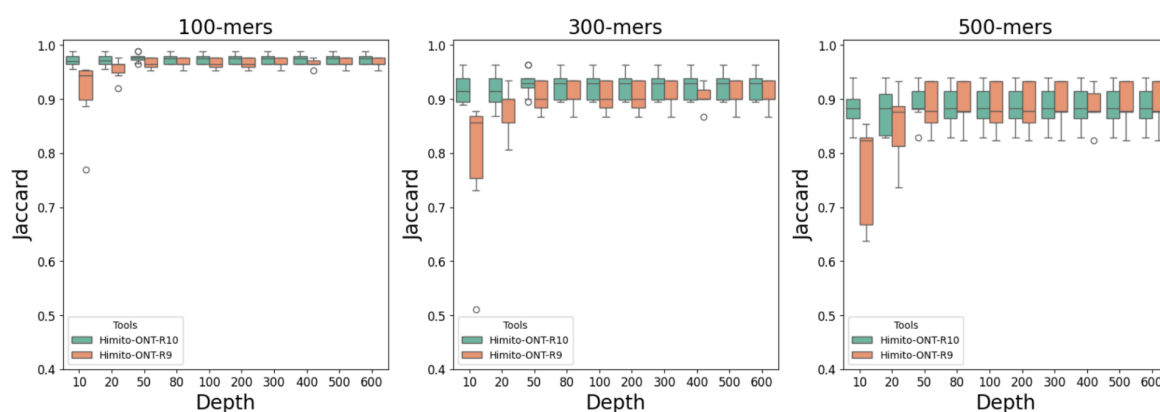
Supplementary Figure 3 Distinct methylation patterns in NUMTs and mtDNA sequences.

We examine the methylation modification likelihood of reads extracted from one of the largest NUMTs sequences in GRCh38 (*chr2:130271785-130283331*) and compare them to reads aligned to chrM in both PacBio and ONT samples ($n = 10$). We plot the single CpG site methylation likelihood for those 2 regions. Secondly, we plot the fraction of methylated CpG per-read in these same 2 regions, where a CpG was considered methylated if the likelihood exceeded 0.5. Our findings indicate distinct methylation patterns between the nuclear DNA and mtDNA, consistent with prior research. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within $1.5 \times$ interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers.



Supplementary Figure 4 Homoplasmic calling evaluation.

The homoplasmic variants ($HF > 95\%$) in the Himito callset are compared with the HPRC assembly-based truth callset (generated by minigraph-cactus). The HPRC truth callset comprises both homoplasmic variants and major heteroplasmic variants. Using a heteroplasmy frequency (HF) cutoff of 95% might miss some major heteroplasmic variants in the assembly-based truth callset, potentially leading to reduced sensitivity. $N = 26$. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within $1.5 \times$ interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers.

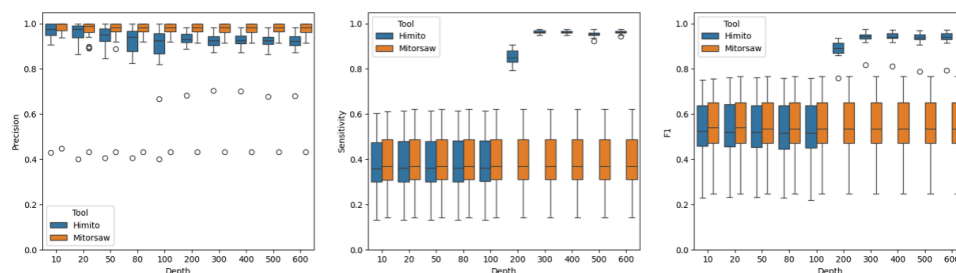
a**PacBio-assembly Benchmarking****b****ONT-assembly Benchmarking**

Supplementary Figure 5. Assembly benchmarking using HPRC PacBio and ONT (R9, R10) long read samples.

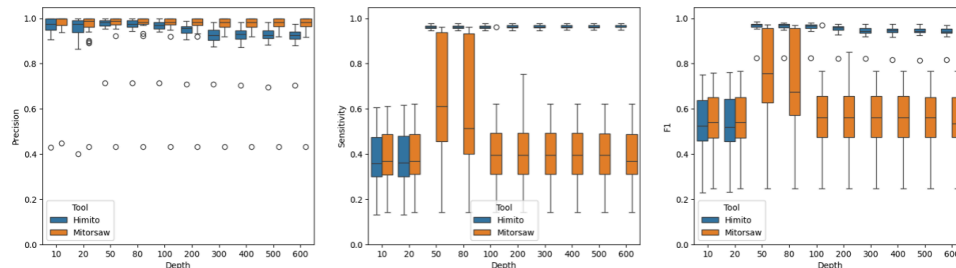
a. To evaluate the accuracy of the Himito-PacBio assembly, the assembly quality was benchmarked against MitoHiFi across different sample coverages ($n = 29$). This was achieved by comparing the Jaccard similarity of k -mers (100-mers, 300-mers, and 500-mers) derived from both the assembled contigs and the ground truth sequence. The resulting Jaccard similarity served as the metric for assessing assembly accuracy. b. Similar approaches are applied to estimate the Himito-ONT (R9 and R10) assemblies ($n = 10$). All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within $1.5 \times$ interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers.

Mitorsaw minimal-MAF = 0.1 (default)

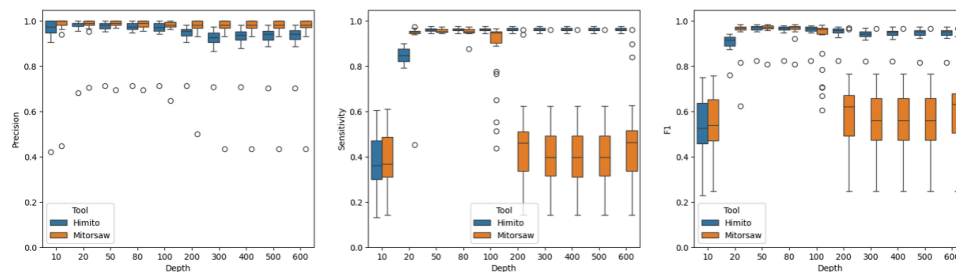
a. 1%



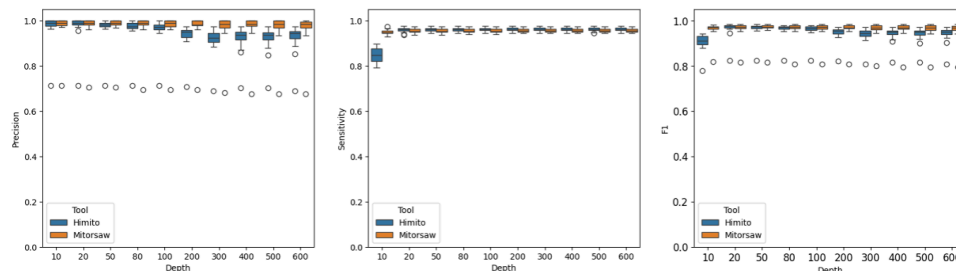
b. 5%



c. 10%



d. 20%

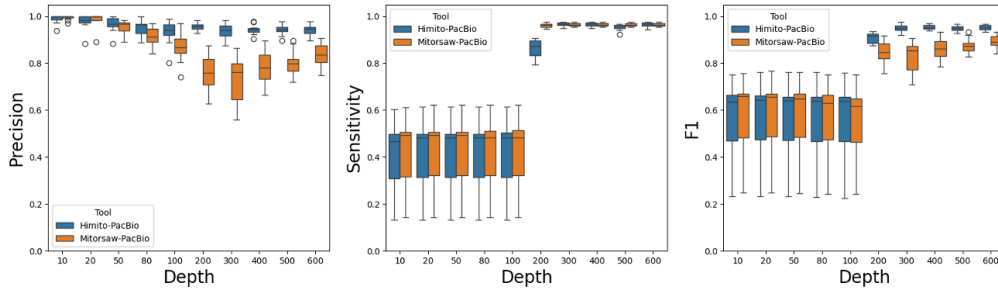


Supplementary Figure 6 Variant calling benchmark against Mitorsaw at default parameters (Mitorsaw minimal-MAF = 0.1, Himito vaf-threshold = 0.01).

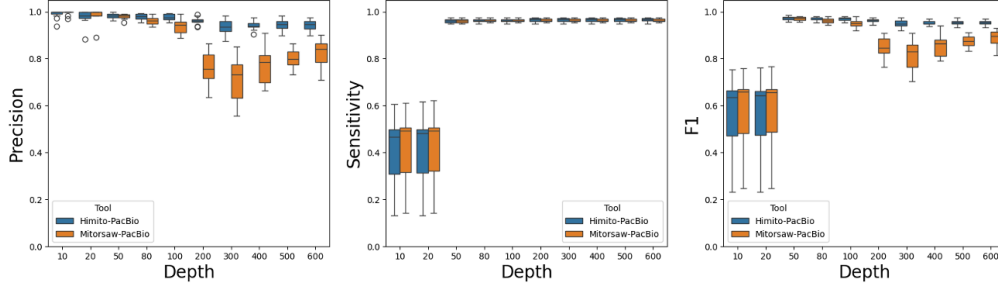
In silico experiments evaluated heteroplasmic frequencies at 0.01 (a), 0.05 (c), 0.1 (d), and 0.2 (e). The precision, sensitivity, and F1 score were calculated for each frequency. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within 1.5 * interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers. Downsampling seed -R = 1. Sample Size N = 25

