NanoMnT: A STR analysis tool for Oxford Nanopore sequencing data driven by comprehensive analysis of error profile in STR regions.

Gyumin Park[1†], Hyunsu Ahn[1], Han Luo[*2], and Jihwan Park[*1]

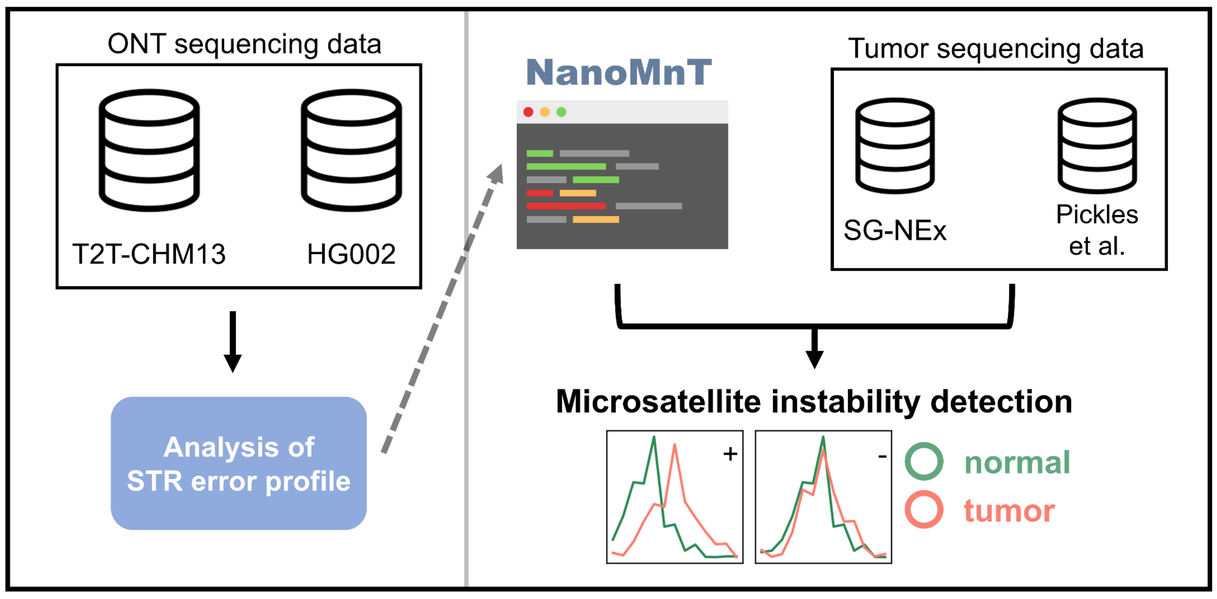
1 School of Life Sciences, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea,
2 Department of Thyroid and Parathyroid Surgery, Sichuan University, Chengdu, Sichuan, China

Functional Genomic Laboratory

ABSTRACT

Nanopore sequencing is a third-generation sequencing technology that offers cost-effective long-read sequencing. Although Nanopore sequencing offers exciting opportunities to research various areas of biology, its elevated sequencing error rate in low-complexity regions hampers its applications in short tandem repeat (STR) related research. To address this issue, we comprehensively analyzed publicly available Nanopore sequencing datasets.[1,2] We show that the sequencing error rate is not only STR length-dependent but also dependent on repeat unit and flanking sequence of STR regions. In particular, some flanking sequences were associated with good sequencing accuracy of STR, implying that certain STR loci are more viable for Nanopore sequencing compared to other loci. Moreover, while the base quality scores of substitution errors within STR regions were markedly lower than correctly sequenced bases, such discerning patterns could not be observed in indel errors. Furthermore, we show that choosing the most up to date basecaller version as well as using the SUP configuration confers significant improvements in STR sequencing accuracy. Finally, we present NanoMnT, a lightweight Python-based tool that corrects STR sequencing errors in ONT data and estimates STR allele sizes. Using NanoMnT, we present the utility of our findings by identifying MSI/MSS status in cancer sequencing data. [3] NanoMnT is available in https://github.com/18parkky/NanoMnT.

Study design / Methods



Findings

In Oxford Nanopore Sequencing (R9.4.1) ...

Indel errors (deletions in particular) are the most prominent error types in STR regions.

STR composed of certain repeat units are better sequenced than other types of STR.

Using the most up-to-date basecaller makes a night-and-day difference when it comes to STR analysis.

STR-flanking sequences profoundly influence STR sequencing accuracy.

While the R10.4.1 flowcell considerably improves STR sequencing accuracy, the overall characteristics of STR error profile is overall similar to that of R9.4.1.

MSI detection is possible from ONT data of tumor samples by strategic bioinformatic approaches that takes ONT error profile into account.

