# Precise characterization of complex repeat regions in cancer genomes

Abstract # 6288
Khi Pin, Chua[1,2], Egor Dolzhenko[1], Tom Mokveld[1], Zev N Kronenberg[1], Seiya Imoto[2], Seiichi Mori[3], Michael A Eberle[1]
*1. Pacific Biosciences of California, Menlo Park, CA, USA 2. Institute of Medical Sciences, University of Tokyo, Japan 3. Japan Foundation for Cancer Research, Japan*

## Somatic repeats variation remains understudied in cancer

The characterization of somatic variation, especially in complex genomic regions, is crucial for understanding the molecular drivers of cancer progression. Accurate PacBio long-read sequencing (HiFi) enables detection of all variant classes, from simple SNVs and INDELs up to complex structural variation, tandem repeats, and changes in epigenetic signatures. Complex and repetitive regions, while fully sequenced by HiFi reads, remain bioinformatically challenging, requiring tailored solutions. Here, we describe new tools to genotype understudied repetitive regions in cancer genomes, a task that has historically posed significant challenges for short-read sequencing.

## HiFi reads allow accurate genotyping of repeats

Recent advances in both the HiFi sequencing systems and bioinformatics algorithms are now allowing researchers to genotype repeats at high accuracy (Dolzhenko E. et. al. 2024). In this work, we attempt to use TRGT and trgt-denovo (in duo-mode) to genotype somatic repeats expansion in 8 cancer cell lines with matching normal (HiFi dataset publicly available from Park et. al. 2024). Next, we applied the workflow to a set of 50 real cancer samples to discover recurrent repeats expansion. We genotyped more than 4.8 millions loci using a catalogue recently published (Weisburd, Dolzhenko, et al. 2024) which included variation clusters determined from a set of healthy samples.

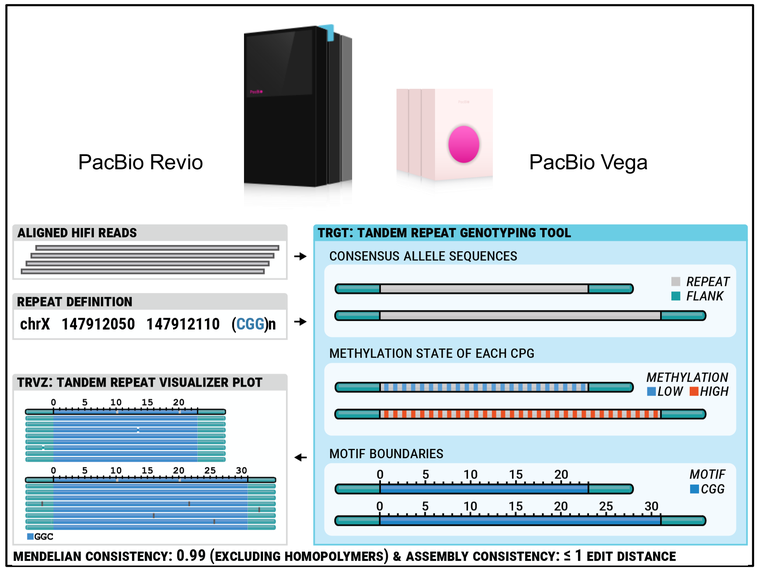


**Figure 1 Overview of TRGT.** (Top) PacBio Revio and Vega instrument. (Bottom) Workflow of genotyping repeats using TRGT. TRGT takes aligned HiFi BAM and a set of repeats definition in BED format, and genotypes each of the polymorphic repeat loci in the BAM file to provide allelic length, methylation state as well as motifs counts. In addition, a companion tool TRVZ can be used to visualize the results.

## Reducing false-positives through comparison to normal(s)

- 8 cancer cell lines with matched normal lymphoblasts were genotyped.
- We hypothesize that truly significant repeats should be expanded compared to both matching normal and unrelated normal.
- Comparing against unrelated normals reduced 65% (Range: 45% to 87%) of somatic repeats variants, resulting in an average of 2,687 somatic repeats variants (>5 supporting reads)

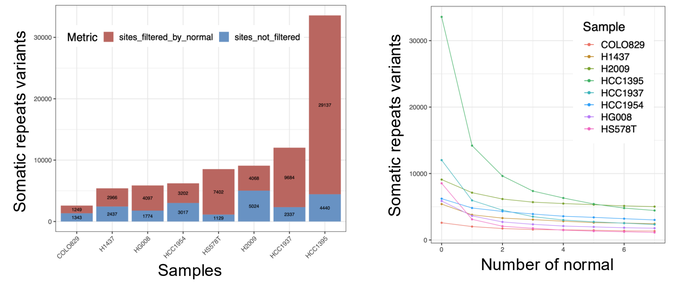


**Figure 2 Approach to reduce false-positives.** (A) Number of somatic repeats expansion before and after requiring any repeats to be expanded comparing to all other normals. Blue represents number after filtering and red represents those that are filtered. (B) Number of somatic repeats reduces with more unrelated normals used for comparison.