minimal-MAF = 0.01

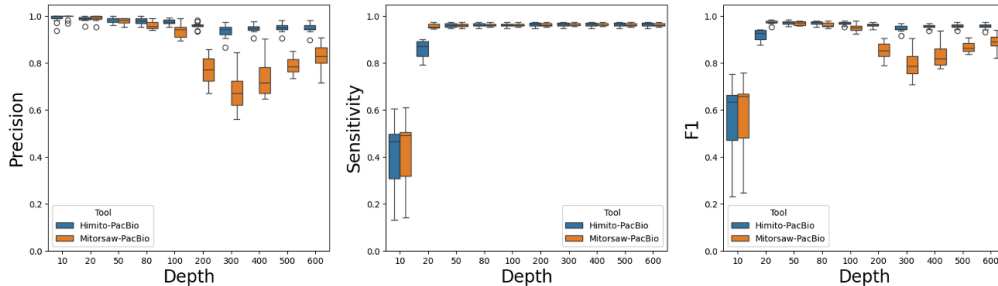
a. 1%



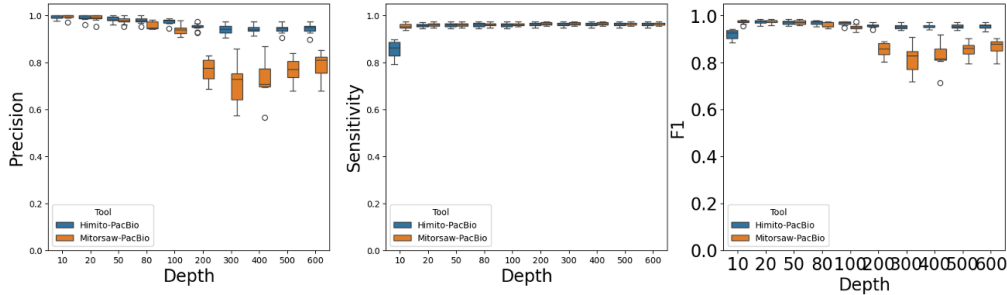
b. 5%



c. 10%



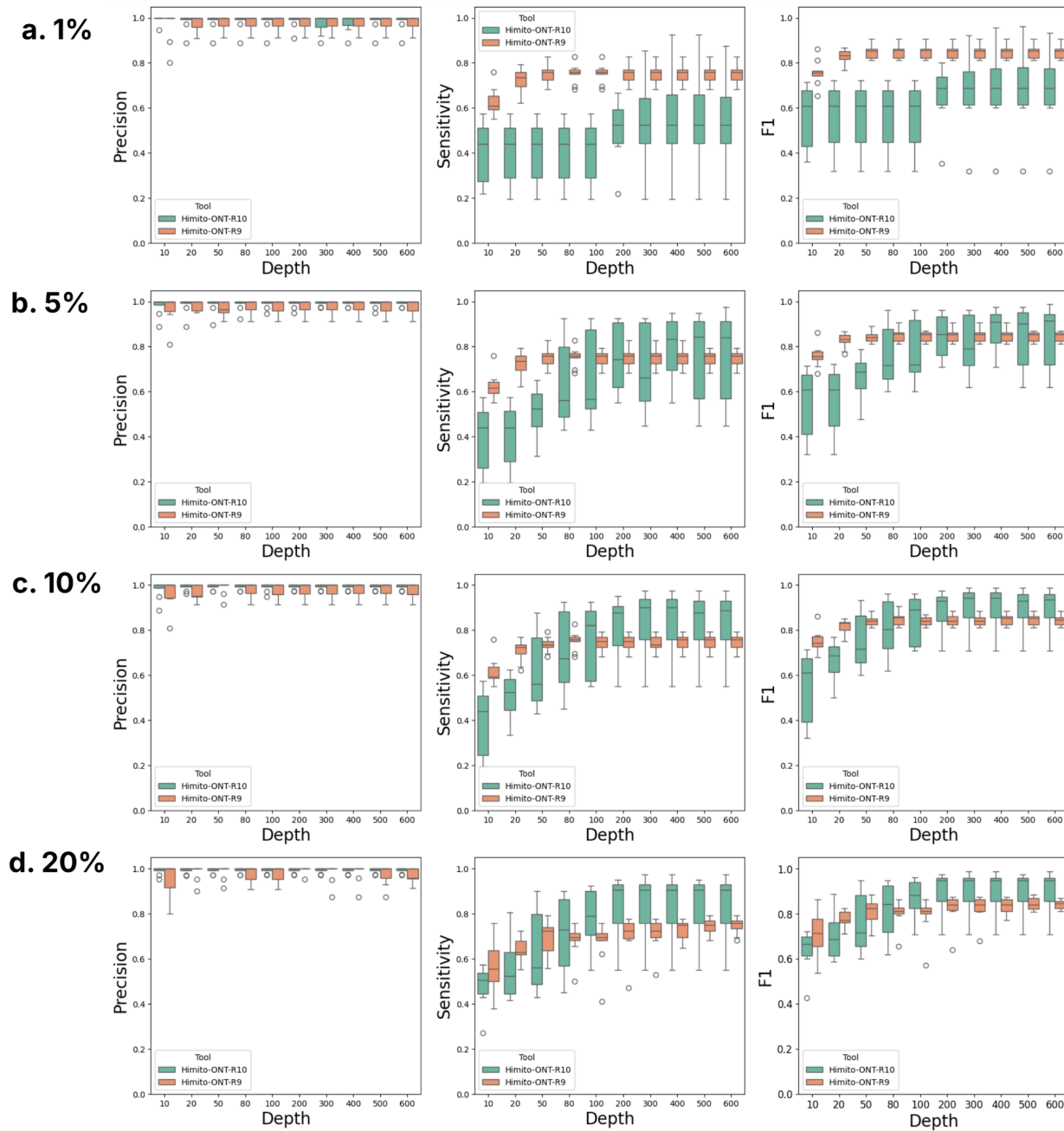
d. 20%



Supplementary Figure 7 Mitosaw and Himito comparison with the same detection limits

By changing minimal-MAF from 0.1 to 0.01, Mitosaw's increased sensitivity comes at the expense of reduced precision. Using in silico experiments, heteroplasmic frequencies at 0.01 (a), 0.05 (c), 0.1 (d), and 0.2 (e) were evaluated. The precision, sensitivity, and F1 score were calculated for each frequency. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within 1.5 * interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers. Downsampling seed -R = 1; Mitosaw —minimum-maf = 0.01, Himito

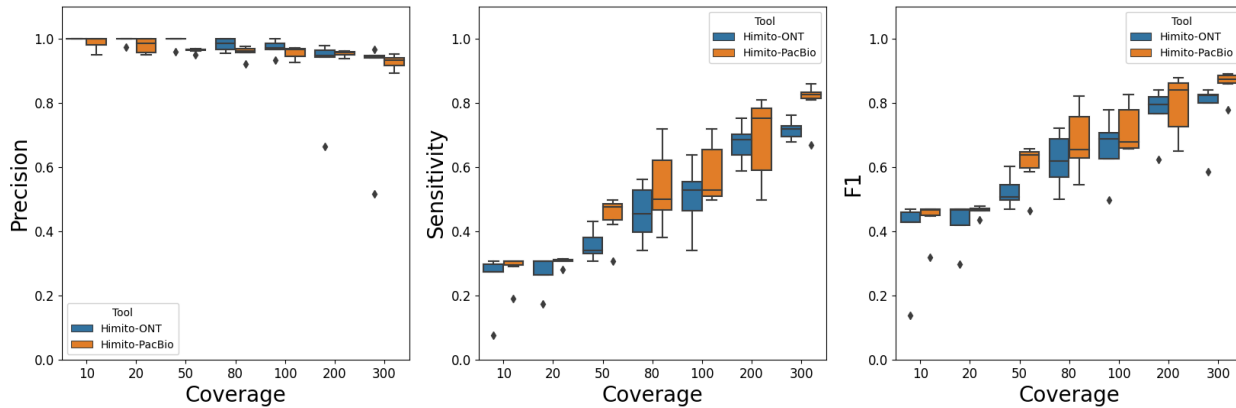
parameter used here is `-vaf-threshold = 0.01`, `--heteroplasmic-frequency-threshold = 0.5`, `--p-value-threshold=0.01`. Sample Size N = 25



Supplementary Figure 8 Himito variant calling performance using ONT-R9 and ONT-R10 data

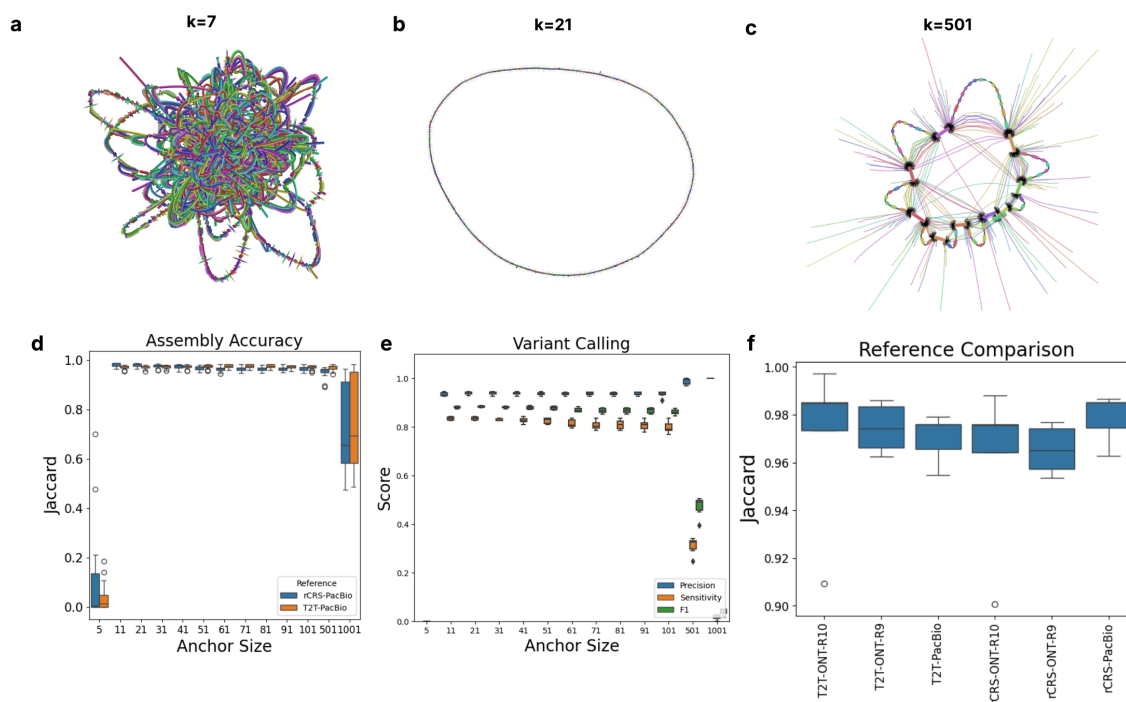
Using the in silico mixture experiments, we optimized the performance for ONT-R9 and ONT-R10 dataset and evaluated the calling performance with heteroplasmic frequencies at 0.01 (a), 0.05 (c), 0.1 (d), and 0.2 (e). Sample Size N = 9. The precision, sensitivity, and F1 score were calculated for each frequency. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within 1.5 * interquartile range from the lower and upper bounds of the box;

individual points beyond the whiskers represent outliers. Downsampling seed -R = 1; Himito default parameter for ONT-R10 `-vaf-threshold = 0.01`, `--heteroplasmic-frequency-threshold = 0.5`, `--p-value-threshold=0.01`. Himito default parameter for ONT-R9 is `-vaf-threshold = 0.01`, `--heteroplasmic-frequency-threshold = 0.5`, `--p-value-threshold=0.0001`.