RESULT

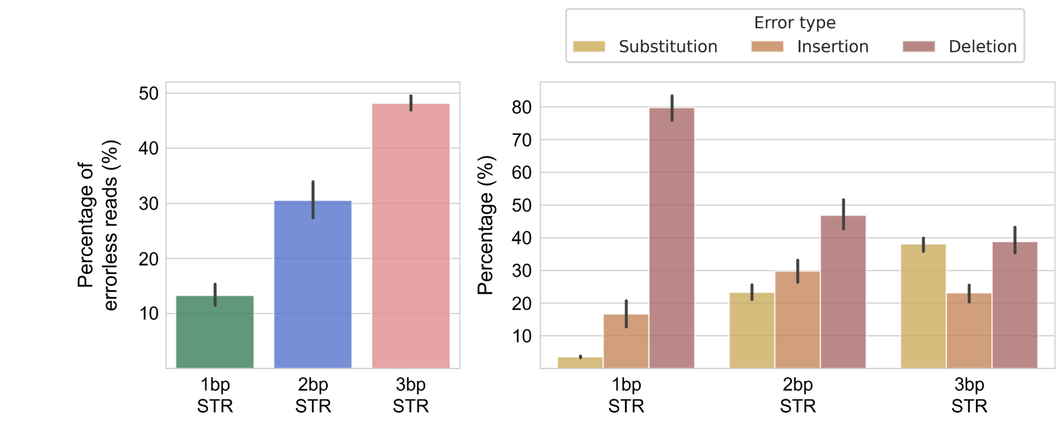


Fig 1. Distribution of sequencing error in STR regions.

Percentage of errorless reads in 1bp-, 2bp- and 3bp-repeat STR (left) and distribution of sequencing error types (right). 1bp-repeats with 10~30 repeats, 2bp-repeats with 7~24 repeats, 3bp-repeats with 5~15 repeats were subjected to analysis.

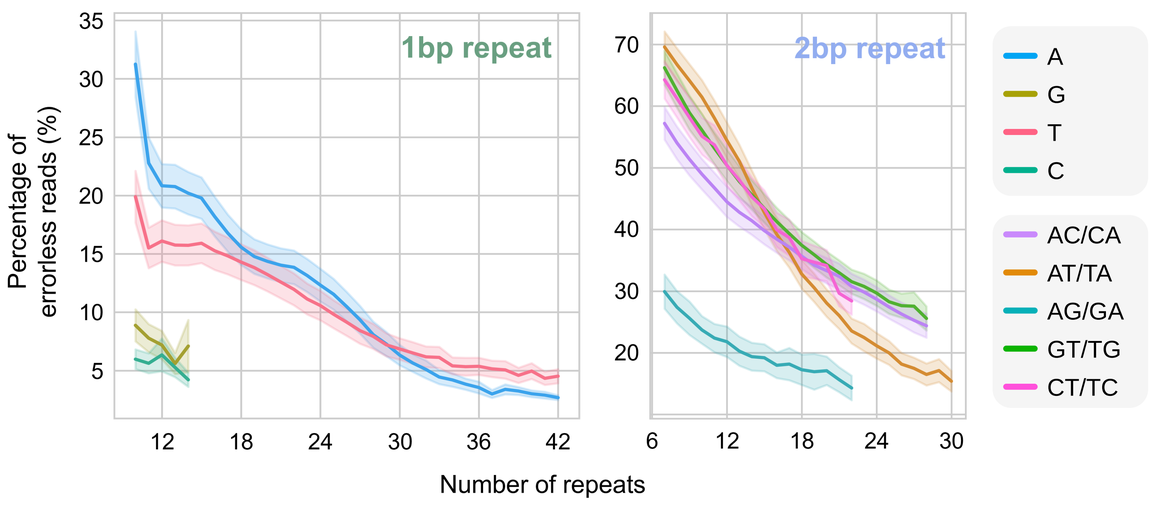


Fig 2. Sequencing accuracy of STR based on its repeat unit.

Sequencing accuracy (assessed by calculating the percentage of errorless reads) of varying lengths of 1bp-repeat STR (left) and 2bp-repeat STR (right) based on its repeat unit.

