P20.043.A

Genome-wide characterization of mosaicism in tandem repeat regions

Egor Dolzhenko[1], Adam English[2], Tom Mokveld[1], Guilherme de Sena Brandine[1], Zev Kronenberg[1], Galen Wright[3], Britt Drögemöller[3], William Rowell[1], Aaron Wenger[1], Xiao Chen[1], Mark Bennett[4], Ben Weisburd[5], Graham Erwin[2], Peng Jin[6], David Nelson[2], Harriet Dashnow[7], Fritz Sedlazeck[2], Michael Eberle[1] [1]PacBio, USA; [2]Baylor College of Medicine, USA; [3]University of Manitoba, Canada; [4]Walter and Eliza Hall Institute of Medical Research, Australia; [5]Broad Institute of MIT and Harvard, USA; [6]Emory University, USA; [7]University of Colorado Anschutz, USA

Repeat mosaicism as a disease modifier

Tandem repeat mosaicism, defined as cell-specific variation in length and sequence of the repeat, is increasingly recognized as a contributor to repeat expansion disorders.

In Huntington disease, somatic mosaicism is associated with earlier disease onset and faster progression, supporting a direct role for mosaicism in pathogenesis[1,2]. As a result, mosaicism is being explored as a candidate biomarker for clinical trials and a potential therapeutic target[3].

Measuring repeat instability in HiFi data

HiFi reads can span entire tandem repeat alleles, enabling direct measurements of their length and sequence heterogeneity. We use TRGT to define the consensus sequence of each allele and assign reads supporting it.

We compare each read with the consensus sequence by computing a divergence rate as the length-normalized edit distance between them (Figure 1). Reads from stable alleles closely match the consensus, while unstable alleles show elevated divergence due to repeat-length or sequence differences.

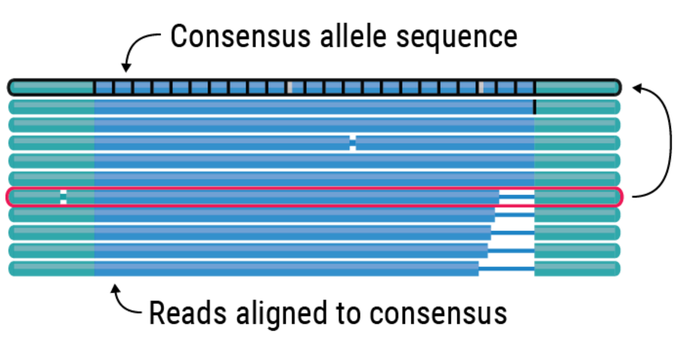

Figure 1. Measuring repeat instability via read-to-consensus divergence. Supporting HiFi reads are compared with the consensus sequence for each allele; greater deviations from the consensus indicate higher instability.

We define repeat instability rate as the mean divergence rate across all reads assigned to alleles at that repeat locus. Instability rate is similar across repeats of different length but is highly correlated with purity, defined as fraction of allele length corresponding to a perfect repeat (Figure 2).

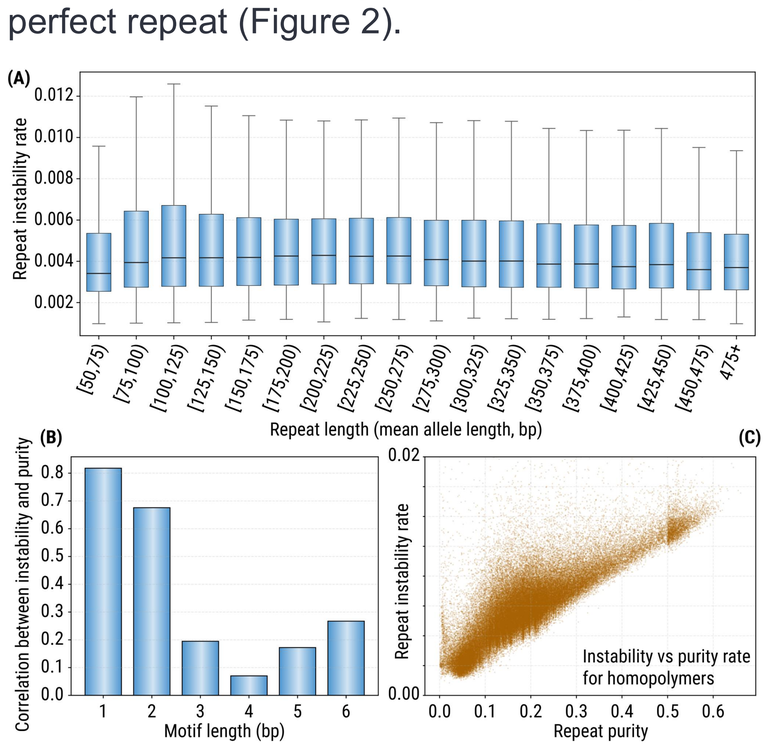

Figure 2. Properties of repeat instability. (A) Instability rate stratified by repeat length. (B) Correlation between instability and purity. (C) Instability vs. purity for homopolymers.