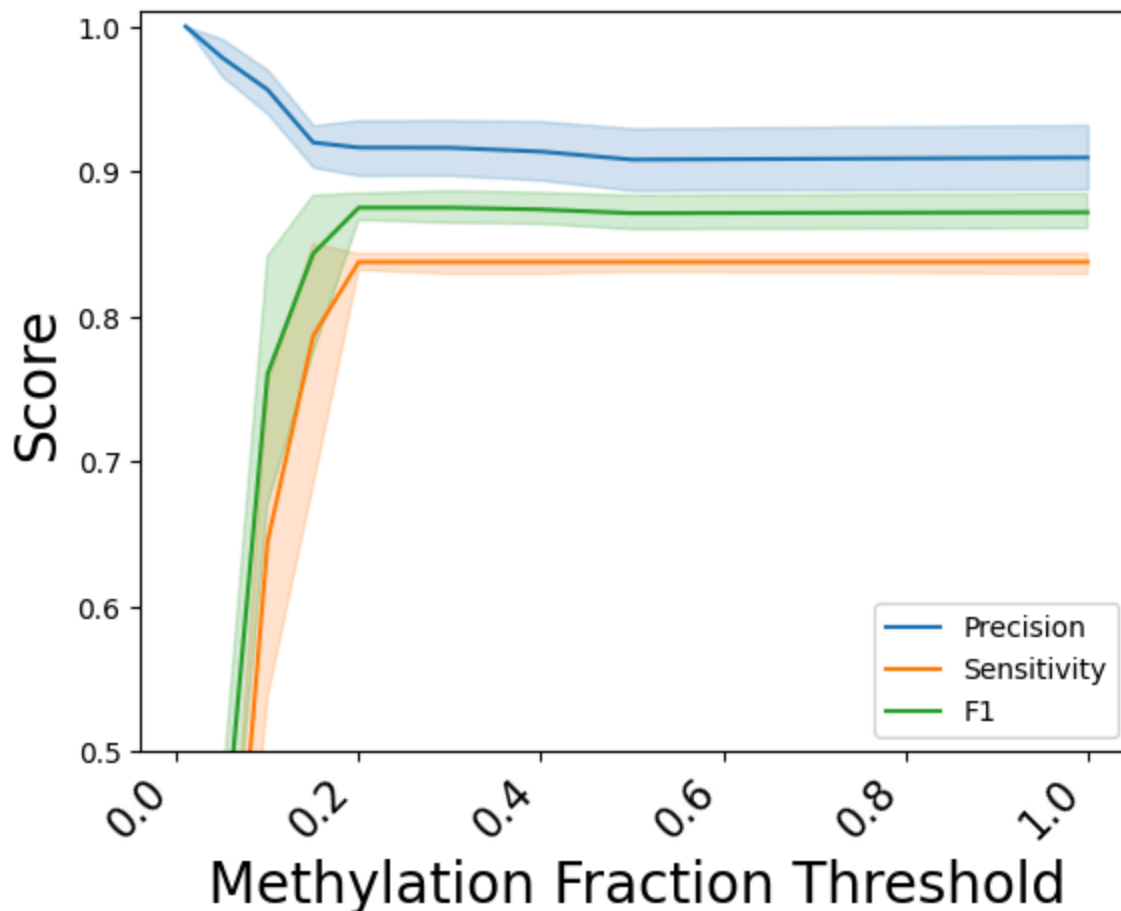
Supplementary Figure 9 HapMap mixture benchmarking using PacBio and ONT long read sequencing technology

HapMap mixture samples, including 6 PacBio and 5 ONT samples are further used to evaluate the Himito variant calling performance at different sample coverages. The Precision, Sensitivity and Recall are shown in the above figures. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within 1.5 * interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers. PacBio: `-vaf-threshold = 0.01`, `--heteroplasmic-frequency-threshold = 0.5`, `--p-value-threshold=0.01`, ONT: `-vaf-threshold = 0.01`, `--heteroplasmic-frequency-threshold = 0.5`, `--p-value-threshold=0.01`).



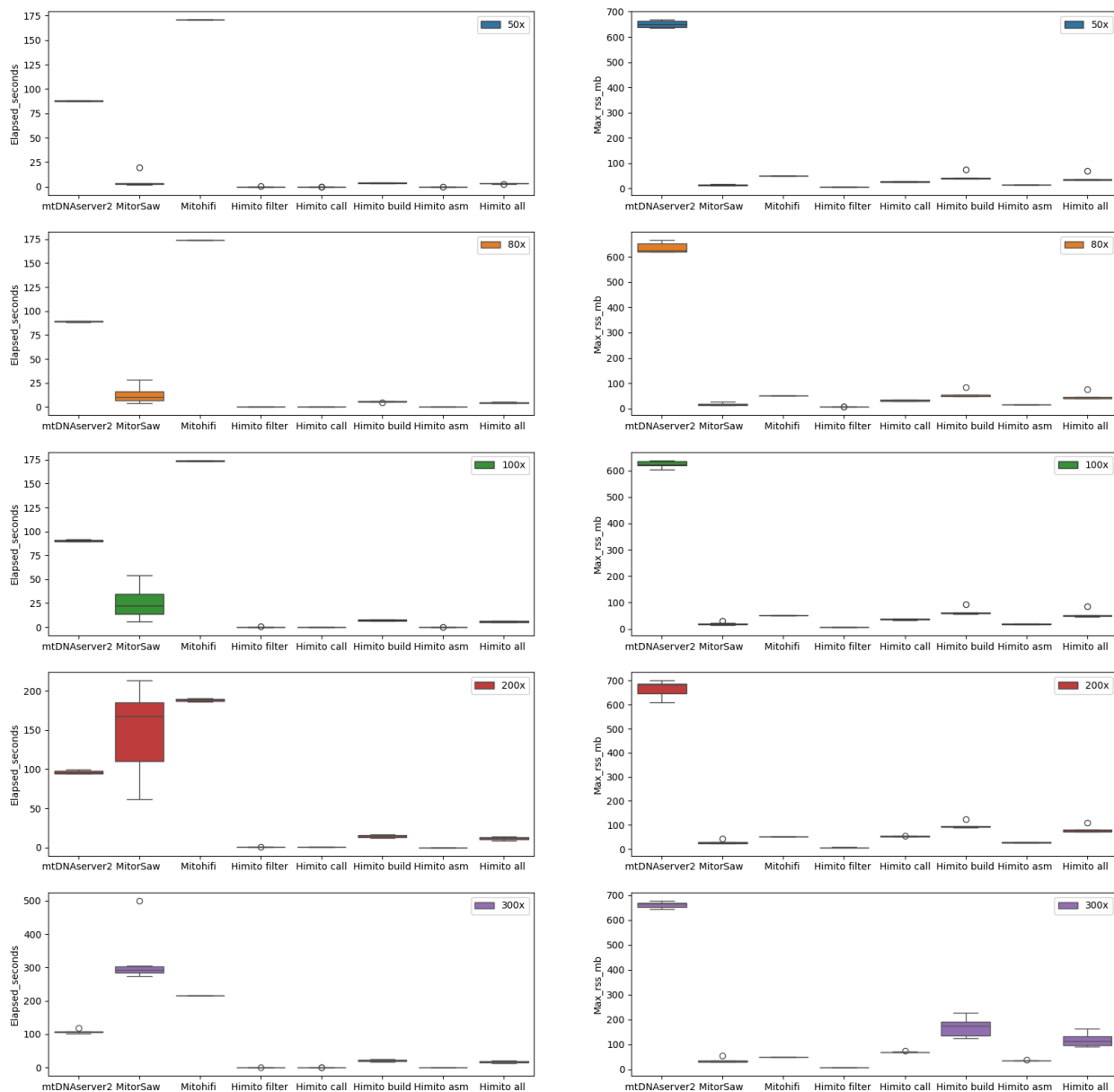
Supplementary Figure 10 The impact of anchor size selection and reference genome choice on graph topology, assembly accuracy, and variant calling performance.

To evaluate the impact of anchor size choices to the graph topology and downstream analysis, we conducted a broad search across a range of kmer-sizes from 5 to 1001 (Supplementary Methods). The assembly quality and variant calling performance are robust when the kmer length ranges from 11-101, and both the assembly and variant calling performance drastically decrease when the kmer size is too small or too large. We further benchmarked the impact of reference choices (rCRS, chm13 chrM) to the assembly quality. Although the use of a different reference genome alters the specific set of anchor sequences and the starting position of the final assembly, the overall assembly accuracy remains consistent. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within 1.5 * interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers.



Supplementary Figure 11. Benchmarking of methylation filtering threshold with the variant calling performance.

