Resolving variation in polymorphic regions of the human genome

Egor Dolzhenko,[1] Graham S Erwin,[2] Katherine Wang,[2] Zev Kronenberg,[1] William J Rowell,[1] Anna C Ferrari,[3] Garrison Pease,[3] Daniel Schwartz,[3] Benjamin Gartrell,[3] Ahmed Aboumohamed,[3] Alex Sankin,[3] Pedro Maria,[3] Kara Watts,[3] John M Greally,[4] Patrick Wilkinson,[5] Yashoda Rajpurohit,[5] John Loffredo,[5] Denis Smirnov,[5] Manuel A Sepulveda,[5] Charles G Drake,[5] Alex Robertson,[1] Michael P Snyder,[2] Michael A Eberle[1] | 1. PacBio, Menlo CA, USA; 2. Stanford University, CA, USA; 3. Montefiore-Einstein Cancer Center, NY, USA; 4. Einstein Epigenomics Center, NY, USA; 5. Janssen Research and Development LLC, PA, USA

Clusters of genomic variation

The human genome contains numerous repeat-rich polymorphic regions whose population structure is just starting to be systematically described. These regions, called **variation clusters**, produce large sets of inconsistent variant calls (fig 1).

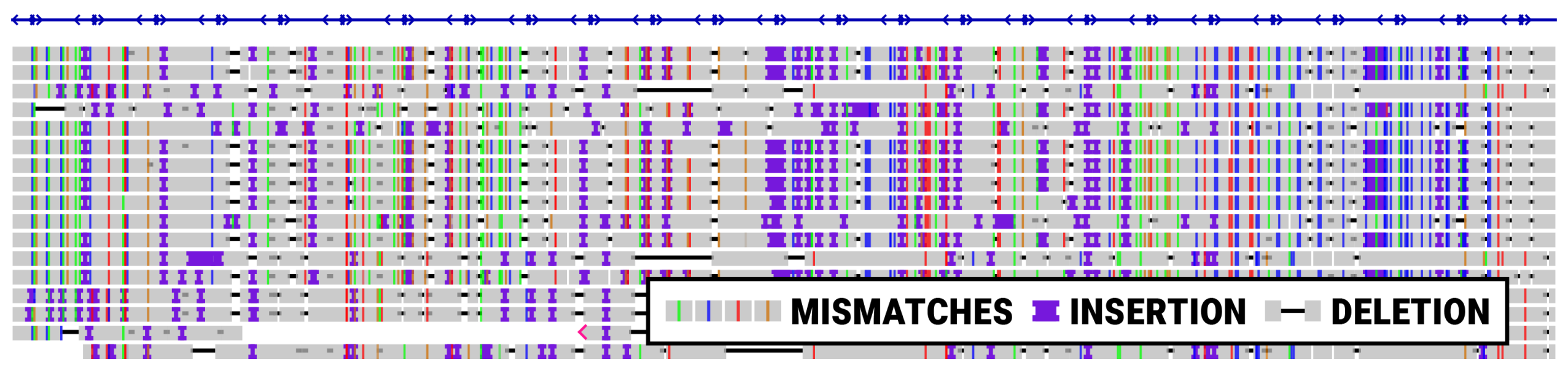

Figure 1. **A variation cluster.** Read alignments in a 3 Kbps variation cluster located in *KCNMB2* gene.

Variation clusters are usually excluded from tertiary analysis because it is difficult to interpret and catalog them.

Annotating variation clusters

We developed a method for cataloguing clusters of genomic variation in WGS HiFi sequencing data. This method works by computing variation profiles of individual HiFi samples and then finding regions where the profiles deviate from the baseline levels (fig 2).