## TRGT identified tumor with micro-satellite instability (MSI)

- Workflow applied to 50 real cancer tumour/normal pairs.
- 2,875 somatic repeats variants identified on average. 1 sample (G41) with significantly higher number of somatic repeats variants (> 10-fold), suggesting micro-satellite instability.
- 2 pathogenic somatic variants in MSH2 identified in MSH2. Importantly, the 2 variants were phased in compound heterozygous configuration, suggesting loss-of-function in both copies of the gene. Short-reads were unable to phase the two variants ~9 kbp apart due to read-length limitation.

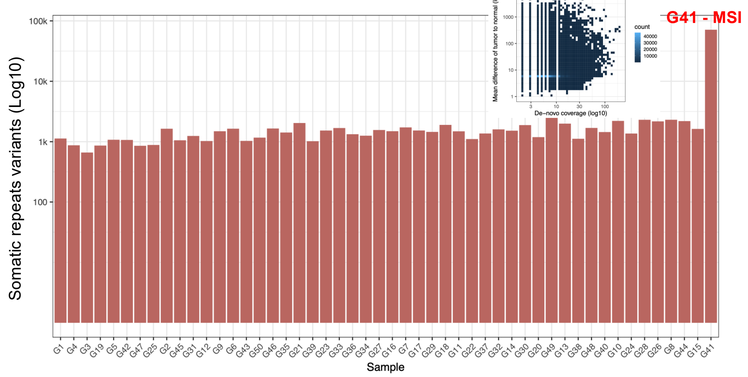


**Figure 3 TRGT effectively separates MSI sample from non-MSI samples.** Figure shows the number of somatic repeats expansion with supporting coverage of at least 5 reads. G41 can be seen to have extraordinarily high number of somatic repeats expansion (~71k) compared to all other samples. Inset figure shows the magnitude of differences in repeat size relative to supporting coverage. Most of the repeats changes are small (<10).

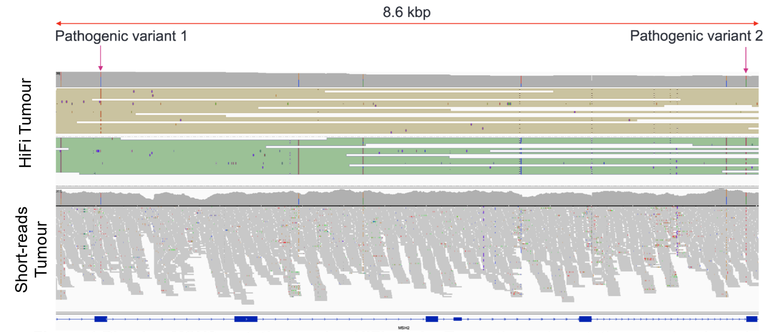


**Figure 4 Phasing MSH2 mutations using HiFi reads**. Two pathogenic mutations can be phased into compound heterozygous configuration, implying loss of function in both copies of the gene. The yellow and green colors represent the two haplotypes in the tumor. Short-read data at the bottom shows that it's not possible to phase the variants across such a long distance.

## HiFi reads identified cancer-relevant somatic expansion

- Interestingly, in the 50 cancer tumors cohort, we found 3% of the loci with somatic repeats expansion are located within genes found in the Compendium of Cancer Genes, suggesting potentially oncogenic significance.

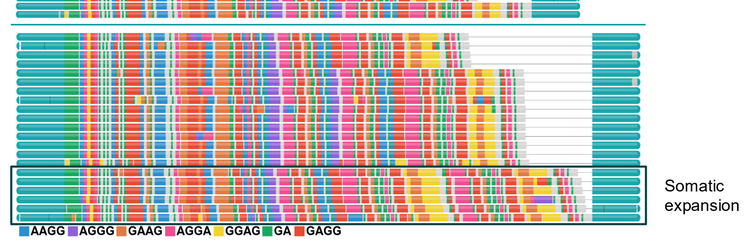


**Figure 5 Complex somatic repeat structures in OR10H1**. On top shows the repeats structure in the HiFi reads in matched normal sample. Bottom shows the tumor reads. The longest reads at the bottom box in the tumor reads pile do not exist in the matching normal.

## Conclusion

- In this work, we demonstrated the ability of HiFi reads to characterize complex repeats in cancer samples and phase somatic variants.
- MSI samples showed significantly elevated count in somatic repeats changes.
- Complex mixture of motifs can be resolved with HiFi reads.

## References

1. Weisburd, B, Dolzhenko, E. et al. Defining a tandem repeat catalog and variation clusters for genome-wide analyses and population databases. Preprint: https://doi.org/10.1101/2024.10.04.615514 (2024)
2. Dolzhenko, E. et al. Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol 42, 1606–1614 (2024).
3. Mokveld, T. et al. TRGT-denovo: accurate detection of de novo tandem repeat mutations. Preprint: https://doi.org/10.1101/2024.07.16.600745 (2024)
4. Park, J. et al. DeepSomatic: Accurate somatic small variant discovery for multiple sequencing technologies. Preprint: https://doi.org/10.1101/2024.08.16.608331 (2024)