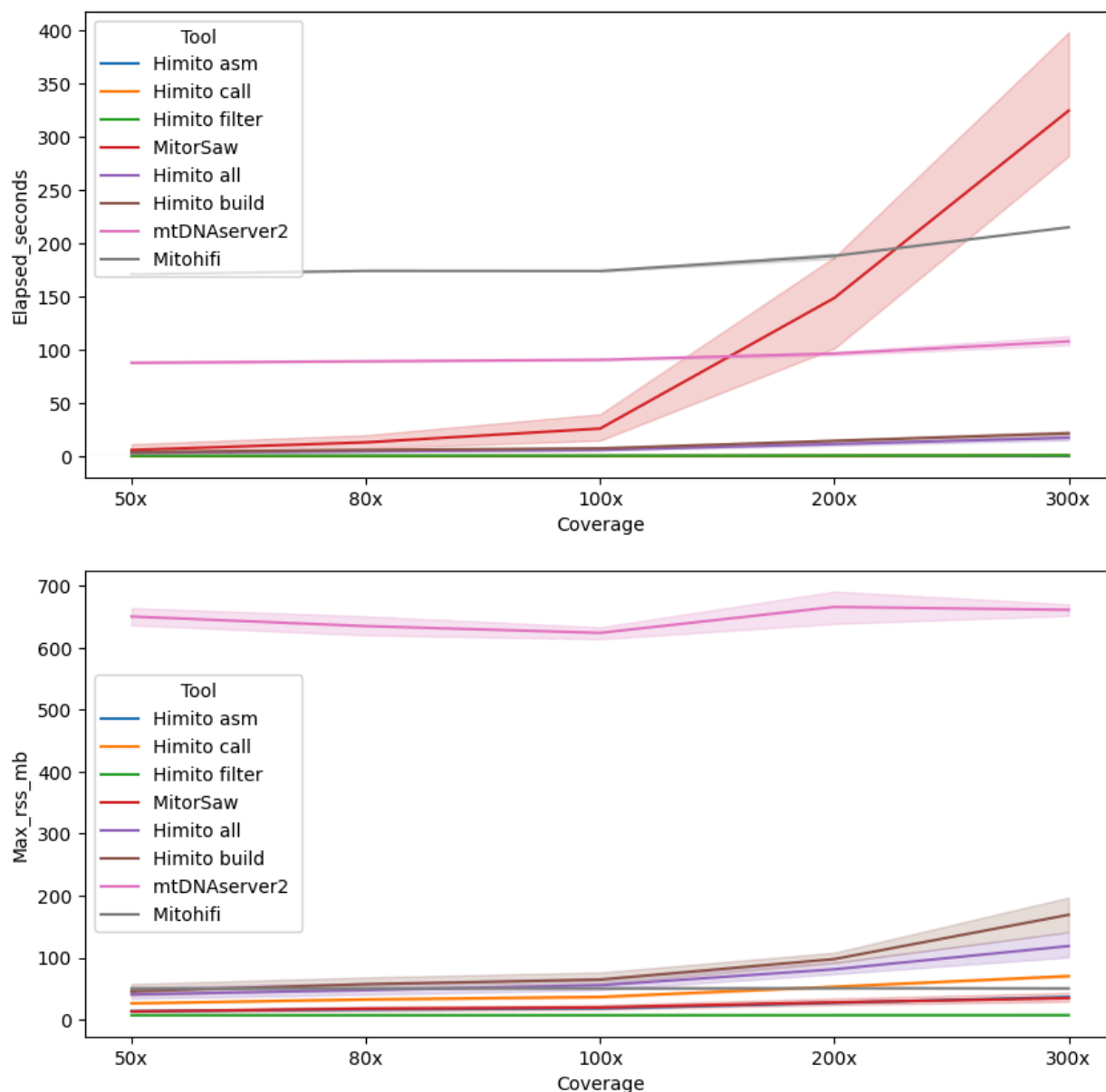
To evaluate the efficacy of our NUMT filtering strategies, we benchmarked Himito's variant calling performance by varying the methylation filtering threshold (`--fraction-max-methylation`) from 0.01 to 1. A threshold of 1 indicates that no reads are filtered based on their methylation status. This experiment was conducted using six HapMap mixture samples with PacBio long read BAMs, downsampled to 600x. Precision slightly improves when the threshold is less than 0.5 and significantly improves when the threshold is less than 0.1. This suggests that filtering by methylation patterns helps reduce false positive calls. However, setting the methylation threshold beyond 0.1 causes a dramatic drop in sensitivity, which implies that the filtering becomes overly aggressive. Solid lines represent the mean value of each metric. Shades represent 95% confidence intervals.



Supplementary Figure 12 Runtime and peak memory usage.

To evaluate Himito's computational efficiency, specifically its run-time and memory usage, we used downsampled HapMap mixture PacBio data at various coverages (10x, 20x, 50x, 100x, 200x, 300x, 400x, 500x) to test the wall clock time (*Elapsed_seconds*) and peak memory usage in Megabyte unit (*Max_rss_mb*) of Himito (both overall and broken down by module), Mitosaw, MitoHiFi, and mtDNAServer2 on a Broad Institute server (see Methods). Himito proved to be the fastest among the tested tools (MitoHiFi, Mitosaw, and mtDNAServer), as measured by elapsed time. In terms of memory consumption, Himito demonstrated relatively low RAM usage, with peak usage consistently remaining below 200Mb and showing only a slight increase with higher sample coverage. In contrast, mtDNAServer2 required significantly more peak RAM. For Himito, the highest RAM requirement occurs during the Build module, which is necessary for loading all the read sequences into memory to construct the sequence graph. HapMap mixture samples $n = 6$. All box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers

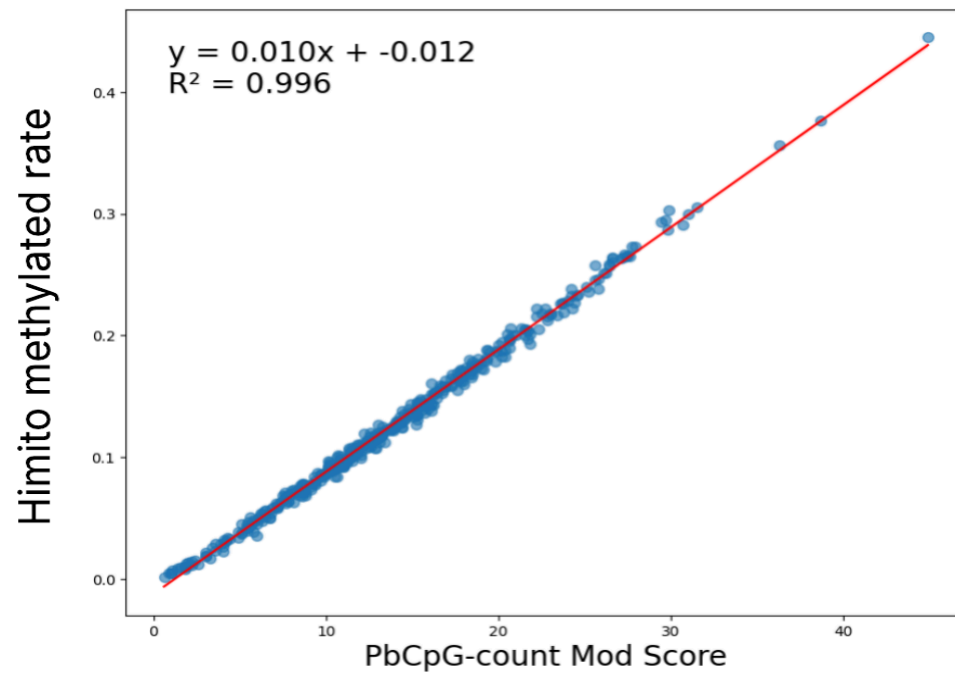
extending to data extremes within 1.5 * interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers.



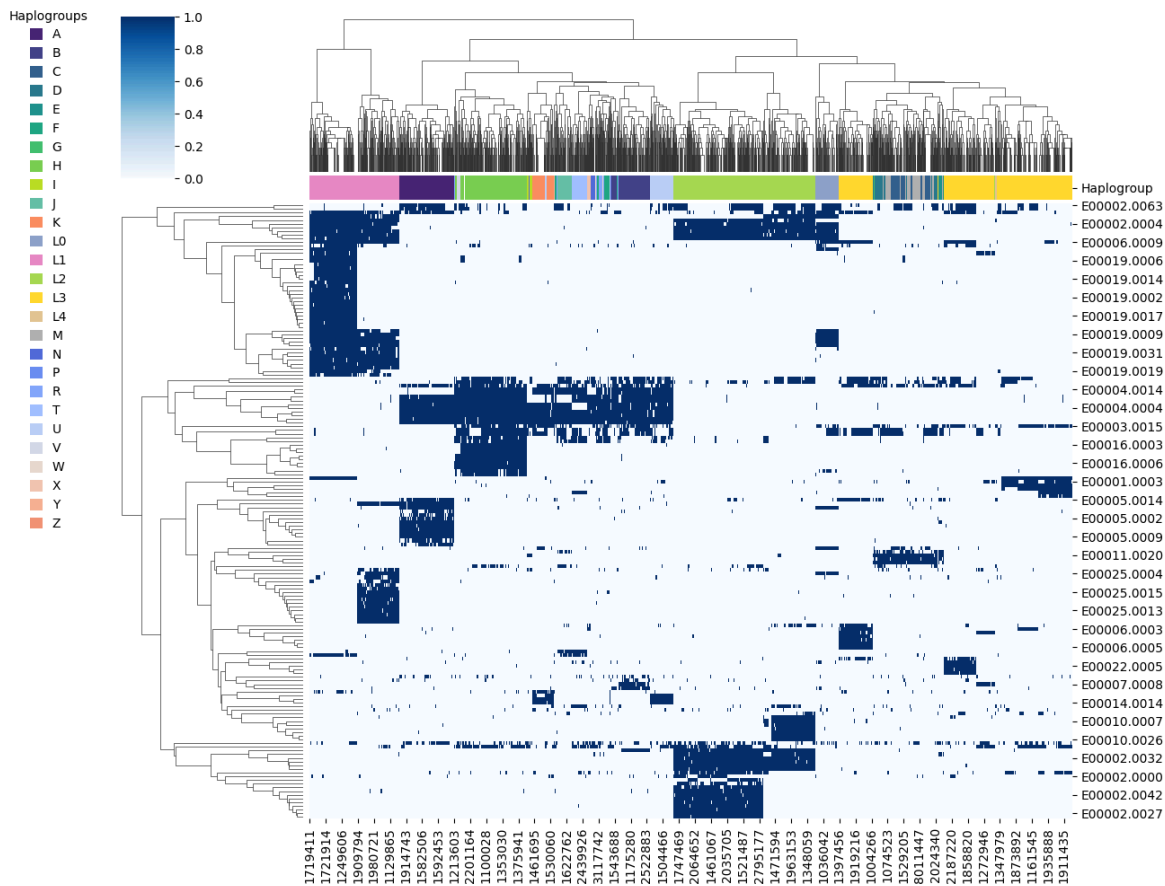
Supplementary File 13 The impact of sample coverage to runtime and peak memory usage.