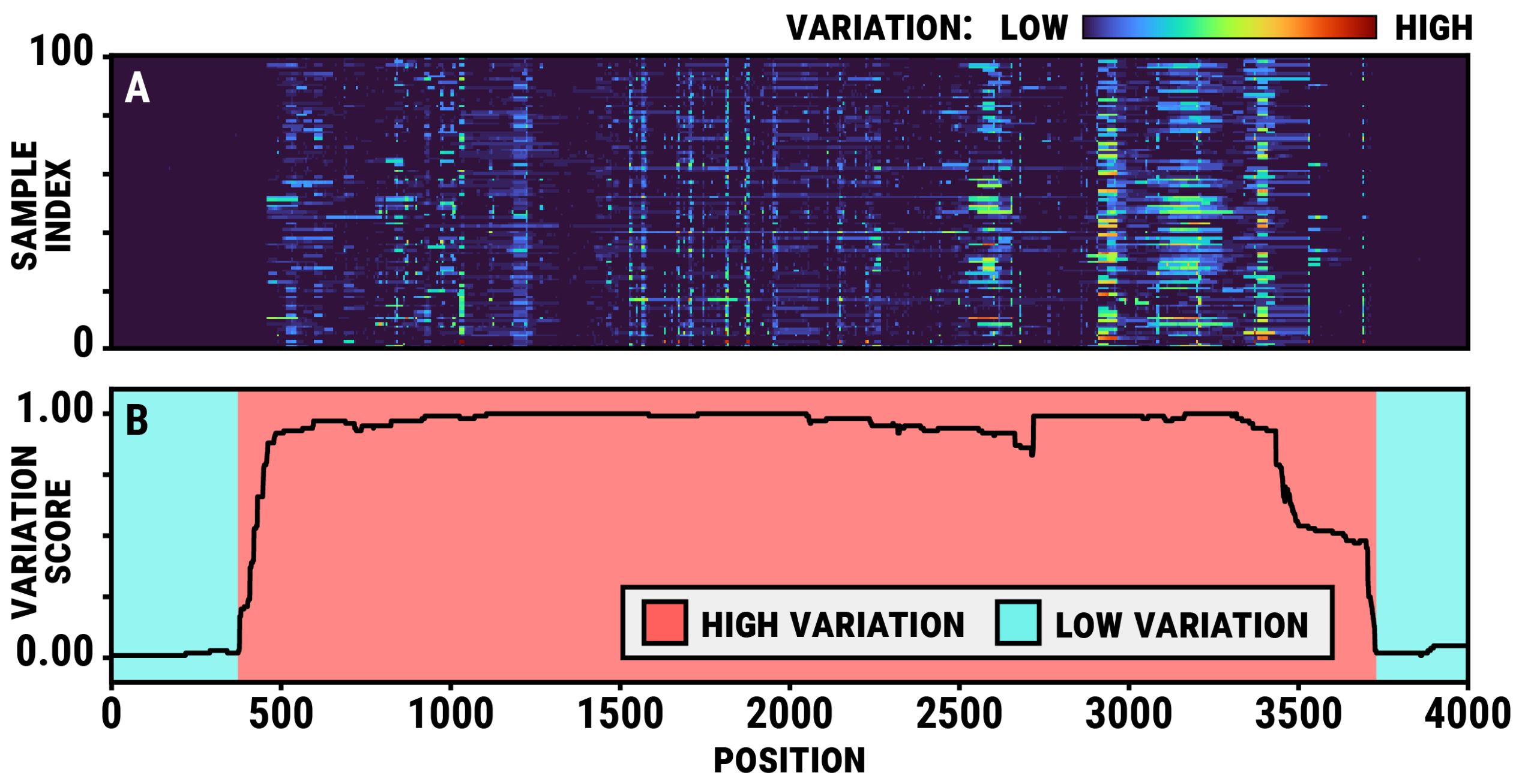

Figure 2. **Segmentation of variation profiles. (A)** Variation profiles of 100 HiFi samples in the *KCNMB2* variation cluster. **(B)** Corresponding variation scores.

Once the exact boundaries of variation clusters have been determined, they can be analyzed with TRGT, a method for genotyping low complexity, repetitive regions that we developed previously.[1]

Variation clusters across the genome

We identified 18,558 variation clusters in the GRCh38 genome across 100 HPRC[2] samples, all of which spanned over 250 bps in the reference genome. About 94% of these clusters contained tandem repeats (TRs). The number of TRs in a cluster correlated with its size (fig 3).