Modeling repeat-specific instability

Given an allele, we calculate divergence rates for all reads supporting it. We then bin these rates, producing a count vector that we call an instability profile of the allele. Stable alleles have most reads in low-divergence bins, whereas unstable alleles show more reads spread across higher-divergence bins.

To model expected instability at each tandem repeat locus, we aggregate allele instability profiles across a cohort of HiFi samples. We then fit a Dirichlet-multinomial model to these profiles. The fitted model captures both the typical instability pattern for that repeat and the amount of variability among alleles. This repeat-specific baseline can then be used to identify alleles whose instability is unusually high for that locus (Figure 3).

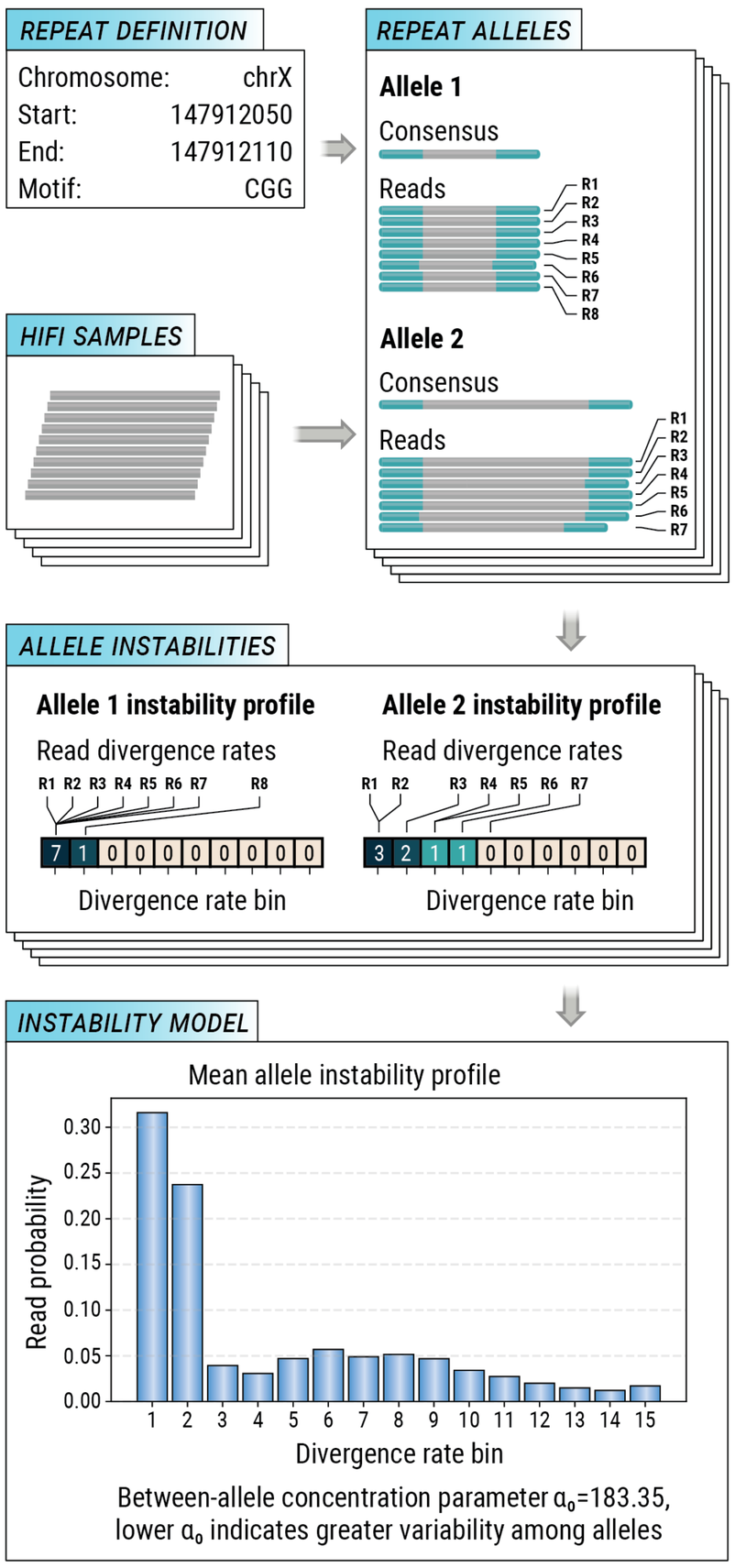

Figure 3. Modeling instability of a repeat. A repeat is genotyped across HiFi samples to obtain allele consensus sequences and supporting reads. Divergence rates are then calculated and binned into instability profiles, and a Dirichlet-multinomial model is fit to these profiles. This model describes the expected instability of the repeat locus.

Instability of pathogenic expansions

We applied our method to long-read targeted sequencing data (PureTarget) from samples with previously identified repeat expansions. We compared the instability of each allele to the baseline for that repeat locus.

