Hunting for microsatellite instability in long-read data with Owl

Abstract #1838

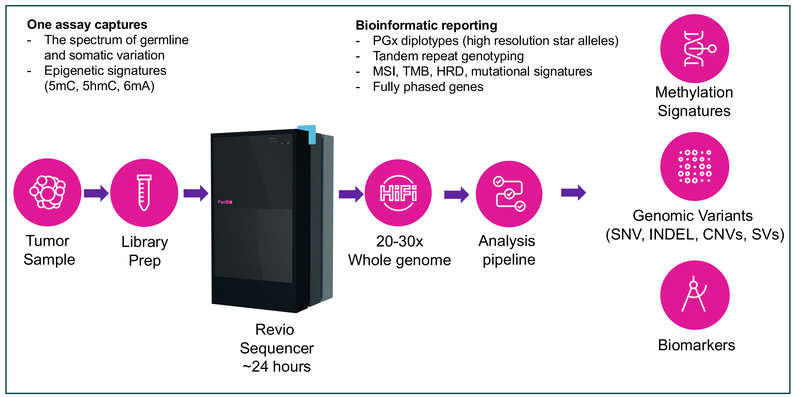
Zev Kronenberg[1], Khi Pin Chua[1], Mark J.P. Chaisson[2,7], Byunggil Yoo[3], Lisa Lansdon[3], William J. Rowell[1], Guilherme de Sena Brandine[1], Egor Dolzhenko[1], Kobe Ikegami[6], Kie Kyon Huang[4], Patrick Tan[4,5,10], Shruti Bhise[6], Everett Fan[6], Mark Mendoza[6], Emily O'Donnell[3], Tomi Pastinen[3], Elizabeth R. Lawlor[6,8,9], Scott N. Furlan[6,9], Midhat S. Farooqi[3], Michael A. Eberle[1]

[1]PacBio, Menlo Park, CA, USA; [2]University of Southern California, Los Angeles, CA, USA; [3]Children's Mercy Hospital, University of Missouri–Kansas City School of Medicine, Kansas City, MO, USA; [4]Duke-NUS Medical School, Singapore; [5]Genome Institute of Singapore; [6]Fred Hutch Cancer Center, Seattle, WA, USA; [7]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA; [8]Ben Towne Center for Childhood Cancer and Blood Disorders Research, Seattle Children's Research Institute, Seattle, WA, USA; [9]Department of Pediatrics, University of Washington, Seattle, WA, USA; [10]Precision Health Research Singapore

Background

Microsatellite instability (MSI) is a key biomarker of mismatch repair deficiency and response to immunotherapy, yet most existing genomic detection methods are optimized for short-read sequencing and rely on small panels of homopolymer markers, limiting the ability to characterize genome-wide and motif-specific patterns of instability. Here we present Owl, a bioinformatic tool for quantifying MSI from PacBio HiFi whole-genome data.

Characterizing cancer genomes with PacBio workflow



Detection of Microsatellite instability

- Use the HiFi Somatic WDL pipeline
  - Map HiFi reads to reference genome
  - Phase reads (haplotag) based on small variants
- Run Owl on the aligned HiFi data (**Figure 1A**)
  - Profile repeats over defined regions (markers and motifs)
  - Measure repeat length variability
  - Summarize variability genome wide

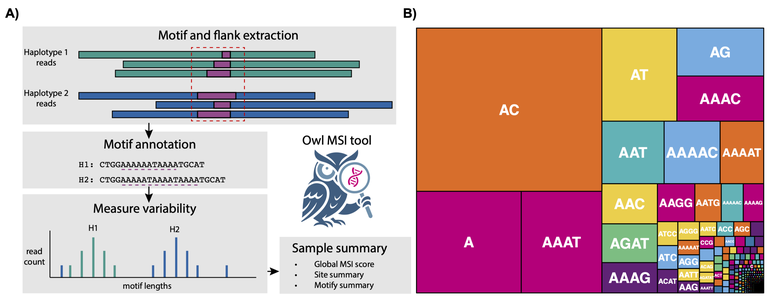


Figure 1. Owl methodology and repeat markers.
(A) Workflow for measuring repeat variability within haplotypes using wrap-around dynamic programming; read length distributions within each allele are summarized using the coefficient of variation. **(B)** Distribution of markers in GRCh38, where each box represents the genomic frequency of a given motif identified with RepeatMasker. Owl considers motifs up to six base pairs in length, not just homopolymers.

