Genome-wide detection of somatic mosaicism at short tandem repeats

Aarushi Sehgal [1], Helyaneh Ziaei-Jam [1], Andrew Shen [1], Melissa Gymrek [1,2]*

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, USA, [2]Department of Medicine, University of California San Diego, La Jolla, USA

Abstract

Motivation: Somatic mosaicism, in which a mutation occurs post-zygotically, has been implicated in several developmental disorders, cancers, and other diseases. Short tandem repeats (STRs) consist of repeated sequences of 1-6bp and comprise more than 1 million loci in the human genome. Somatic mosaicism at STRs is known to play a key role in the pathogenicity of loci implicated in repeat expansion disorders, and is highly prevalent in cancers exhibiting microsatellite instability. While a variety of tools have been developed to genotype germline variation at STRs, a method for systematically identifying mosaic STRs (mSTRs) is lacking.

Results: We introduce prancSTR, a novel method for detecting mSTRs from individual high-throughput sequencing datasets. Unlike many existing mosaicism detection methods for other variant types, prancSTR does not require a matched control sample as input. We show that prancSTR accurately identifies mSTRs in simulated data and demonstrate its feasibility by identifying candidate mSTRs in whole genome sequencing (WGS) data derived from lymphoblastoid cell lines for individuals sequenced by the 1000 Genomes Project.

Our analysis identified an average of 76 and 577 non-homopolymer and homopolymer mSTRs per cell line as well as multiple cell lines with outlier mSTR counts more than 6 times the the population average, suggesting a subset of cell lines have particularly high STR instability rates. **Availability:** prancSTR is freely available at https://github.com/gymreklab/trtools.

Documentation: Detailed documentation is available at https://trtools.readthedocs.io/

Contact: mgymrek@ucsd.edu

Introduction

Population-level heterogeneity generally arises due to germline mutations that occur before the formation of the zygote and are inherited by all cells in the offspring. However, heterogeneity within an individual may also exist due to somatic mutations that occur post-zygotically in only a sub-population of cells (reviewed in (Youssoufian and Pyeritz, 2002)). Somatic mosaicism has long been known to play a key role in cancer (reviewed in (Stratton *et al.*, 2009)), and has also been implicated in a range of non-neoplastic disorders (e.g., Proteus Syndrome (Cohen, 1993), Neurofibromatosis Type 1 (Ruggieri and Huson, 2001) and CLOVES syndrome (Kurek *et al.*, 2012)). Somatic mosaicism is also a hallmark of conditions resulting in DNA repair deficiencies, such as Xeroderma Pigmentosum (Cleaver, 1969). Beyond its role in disease, accumulation of somatic mutations is likely a widespread phenomenon occurring in healthy individuals across the course of their lifetime Fernández *et al.* (2016).

High-throughput sequencing offers the potential to perform genome-wide detection of somatic mosaicism, but also presents important technical challenges (Dou *et al.*, 2018). To distinguish somatic mutations from germline variants or technical artifacts, a matched control sample is often required to serve as a baseline. Further, in cases where the somatic mutation is present in a small fraction of cells, ultra high coverage data is needed to detect the event (Breuss *et al.*, 2022). A variety of methods have been developed to address these challenges (e.g. MrMosaic (King *et al.*, 2017), MosaicForecast (Dou *et al.*, 2020), and DeepMosaic (Yang *et al.*, 2023)). These methods leverage allele fractions, read-based phasing, other read-level features to accurately distinguish true mosaic variants. However, existing methods in some cases still require matched control samples and focus largely on detecting mosaic single nucleotide polymorphisms (SNPs) or in some cases mosaic copy number variants (e.g. Montage (Glessner *et al.*, 2021)).

Short tandem repeats (STRs), consisting of 1-6bp sequences repeated in tandem, occur at more than 1.5 million loci in the human genome (Lander *et al.*, 2001) and exhibit rapid germline mutation rates (Sun *et al.*, 2012). Somatic instability of STRs, also known as microsatellite instability (MSI), is a hallmark of certain cancers such as Lynch Syndrome (reviewed in (Lynch *et al.*, 2009)). Recent work suggests mutation rates of approximately 10^{-4}-10^{-3} mutations per STR in non-MSI cancers, with rates more than 0.03 in the case of MSI (Fujimoto *et al.*, 2020) Additionally, somatic mutation of STRs in the brain has been implicated as a key driver of

pathogenicity in some repeat expansion disorders including Huntington's Disease (Swami *et al.*, 2009).

Detection of somatic mosaicism at STRs from sequencing data is particularly challenging, as these regions may exhibit high error rates due to PCR artifacts (Raz *et al.*, 2019) making it difficult to distinguish true somatic mutations from errors. STR-specific genotyping methods have been developed for germline genotyping that address this challenge (e.g. HipSTR (Willems *et al.*, 2017) and ExpansionHunter (Dolzhenko *et al.*, 2017)), but these are not designed to detect somatic events. Previous studies performed genome-wide analysis of somatic STR instability in the context of cancer (Hause *et al.*, 2016; Kim *et al.*, 2013; Fujimoto *et al.*, 2020), but relied on comparing sequencing from tumors with matched normal samples. Further, somatic events were detected either using custom analysis pipelines not packaged as a separate tool (Kim *et al.*, 2013) or were based on heuristics rather than hypothesis testing frameworks (Salipante *et al.*, 2014).