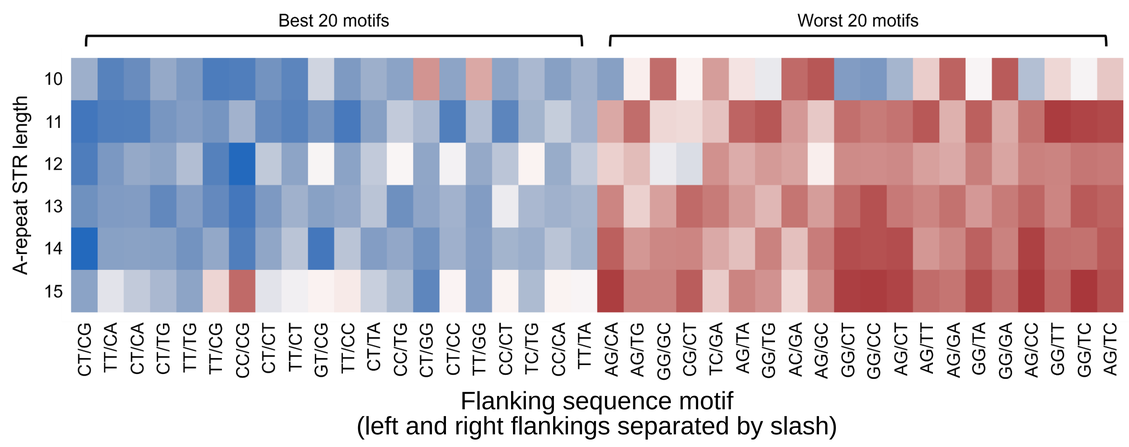


Fig 3. Identification of flankings associated with good/bad sequencing accuracy of A-repeat STR

Flanking sequences of (2 nucleotides in each direction, 4 nucleotides total) A-repeat STR associated with good and bad sequencing accuracy. 2 nucleotides in each direction are separated by slash (e.g., CT/CG motif indicates CT-(A)n-CG).

