Mitorsaw: Long-read sequencing secondary analysis optimized for the mitochondrial genome

Abstract # 4050F

J. Matthew Holt, Christopher T. Saunders, William J. Rowell, and Michael A. Eberle
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Background

- The human mitochondria is a circular fragment of DNA that is approximately 16 kb long. Despite the relatively short length, it contains over 30 genes, many of which are associated with human disease.
- Mitochondria have several genomic features that researchers are interested in identifying:
  - Small variants: ~130 confirmed pathogenic mutations[1]
  - Structural variants: ~200 structural variants in the MITOMAP database[1]
  - Heteroplasmy: Variants that are not present in all copies of the mitochondria
  - NUMTs: Fragments of mitochondrial DNA that have been incorporated into the nuclear genome
- Long-read sequencing captures full-length mitochondrial DNA sequences[2], enabling easier identification of interesting genomic features.

Mitochondrial analysis of long-read sequencing with Mitorsaw

- **Variant detection**: Identifies SNVs, indels, and structural variants from HiFi sequencing, reporting both homoplasmic and heteroplasmic variants with their observed allele fractions
- **Haplotype analysis**: Identifies the full-length mitochondrial sequences from the HiFi reads
- **Custom visualizations**: Generates custom IGV sessions with re-alignments against cyclic reference genome and consensus mitochondrial sequences
- **Avoids pitfalls of mitochondrial analysis**:
  - Filters NUMTs prior to variant calling, removing many false positives
  - Re-maps reads to cyclic genome, recovering variants near the ends of the reference

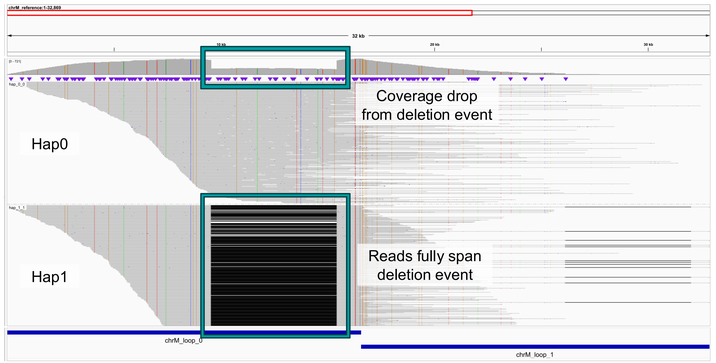


Figure 1. Example heteroplasmic deletion. This is a custom IGV session generated by Mitorsaw, where the HiFi reads have been re-mapped to a cyclic version of the reference mitochondria (tracks "chrM_loop_0" and "chrM_loop_1"). The heteroplasmic deletion event is visually apparent in the second haplotype (Hap1) reported by Mitorsaw with several reads fully spanning the deletion event.

Mitorsaw improves on DeepVariant mitochondrial calls

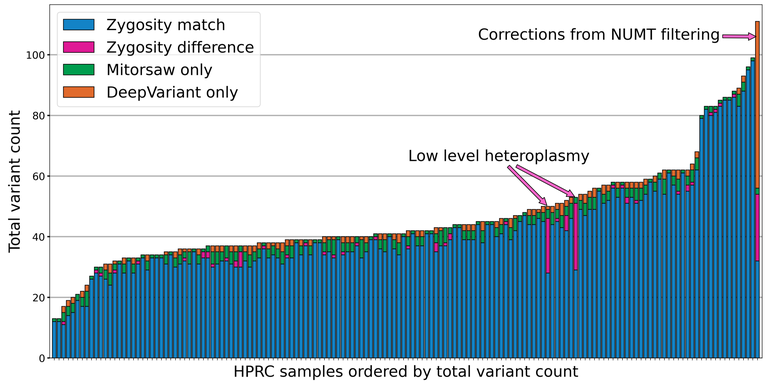


Figure 2. HPRC dataset analysis. We compared the variant calls from Mitorsaw (heteroplasmy-aware caller) and DeepVariant (diploid caller) for 152 HPRC samples sequenced with PacBio HiFi sequencing. For all shared variants, the zygosity of the call is compared to determine if they are equivalent. Three datasets show elevated discordance, two of which are caused by heteroplasmy levels below the default cutoff (10%). The last is enriched for false positives from DeepVariant that are caused by unfiltered reads from a NUMT event in the dataset.