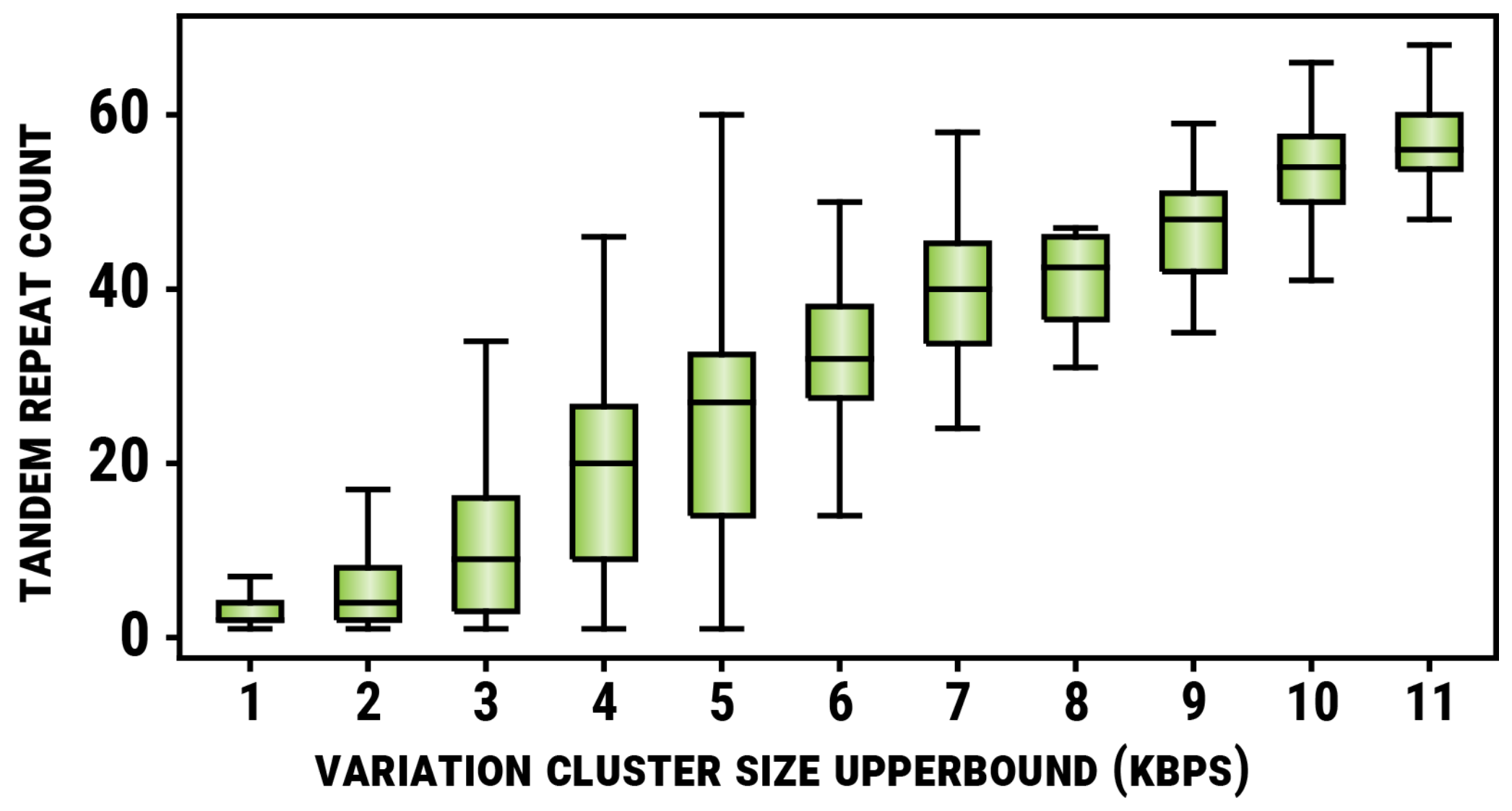

Figure 3. **Tandem repeats in variation clusters.** The number of distinct simple tandem repeats contained in a cluster stratified by the cluster length.

The variation clusters have polymorphic length across the 100 HPRC samples (fig 4), with over 76% of loci achieving the mean average deviance score (MAD) of at least 50.