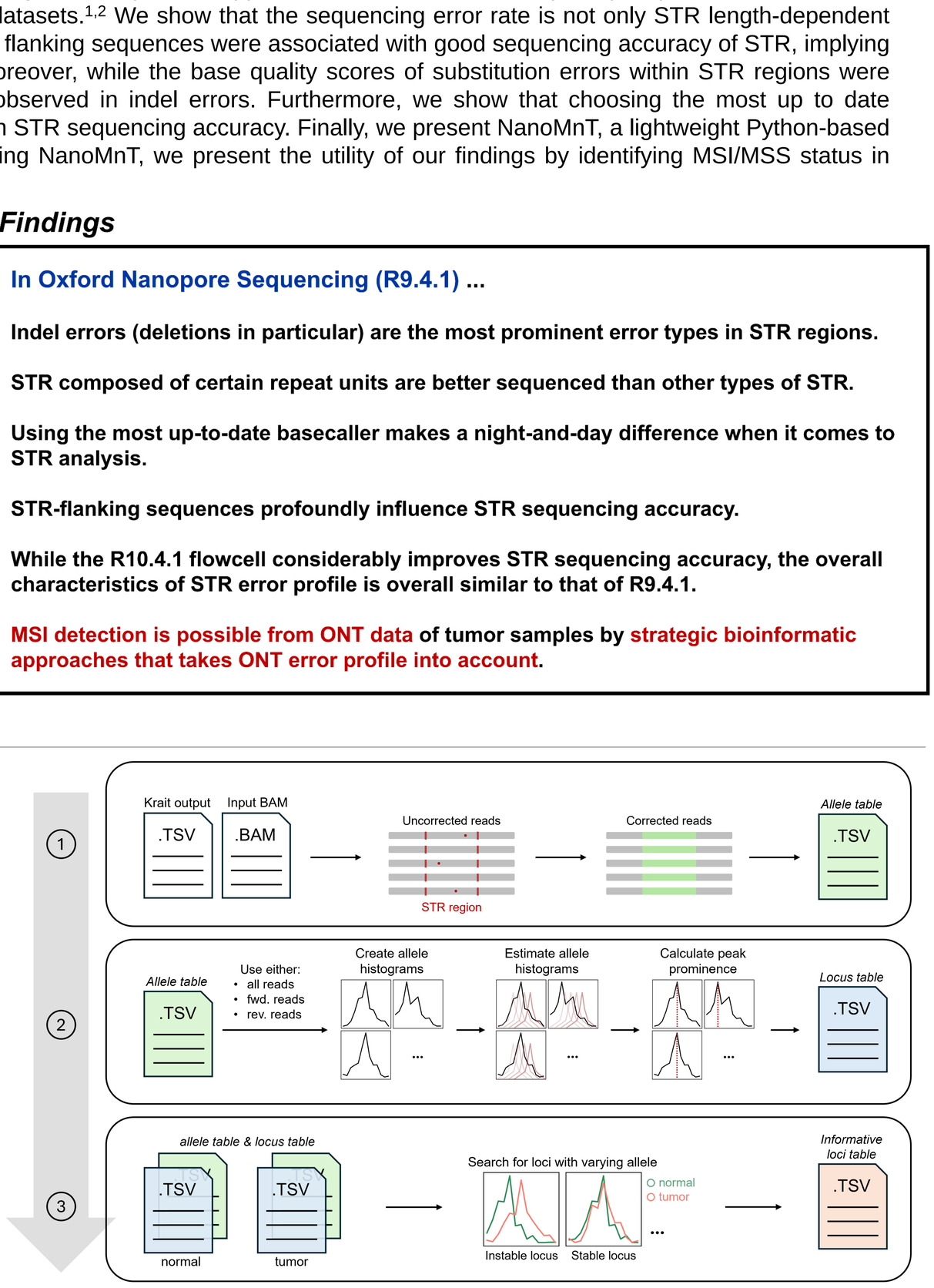


Fig 4. Functionality of NanoMnT

Schematic overview of NanoMnT functionality. First, NanoMnT performs rudimentary STR error correction for reads and generates a tab-delimited file (TSV) named Allele Table. Allele table is then used to estimate STR allele size of user-specified STR loci by comparing the observed STR allele size histogram against many synthetic STR allele size histograms, generating another TSV file, namely Locus Table. Given Allele Table and Locus Table of paired normal and tumor samples, NanoMnT compares the STR allele size histogram of commonly captured STR loci to search for loci that could provide useful information regarding the tumor's MSI status.

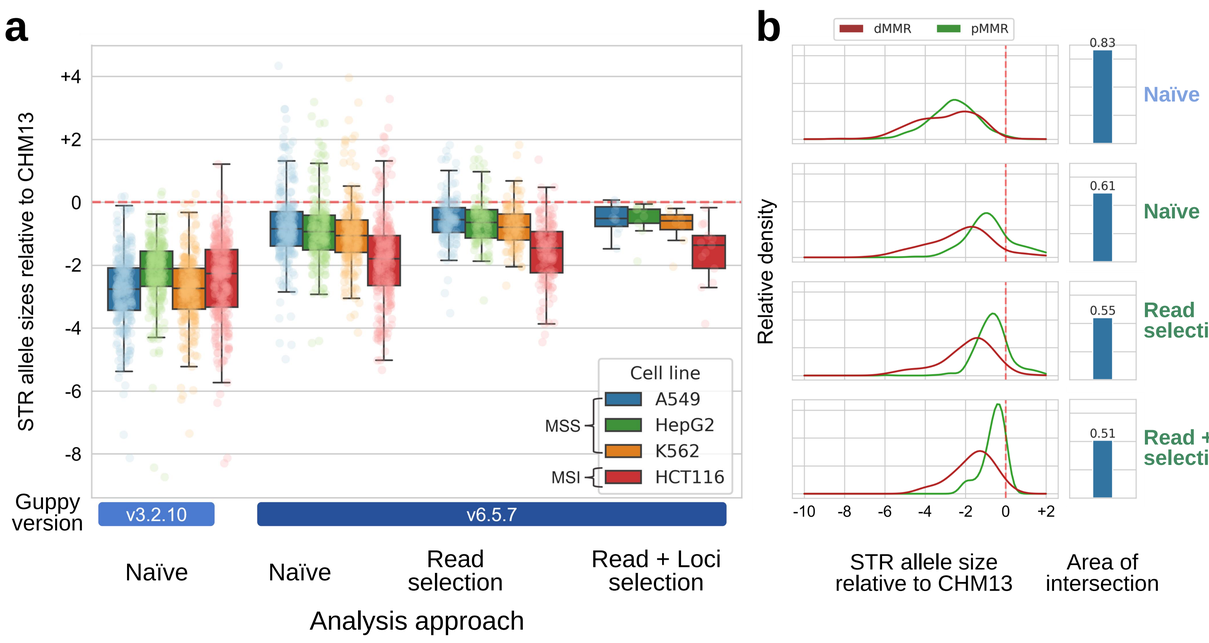


Fig 5. MSI identification results of bulk RNA sequencing data (SG-NEx)

(**a**) Distribution of STR allele size relative to CHM13 in 4 cancer cell lines, visualized by box plots and strip plots. 4 analysis approaches are compared; Naïve (Guppy v3.2.10), Naïve, read selection approach, and read + loci selection approach. (**b**) Comparison of STR allele size distribution between MSS cell lines and MSI cell line when employing the 4 different analysis approaches, visualized by kernel density estimate plots (left) and the area of intersection between the kernel density estimate plots of MSS and MSI. The red dashed vertical lines in the left figure represents the reference STR allele size (zero).

REFERENCE

1. Sergey Nurk et al. ,The complete sequence of a human genome.Science376,44-53(2022).
2. Wright, C. (2020, September 22). GM24385 Dataset Release. EPI2ME.
3. Chen, Ying, et al. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. bioRxiv (2021).

한국생명정보학회 KSBI KOREAN SOCIETY FOR BIOINFORMATICS