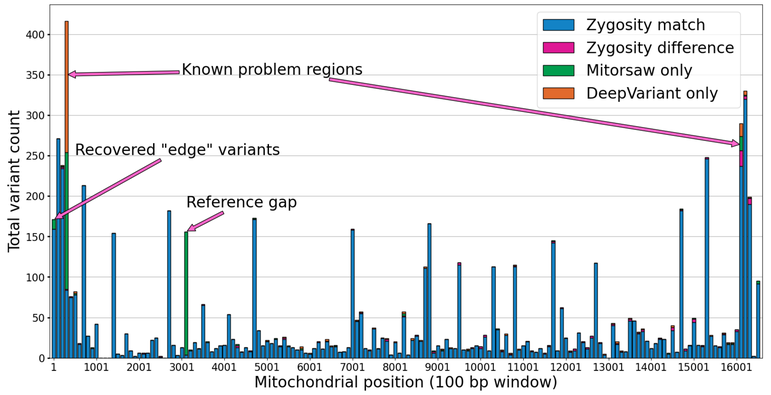
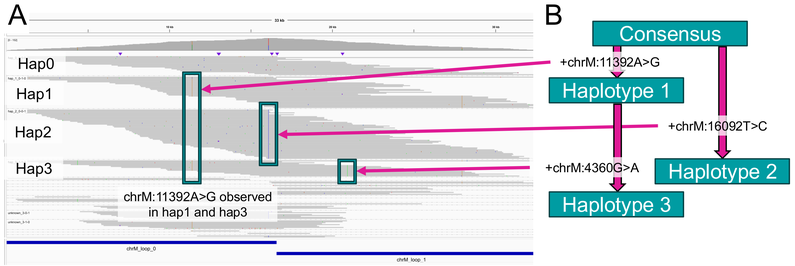


Figure 3. Spatial analysis of variant calls. We sorted the same variants from Figure 2 by their positions along the reference mitochondria. The vast majority of discrepancies fall into four regions, three of which are known problematic regions or reference gaps that are generally ignored. The fourth region is at the start of the reference mitochondria, where DeepVariant often misses variants (false negatives) due to technical limitations of the algorithm. Mitorsaw's algorithms enable these edge variants to be recovered.

Mitorsaw constructs heteroplasmic haplotypes



C

Variant	Allele fraction
Homoplasmic variants (34)	~100%
chrM:4360G>A	15.6%
chrM:11392A>G	43.5%
chrM:16092T>C	42.0%

D

Haplotype	Heteroplasmies	Sample fraction
Consensus (Haplotype 0)	--	16.4%
Haplotype 1	chrM:11392A>G	27.1%
Haplotype 2	chrM:16092T>C	40.9%
Haplotype 3	chrM:11392A>G + chrM:4360G>A	15.6%

Figure 4. HG02300 Mitorsaw results. (A) An image of an IGV session created from Mitorsaw results. The reads are re-mapped to a cyclic consensus sequence and grouped by the haplotype assignment from Mitorsaw. This grouping highlights three heteroplasmic variants within this dataset (teal boxes), including one variant that is shared by both haplotype 1 and 3. **(B)** The inferred phylogeny for the mitochondrial sequences present in this sample. **(C)** The variants detected in the sample with their observed allele fraction. **(D)** The inferred population frequency of each mitochondrial haplotype detected in the sample based on the read observations. These fractions correlate to the read depth observations of the IGV visualization in (A).

Conclusion

Mitorsaw is an analysis tool for HiFi sequencing data that:

- Generates full-length mitochondrial haplotype sequences
- Reports small and structural variants with allele fractions
  - Recovers variants near the edge of the reference
  - Removes false positives caused by NUMTs
- Creates custom visualizations for manual inspection

Mitorsaw

https://github.com/PacificBiosciences/mitorsaw

References

1. Lott, M. T., et al. (2013). mtDNA variation and analysis using mitomap and mitomaster. Current protocols in bioinformatics, 44(1), 1-23.
2. Macken, W. L., et al. (2023). Enhanced mitochondrial genome analysis: bioinformatic and long-read sequencing advances and their diagnostic implications. Expert Review of Molecular Diagnostics, 23(9), 797-814.