We further benchmarked the impact of sample coverage on runtime (elapsed-seconds) and peak memory usage (Max_rss_mb). Himito consistently demonstrated the fastest runtime across all tested sample coverages. MitoHiFi and mtDNAserver2 required significantly longer processing times. Notably, Mitosaw's runtime increased considerably for samples with coverage exceeding 100x. mtDNAserver2 consumed significantly more memory compared to the other tools. Both MitoHiFi and Himito (all pipeline) showed a minor increase in memory usage when the sample coverage exceeded 200x. (HapMap mixture samples n = 6) All

box plots show the median (center line), interquartile range for bounds of box (25th to 75th percentiles), and whiskers extending to data extremes within $1.5 \times$ interquartile range from the lower and upper bounds of the box; individual points beyond the whiskers represent outliers.

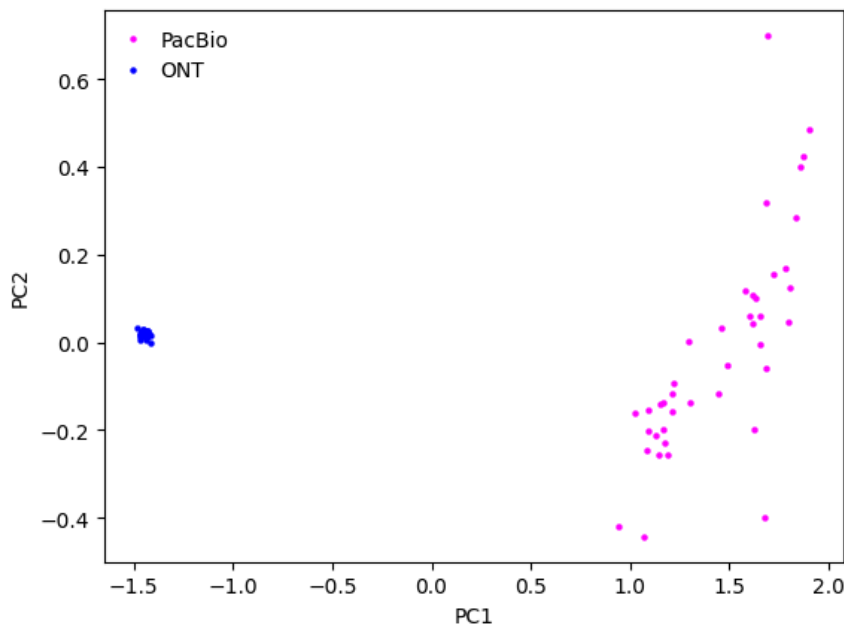


Supplementary Figure 14, Benchmark of Himito Methyl module against PbCpG-count Mode.

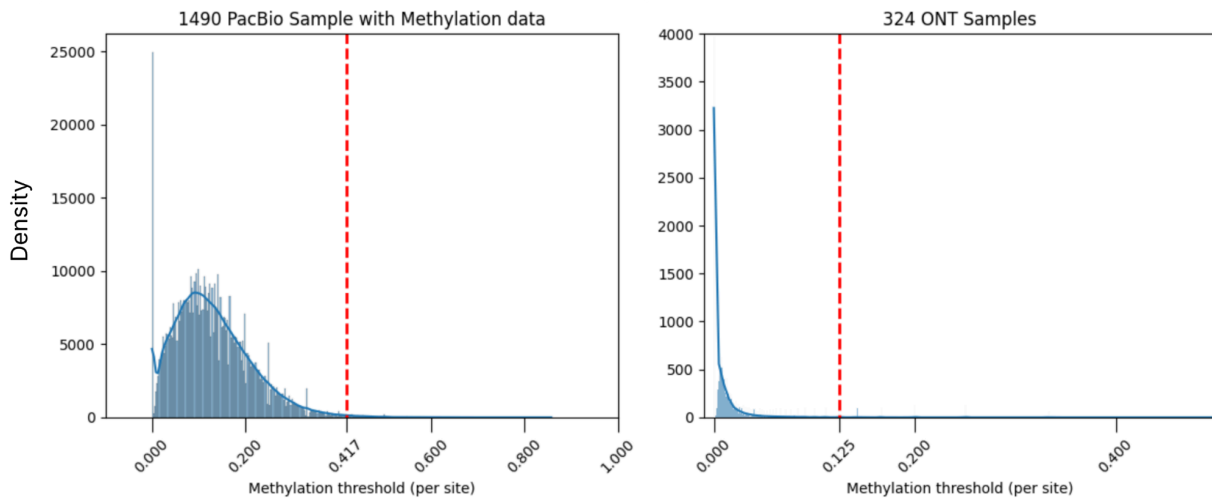


Supplementary Figure 15 Hierarchical clustering of the graph vertices presence-absence matrix in each assembly.

A binary matrix was created to represent the presence or absence of each graph vertex within individual assemblies. Hierarchical clustering of this matrix revealed an automatic grouping of samples according to their haplogroups. Graph vertices with coverage below 100 or above 1000 were excluded from this analysis (sequences less than 100 individuals supported or shared by more than 1000 individuals).

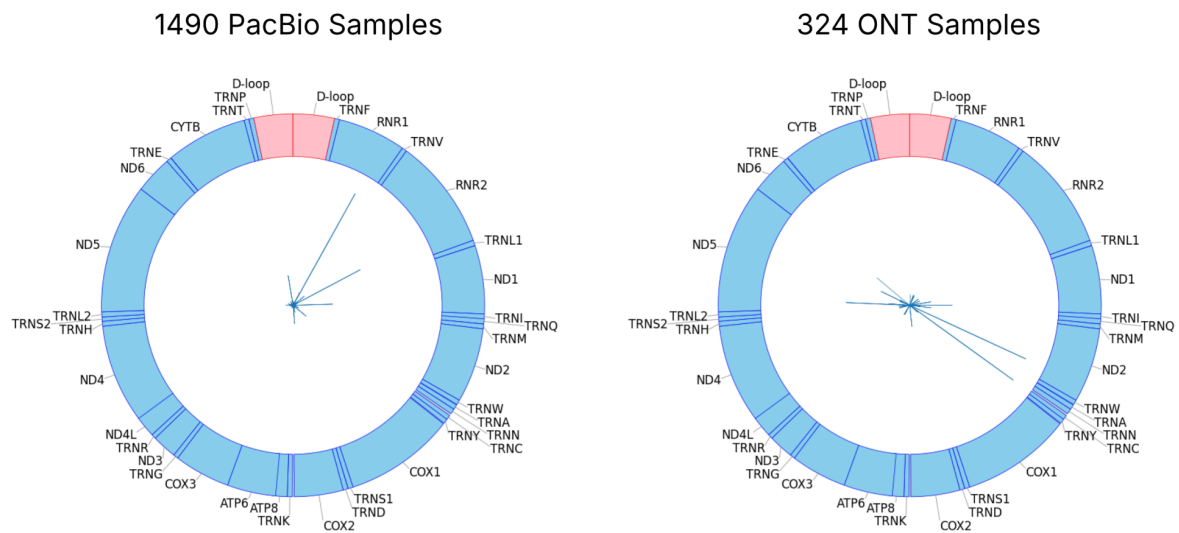


Supplementary Figure 16, PCA plot for CpG modification rate in 41 paired PacBio and ONT samples in AoU v8 cohort.