In total, 27 of 881 profiled alleles of known pathogenic repeats exceeded their repeat-specific instability baseline. This included 17 of 21 pathogenic expansions, consistent with the expectation that pathogenic expansions tend to be unstable. In the remaining 10 alleles, the instability was driven by reads containing slightly longer or shorter repeats compared to the consensus sequence.

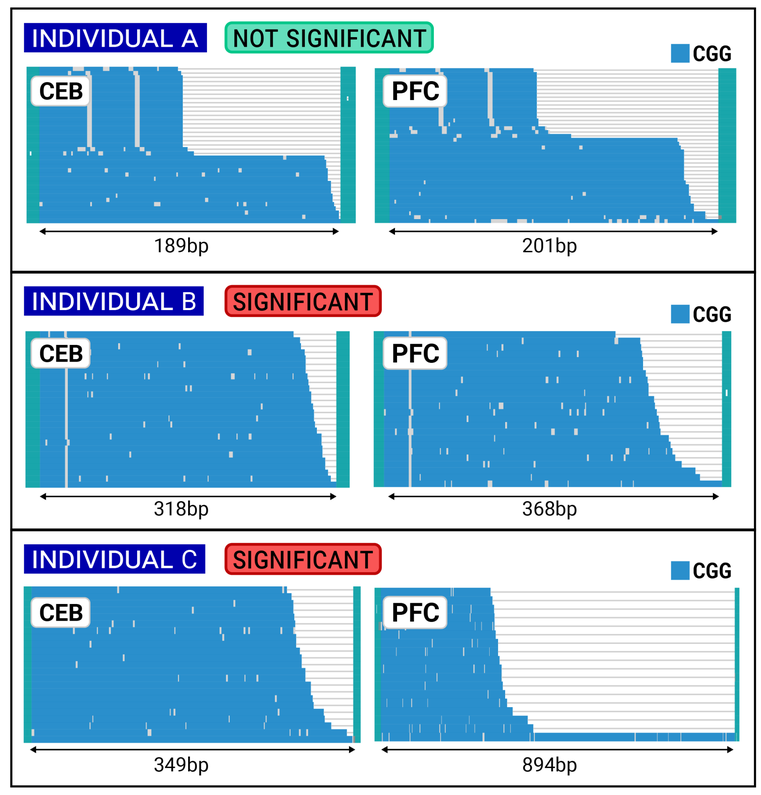

Figure 4. *FMR1* repeat instability. Matched cerebellum (CEB) and prefrontal cortex (PFC) read pileups reveal region-specific differences in instability.

Cross-tissue analysis of matched brain samples revealed two *FMR1* premutation alleles with significantly higher instability in prefrontal cortex than cerebellum (Figure 4).

Conclusions

Read-to-consensus divergence rates obtained from HiFi sequencing data provide a scalable foundation for measuring repeat instability. Repeat-specific models of baseline repeat instability allow identification of abnormally unstable alleles, enabling genome-wide studies of repeat instability and prioritization of potentially pathogenic repeats.

Availability

The method is implemented in TRGT-instability tool

- https://github.com/PacificBiosciences/trgt-instability
- https://pubmed.ncbi.nlm.nih.gov/41993463/

Conflicts of interest

ED, TM, GDSB, ZK, WR, AW, XC, and ME are employees and shareholders of PacBio. ZK is also a shareholder of Phase Genomics and Cyroterra. FS received research support from Illumina, PacBio, and ONT.

References

1. Handsaker et al., Cell, 2025; PMID: 39824182
2. Scahill et al., Nat Med. 2025; PMID: 39825149
3. Wright et al., Lancet Neurol. 2020; PMID: 33098802