HiFi reads characterize somatic repeat variability

- **HiFi reads span entire repeat region (Figure 2).**
- **HiFi reads are very accurate (99.9%)**
- **Repeat lengths are directly measured by HiFi reads**

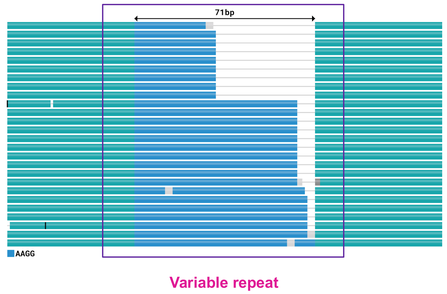


Figure 2. Repeat length distribution for an AAGG locus. Reads for the TC32 Ewing sarcoma cell line (haplotype 2) are shown. Each row is a HiFi read colored by repeat content. Mean repeat length = 58.5 (blue); coefficient of variation = 29. The locus overlaps a distal enhancer-like element (EH38E2599658) in intron 3 of *EZH2*

Characterizing MSI across control and cancer samples

- 131 diverse samples from the Human Pangenome Reference Consortium (**controls**)
- 9 Ewing sarcoma cell lines (**2 MSI**)
- 2 lung cancer cell lines
- 2 gastric cancers **(2 MSI)**
- 1 diffuse astrocytoma (**MSI**)
- 1 melanoma

Empirical thresholding (Figure 3A) shows that an MSI score of 10% clearly stratifies the five suspected MSI samples from healthy controls.

The diffuse astrocytoma sample had a high Owl score of 17%. This sample also had a high DRAGEN score, providing orthogonal validation of our approach.

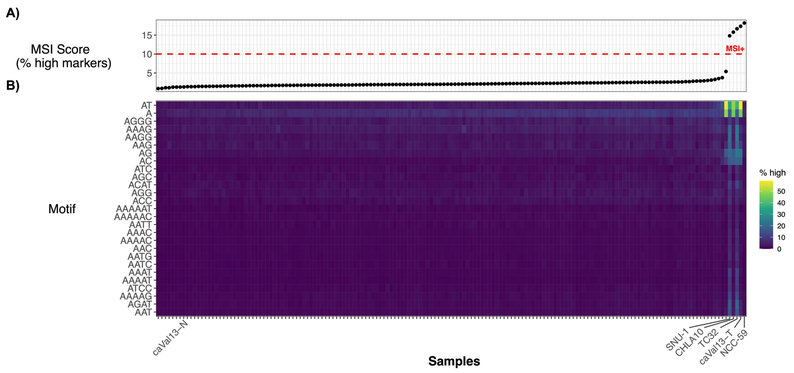


Figure 3. Per-sample MSI scores and motif-specific microsatellite variability.
(A) Owl MSI scores for each sample, sorted from low to high. Two Ewing sarcoma cell lines (CHLA10 and TC32), two gastric cancers (SNU-1 and NCC-59), and one diffuse astrocytoma (caVal13) exceed the 10% MSI threshold. **(B)** Heatmap showing the percentage of high-variability sites per motif, restricted to the most frequent motifs (>500 marker loci). Samples along the x-axis follow the order from the panel above, while motifs along the y-axis are clustered by similarity using `hclust`. The five MSI-high samples are labeled and the diffuse astrocytoma control (N).

Motif specific patterns of instability

- **Ewing sarcoma cell lines (T32, CHLA10) showed numerous outlier motifs (Fig. 3)**
- **Gastric cancers (NCC-59, SNU-1) and astrocytoma (CaVal-13) had fewer, shorter motif outliers**
- **Ewing sarcoma enriched for GGAA unstable repeat motifs**
- **GGAA unstable repeats were enriched for regulator element overlap (permutation test, p = 0.0005)**
- **GGAA motifs are EWS::FLI1 binding sites linked to transcriptional reprogramming**
- **Conventional MSI assays (short-read) focus on homopolymers and missed these patterns**

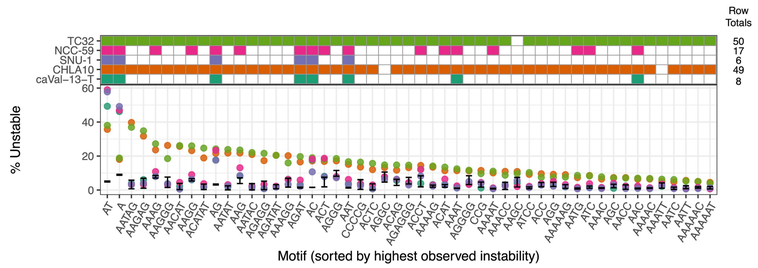


Figure 3. Outlier motifs in the five MSI-high samples.
The top panel marks which sample contains a Bonferroni-corrected p-value (p<0.05) for a beta-binomial one-sided test. The bottom panel shows, for each motif (lexicographically and rotationally minimized), the percentage of sites with high CV. Each point represents one of the five samples. The beta-binomial credible intervals (95%) for the HPRC samples are shown as black lines. Only motifs with >100 markers across the genome are shown.

Conclusion

Owl is a versatile tool for detecting MSI and characterizing motif-specific instability patterns using PacBio HiFi data.

Questions and comments can be directed to zev@pacbio.com

Owl GitHub software

Owl BioRxiv preprint

HiFi Somatic WDL Github

Acknowledgements

The authors would like to thank everyone who helped generate data for the poster.