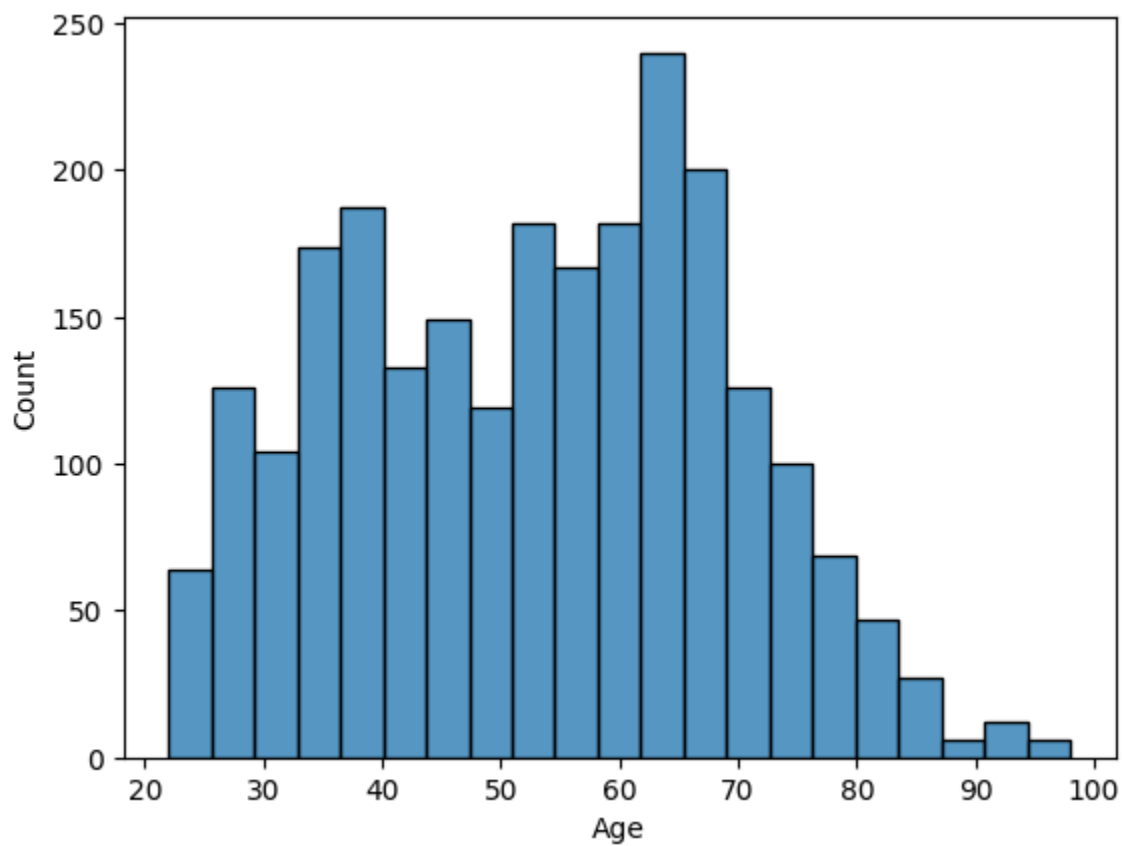


Supplementary Figure 17, The methylation modification score distribution in 1490 All of us v8 PacBio samples and 324 ONT samples.

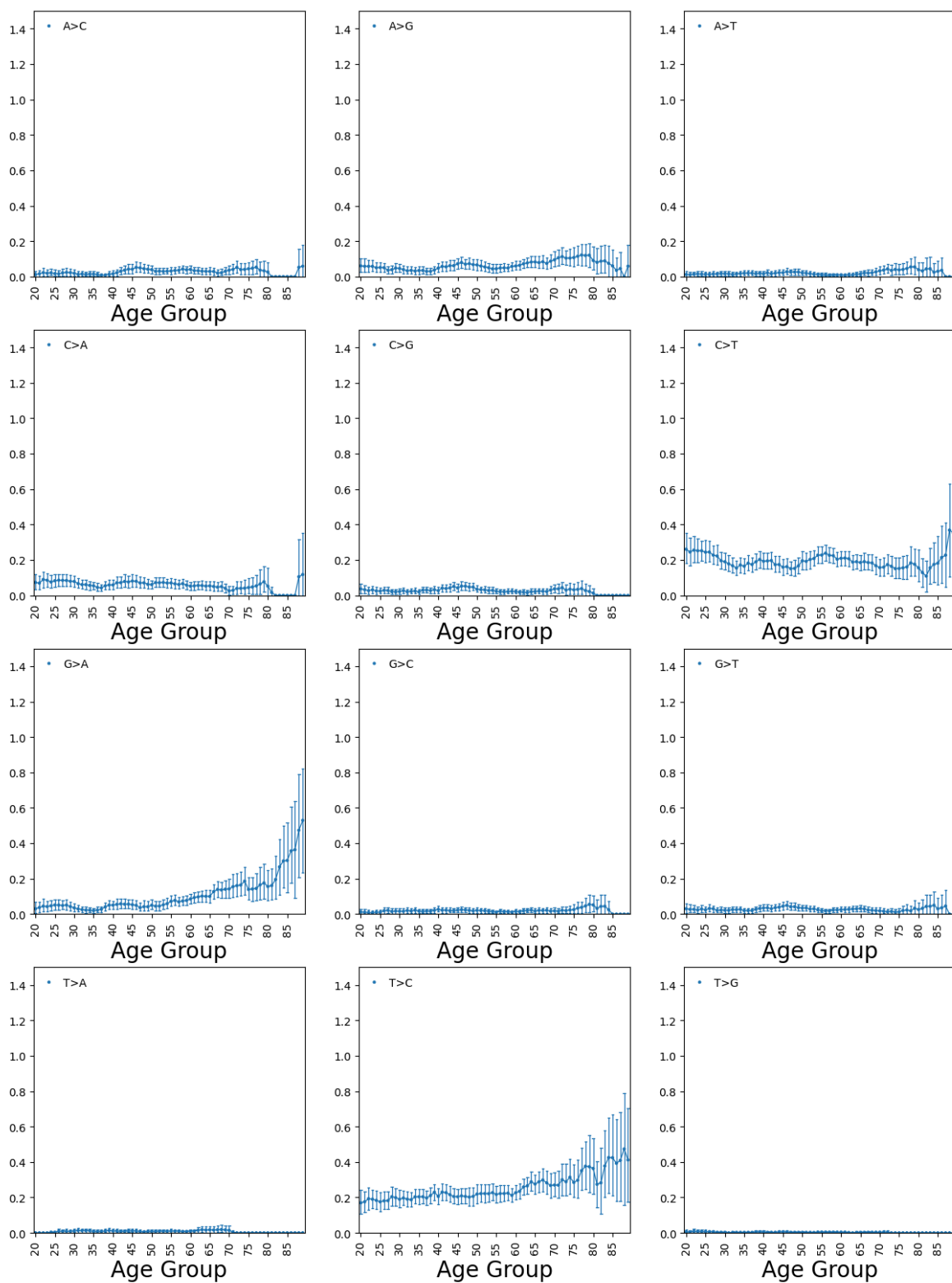
Methylated count normalized by coverage of the CpG sites across all the 1490 PacBio long-read samples and 324 ONT samples. The red lines denote the threshold determining the 5mCpG is hyper-methylated (99.5th percentile of the distribution).



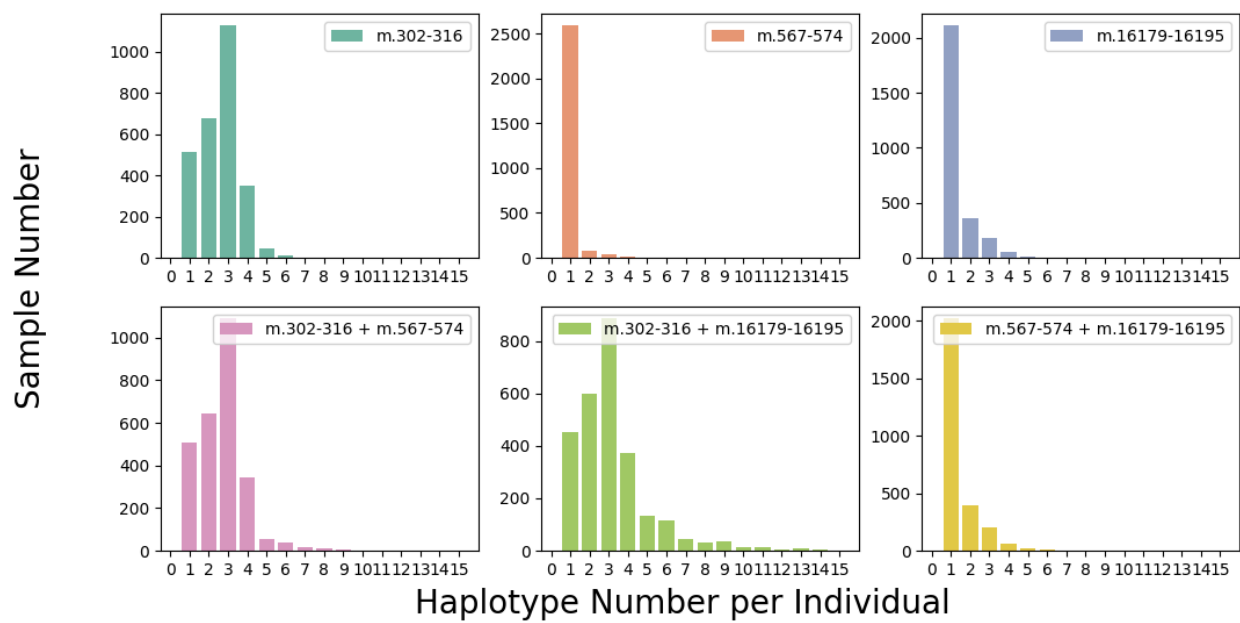
Supplementary Figure 18, Recurrent hypermethylated CpG sites across PacBio (modification rate threshold = 0.417) and ONT samples (modification rate threshold = 0.125). N = 2422.



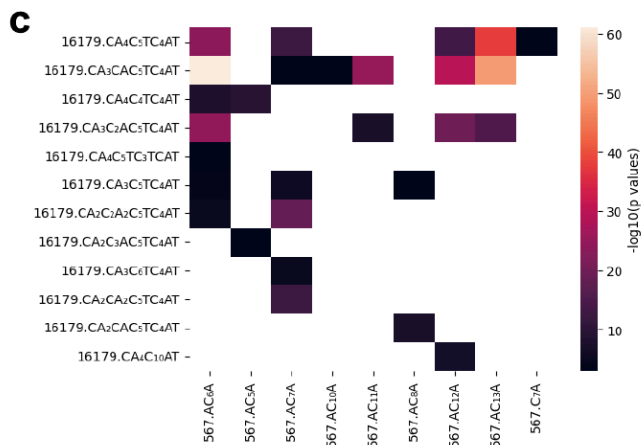
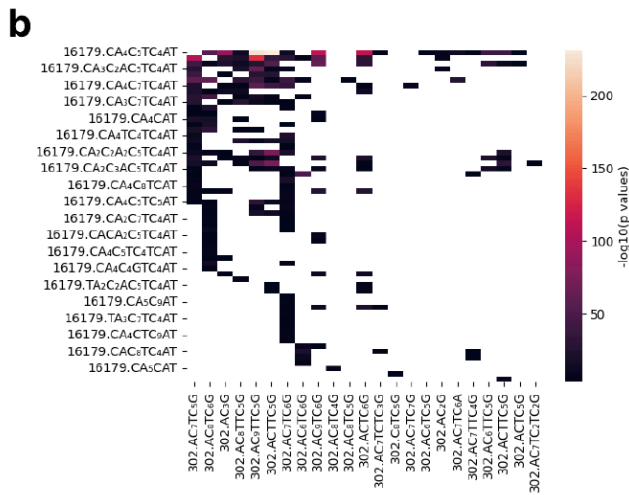
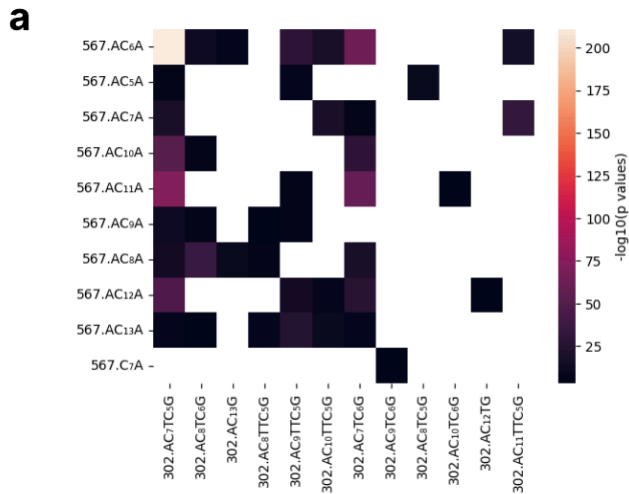
Supplementary Figure 19 Age distribution in AoU v8 PacBio long read samples (n = 2422).