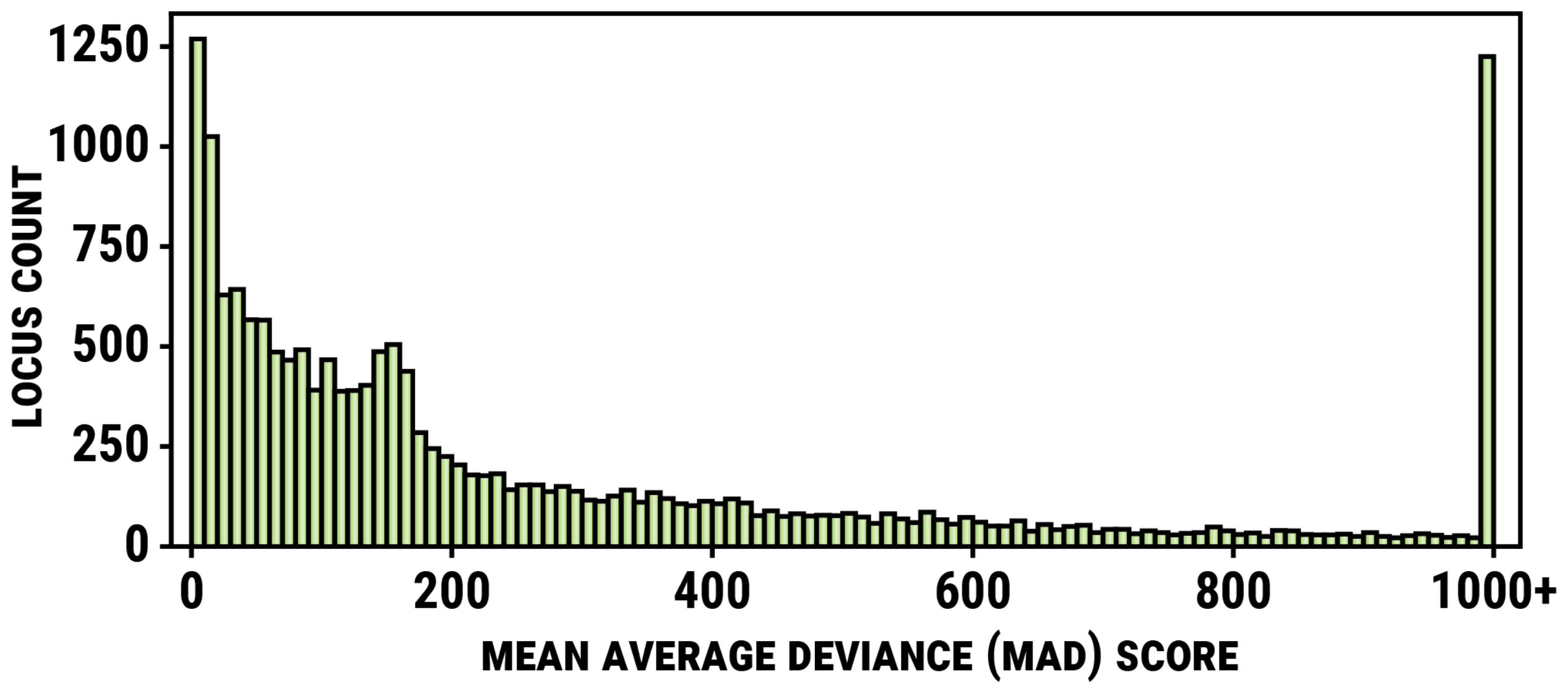

Figure 4. MAD scores. Distribution of variation cluster mean average deviance scores for HPRC samples.

One example is a variation cluster on chr14 which is composed of three motifs in various configurations (fig 5).

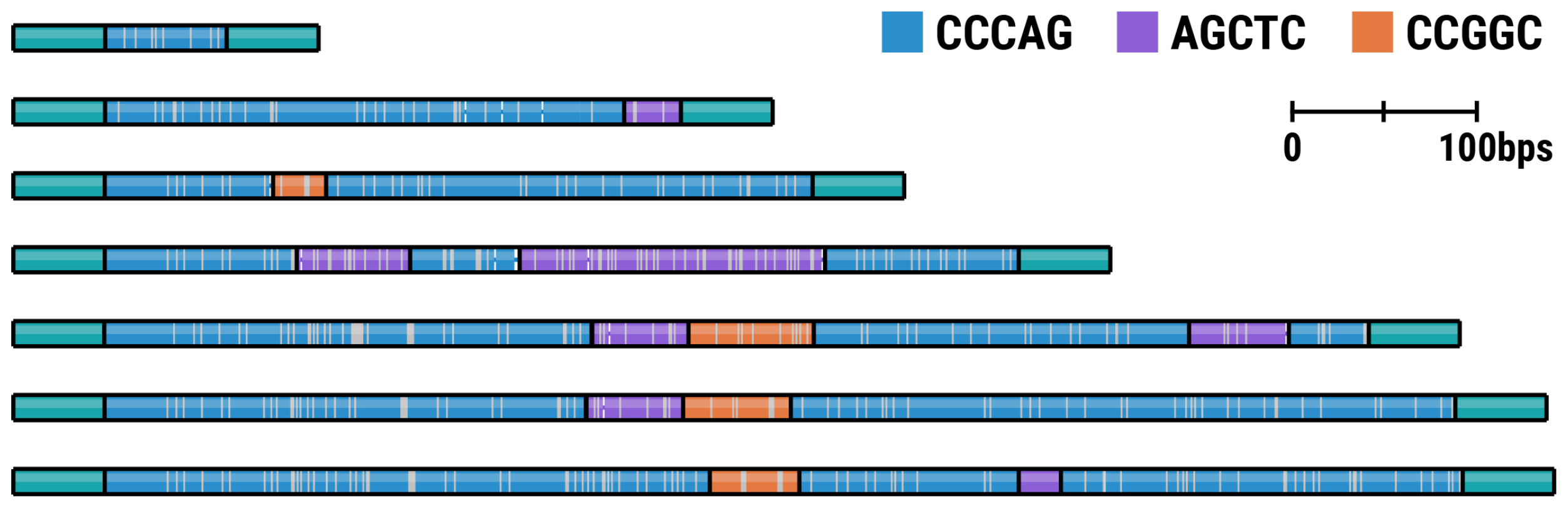

Figure 5. A polymorphic variation cluster. Alleles of a variation cluster composed of three distinct TRs.

Variation clusters in cancer

To demonstrate that our method can be used to find cancer-specific variation, we applied it to three prostate tumor-normal pairs. We identified 95 tumor-specific expansions that were also expanded relative to the 100 HPRC controls (fig 6).