Supplementary Figure 20. The number of different variant types in Age groups (sliding window of 10). The variant types are shown in the legend. The solid line represents the mean values, error bars represent the 95% confidence interval.

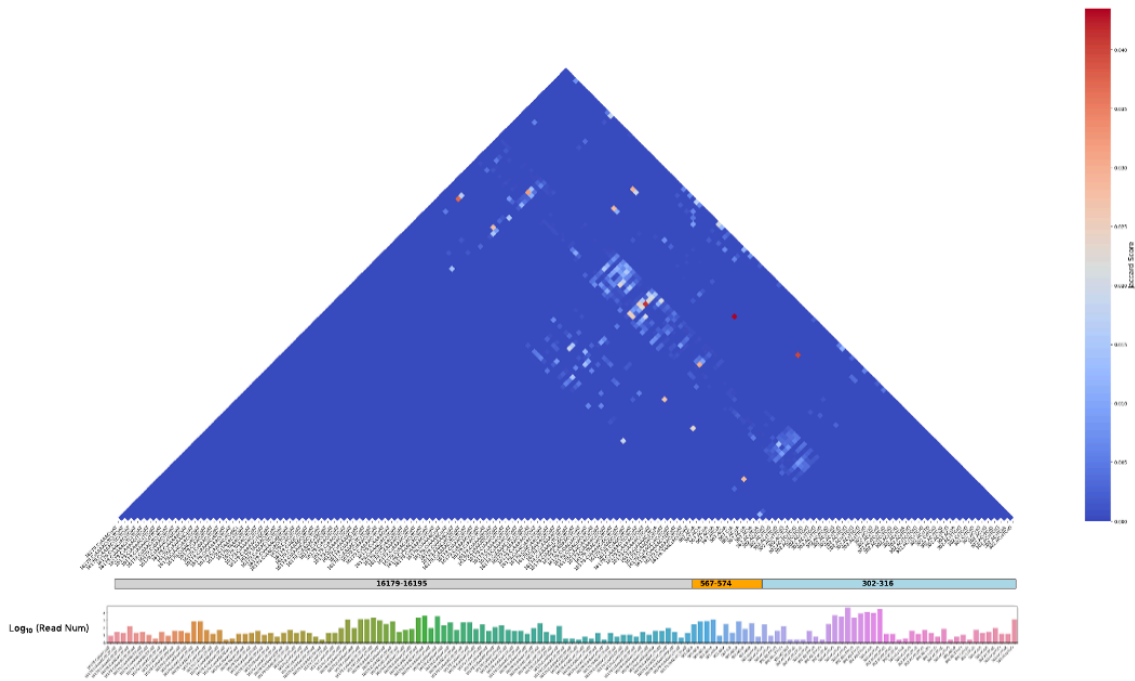


Supplementary Figure 21. Number of length heteroplasmic variants (*chrM*:302-301, *chrM*:567-574, *chrM*:16179-16195) per individual and the joint allele distribution.



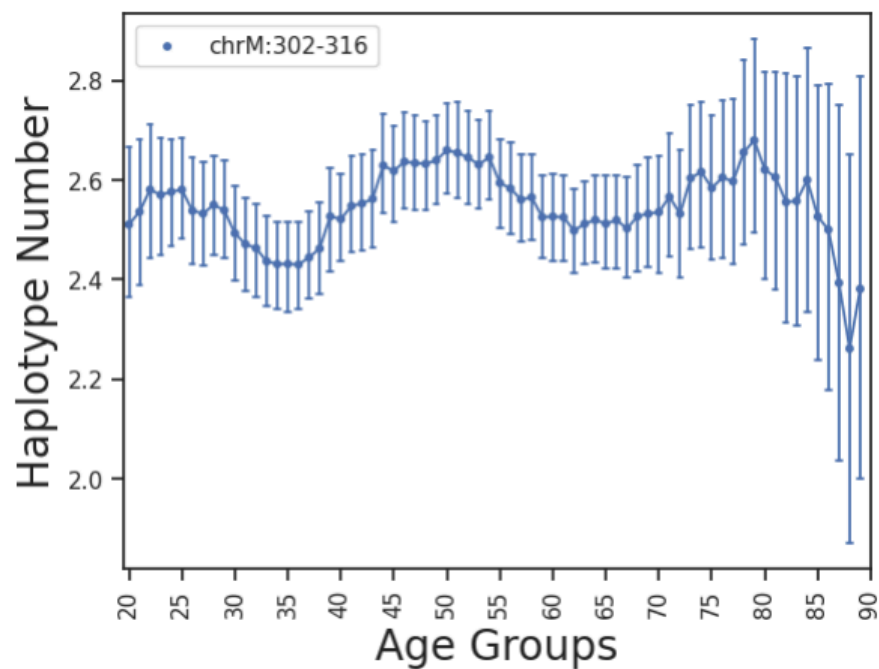
Supplementary Figure 22. Fisher's Exact test for the association for length heteroplasmic alleles (*chrM*:302-301, *chrM*:567-574, *chrM*:16179-16195) across 229985 reads.

Allele pairs with p -value < 0.001 are included in these plots.



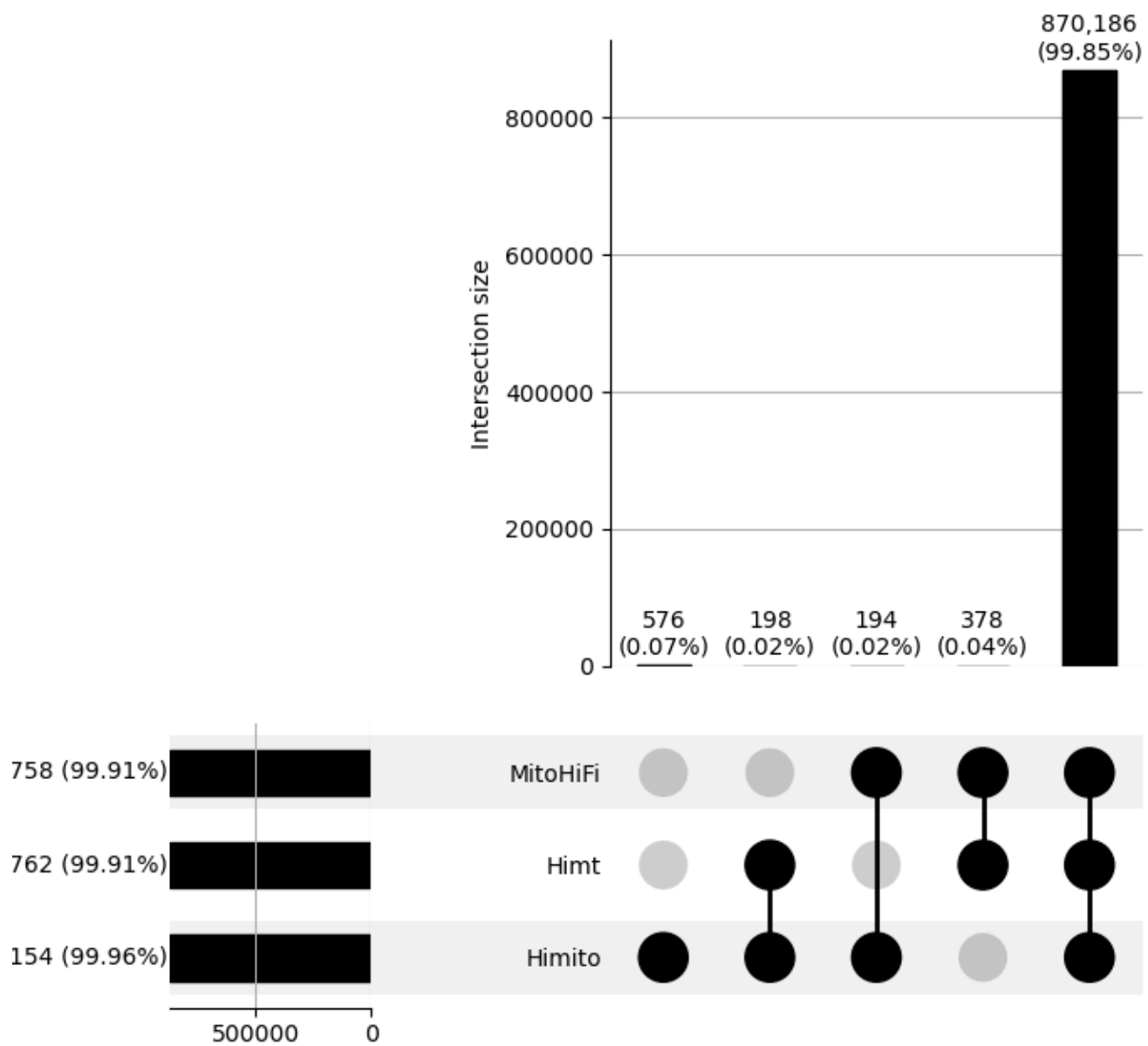
Supplementary Figure 23 Co-occurrence of length heteroplasms.

We aggregated the binary matrix of variants versus reads in the 2422 AoU v8 PacBio samples and test the co-occurrence of 3 length heteroplasms in *chrM*:302-301, *chrM*:567-574, *chrM*:16179-16195. The Jaccard score is calculated by the number of reads supporting the co-occurrence of two length heteroplasms, divided by the total number of reads supporting each length heteroplasmy. The bar plot shows the number of reads supporting each allele present in the heatmap.



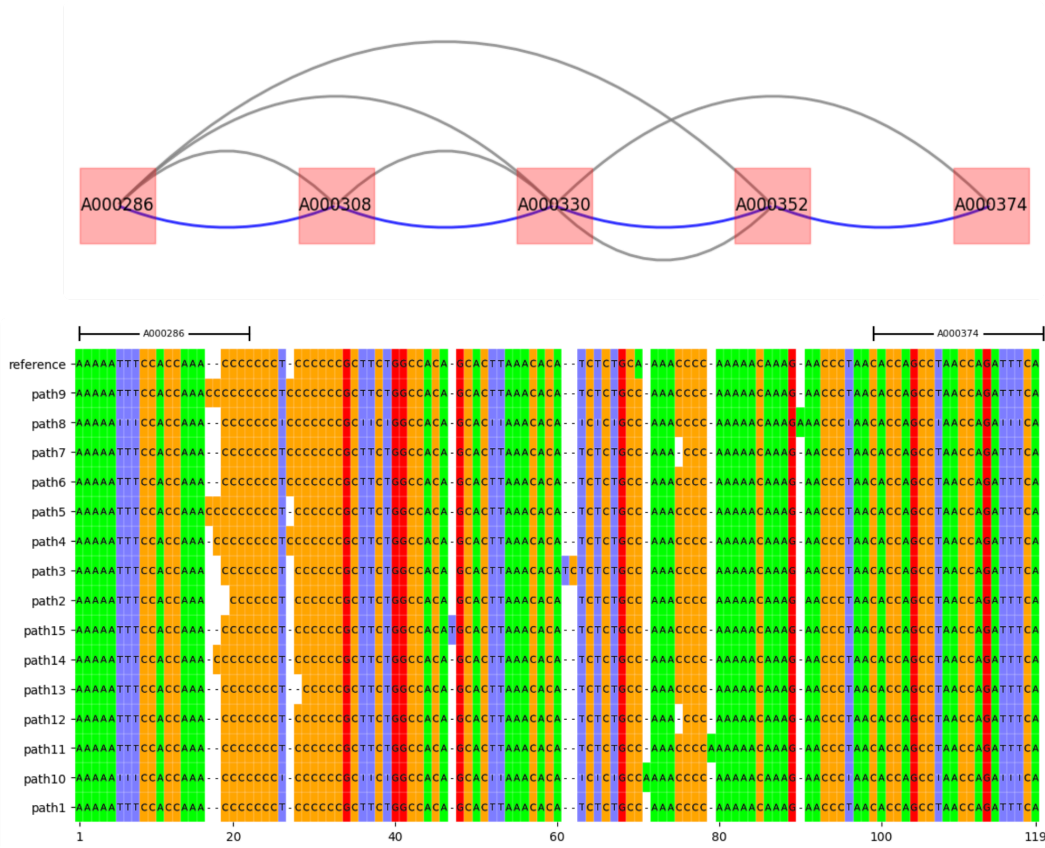
Supplementary Figure 24 Haplotype Number of chrM:302-316 in Age groups (sliding window of 10).

The haplotype is reconstructed by stitching the alternative alleles to the reference sequences according to the variant-read binary occurrence matrix of each sample ($n = 2422$). The solid line represents the mean value, and the error bars represent 95% confidence intervals.

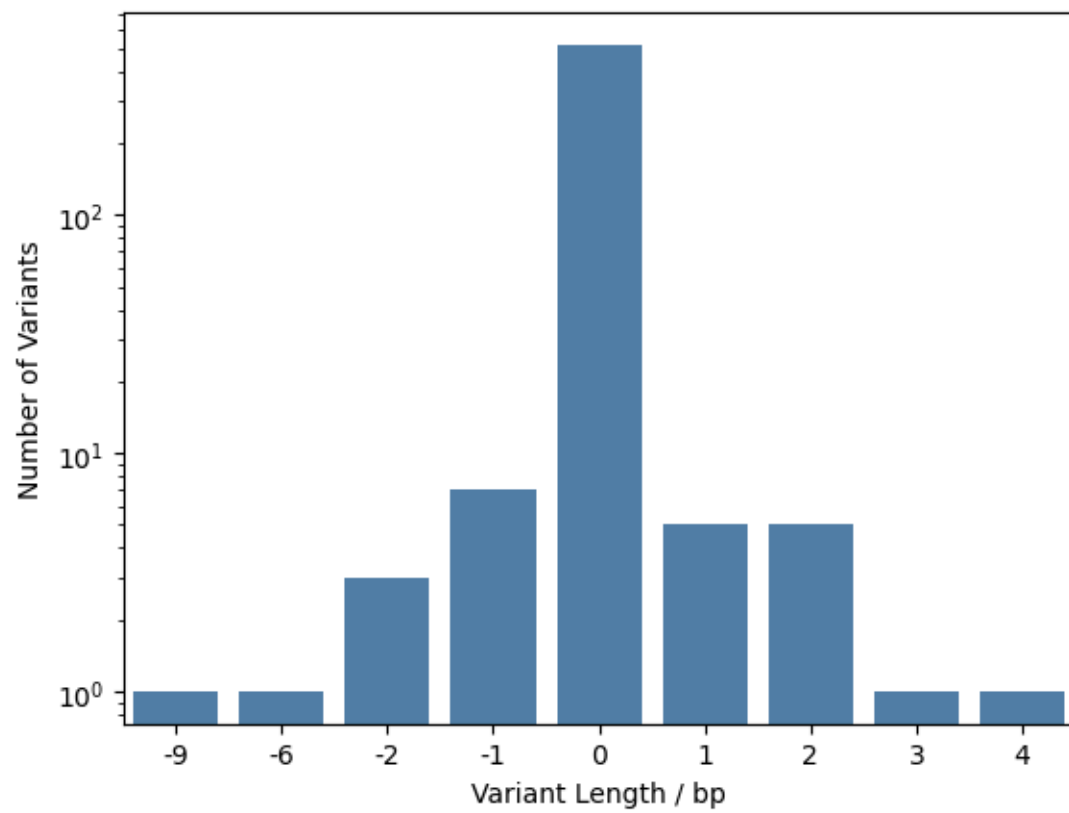


Supplementary Figure 25 Kmer composition of Himito, MitoHiFi and HiMT assemblies for *Lycopodium japonicum* mitochondrial genome (n = 1, SRA accession: SRX20558281)

HG005 chrM:302-315



Supplementary Figure 26 Graph topology and sequence variations in *chrM*: 302-315 in HG005.



Supplementary Figure 27 Variant length distribution in All of Us v8 PacBio samples.

Supplementary Table 1 In silico mixture experiment design (PacBio samples)

In silico mixture experiments were designed to evaluate the heteroplasmic variant calling performance at predefined frequencies. We mixed HG01109 to 25 HPRC Y1 samples at predefined frequencies (0.01, 0.05, 0.1, 0.2). The union of donor's minigraph-cactus chrM calls were merged as the truth set.

Donor1 ID	Donor1 Coverage	Donor2 ID	Donor2 Coverage	Donor2 Fraction
HG00621	622.692	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG00673	644.692	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG00733	2571.58	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG01123	1893.54	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG01243	1108.26	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG01258	1636.24	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG01358	2452.52	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG01361	1422.07	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG01891	975.373	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02055	1261.75	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02145	655.189	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02148	1178.57	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02257	627.417	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02486	1442.33	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02572	1669.67	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02630	2604.94	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02717	2454.54	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02723	1599.15	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02818	2801.48	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG02886	2732.39	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG03453	1156.96	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG03486	1497.44	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG03492	739.205	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG03540	1357.33	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
HG03579	1737.04	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]
NA20129	1035.53	HG01109	164.543	[0.01, 0.05, 0.1, 0.2]

Supplementary Table 2 In silico mixture experiment design (ONT-R10)

In silico mixture experiments were designed to evaluate the heteroplasmic variant calling performance at predefined frequencies. We mixed HG00128 to 9 HPRC Y2 ONT R10 samples at predefined frequencies (0.01, 0.05, 0.1, 0.2). The union of donor's minigraph-cactus chrM calls were merged as the truth set.

Donor1 ID	Donor1 Coverage	Donor2 ID	Donor2 Coverage	Donor2 Fraction
HG00097	7954.9	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00126	7607.85	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00133	6597.09	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00146	8416.31	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00232	8211.48	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00235	8243.61	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00253	8462.27	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00290	8772.94	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]
HG00097	7954.9	HG00128	8360.89	[0.01, 0.05, 0.1, 0.2]

Supplementary Table 3 In silico mixture experiment design

In silico mixture experiments were designed to evaluate the heteroplasmic variant calling performance at predefined frequencies. We mixed HG00099 to 9 HPRC Y2 ONT R9 samples at predefined frequencies (0.01, 0.05, 0.1, 0.2). The union of donor's minigraph-cactus chrM calls were merged as the truth set.

Donor1 ID	Donor1 Coverage	Donor2 ID	Donor2 Coverage	Donor2 Fraction
HG00140	8110.99	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00280	8230.42	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00323	7946.57	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00408	7848.75	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00423	7839.55	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00544	7852.79	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00558	7821.35	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00597	7848.9	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]
HG00609	8116.88	HG00099	7884.13	[0.01, 0.05, 0.1, 0.2]