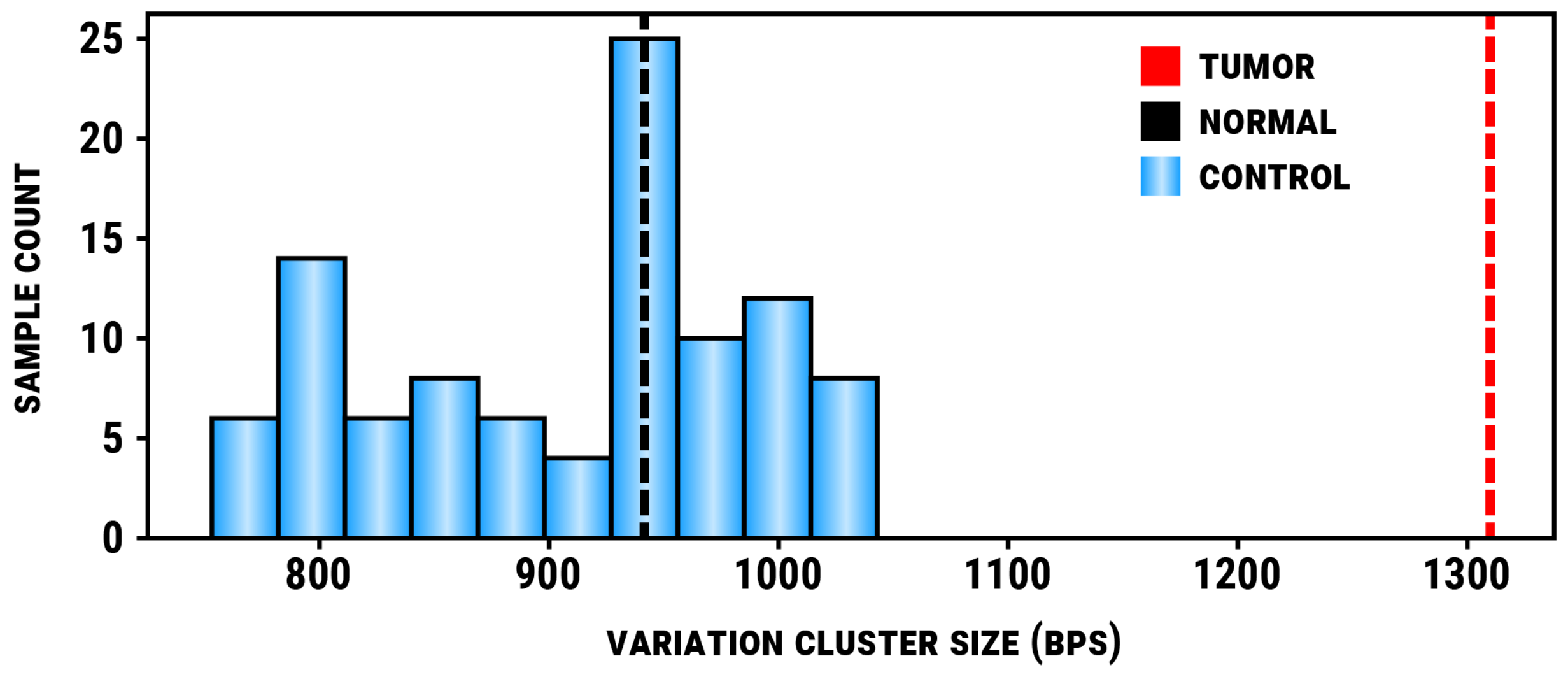

Figure 6. A tumor-specific expansion. A variation cluster containing a tumor-specific expansion that is highly polymorphic in the 100 HPRC controls.

We also confirmed that our method can identify a repeat that expands in length from normal to polyp to adenocarcinoma samples originating from the same individual (discovered by Katherine Wang and Graham Erwin) in *LIMD1*, a reported tumor suppressor gene (fig 7).

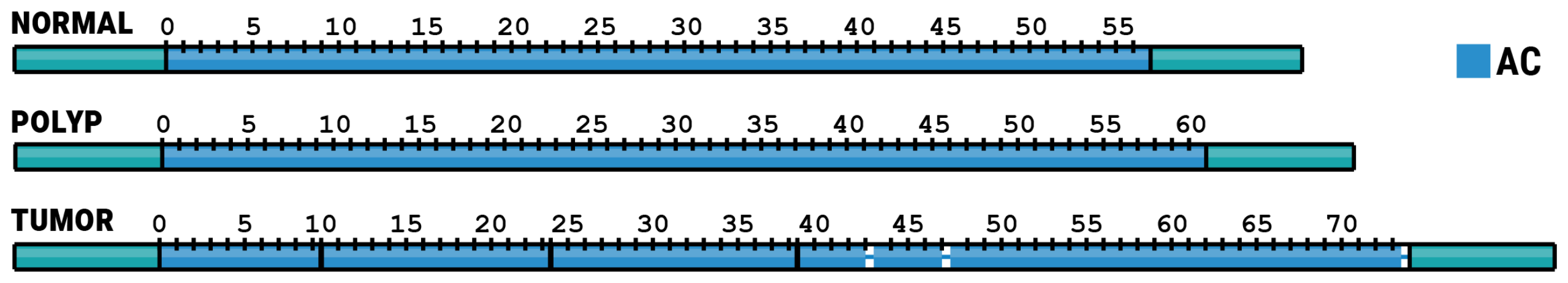

Figure 7. A progressive expansion. A progressive repeat expansion in the 5’ UTR of the *LIMD1* gene.

Conclusion

We propose a method for annotation and subsequent analysis of the entire variation clusters that overcomes typical problems associated with inconsistent variant calls inherent to these regions of the genome.

References

1. Dolzhenko E, English A, Dashnow H, et al. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol* (2024). https://doi.org/10.1038/s41587-023-02057-3
2. Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature* (2023). https://doi.org/10.1038/s41586-023-05896-x