Here, we introduce prancSTR, a novel method for detecting mosaic STRs (mSTRs) from next-generation sequencing data without the need for a matched control sample. prancSTR models observed reads as a mixture distribution and infers the maximum likelihood mosaic fraction and the copy number of the mosaic vs. germline alleles. We show that prancSTR accurately identifies mSTRs in simulated data and validate mSTRs inferred from short reads with orthogonal long read data. Finally, we apply prancSTR to 460 whole genome sequencing (WGS) datasets from the 1000 Genomes Project derived from lymphoblastoid cell lines (LCLs) to characterize genome-wide mSTR mutations in different populations. Overall, prancSTR provides a robust method to identify mSTRs from existing high throughput sequencing datasets.

Methods

prancSTR overview

Baseline model

prancSTR is designed to identify mSTRs at one locus at a time. It takes as input STR genotypes and metadata computed by an existing genotyper and outputs candidate mSTRs (**Fig 1A**). While designed to work downstream of HipSTR (Willems *et al.*, 2017), prancSTR can theoretically process output from any STR genotyping tool as long as it returns estimated diploid repeat lengths and the observed distribution of copy numbers across all reads aligning to a locus.

At each STR locus, prancSTR takes as input a vector of the observed repeat copy number in each read, $\vec{R} = \{r_1, r_2, \ldots, r_n\}$, where r_i is the number of copies of the repeat observed in the ith read. For each locus, let $\langle A, B\rangle$ denote the diploid germline genotype, where A and B give the copy number of the repeat unit on each allele. Let f denote the fraction of chromosome copies harboring an additional allele C resulting from a mosaic mutation, and Θ represent additional error parameters described below. If the somatic mutation occurred on the haplotype containing allele B, we would expect $\frac{1}{2}$ of chromosome copies to contain allele A, $\frac{1}{2} - f$ to contain allele B, and f to contain allele C. Assuming each observed read is independent, we can then write the following likelihood equation:

$$L_B(C, f|\vec{R}; \langle A, B\rangle, \Theta) = \prod_{r\in\vec{R}} \frac{1}{2}S(r|A;\Theta) + (\frac{1}{2} - f)S(r|B;\Theta) + fS(r|C;\Theta) \quad (1)$$

where L_B denotes the likelihood of C and f in the case that the mosaic allele occurred on the haplotype with allele B. $S(r|G;\Theta)$ gives the probability to observe r copies of the repeat in a read given it originated from an allele with G copies assuming stutter error model Θ. This term is computed based on the error model used in HipSTR (Willems *et al.*, 2017):

$$S(r|G;\Theta = \{u, d, \rho\}) = \begin{cases} 1-u-d & r == G \\ u\rho(1-\rho)^{r-G-1} & r > G \\ d\rho(1-\rho)^{G-r-1} & r < G \end{cases} \quad (2)$$

where u and d give the probability for a read to contain a stutter error resulting in a repeat expansion or contraction, respectively, and error step sizes are assumed to follow a geometric distribution with parameter ρ. We assume here that u, d, and ρ are known for each locus as these can be estimated from existing data using other methods (Willems *et al.*, 2017; Kristmundsdottir *et al.*, 2020).

In practice with short reads we are unable to determine the haplotype of origin (either A or B) of the mosaic allele. Therefore below we aim to identify C and f that maximize the log likelihood over two possible cases:

$$\log L(C, f|\vec{R}) = max\{\log L_A(C, f|\vec{R}), \log L_B(C, f|\vec{R})\} \quad (3)$$

Likelihood maximization and hypothesis testing

The goal of prancSTR is to find values for C and f that maximize Equation 3. We assume the underlying stutter model Θ and diploid genotype $\langle A, B\rangle$ are known and can be obtained from HipSTR's output. We then use an iterative algorithm to estimate C and f:

1. Initialize the value of f to 0.01.
2. Compute the log-likelihood for each possible value of C, given f from step 1. We restrict our search for C to $(\min \vec{R}-3, \max \vec{R}+3)$. Return the value of C that maximizes the likelihood.
3. Find the value of f that maximizes the log-likelihood given C from step 2. This step is performed using Sequential Least Squares Programming (SLSQP) (Kraft, 1988) restricting f to be between 0 and 0.5.
4. Repeat until convergence.

In practice, the read vector $\vec{R}$ is obtained from the MALLREADS format field from HipSTR VCF files. We exclude STR calls from analysis if: they have coverage of 0, have missing genotypes, have 0 read support in MALLREADS for the called diploid genotype, or if there is only evidence in MALLREADS of reads from a single allele.

After obtaining the maximum likelihood estimates $\hat{f}$ and $\hat{C}$, prancSTR tests the null hypothesis $H_0 : f = 0$ (no mosaicism) at each STR genotyped in each sample. We compute the likelihood ratio test statistic λ_{LR}:

$$\lambda_{LR} = -2\ln\frac{L(\hat{C}, f=0|\vec{R}; \langle A, B\rangle, \Theta)}{L(\hat{C}, \hat{f}|\vec{R}; \langle A, B\rangle, \Theta)} \quad (4)$$

Finally, we use the fact that $\lambda_{LR} \sim \chi^2(2)$ to obtain a P-value testing $H_0 : f = 0$ at each STR in each sample.

Simulating vectors observed repeat counts

In our first simulation strategy, we simulated vectors of observed repeat counts for a single locus according to the baseline model described above under various parameter settings. The resulting read count vectors, as well as the known values of A, B, and Θ were used as input to prancSTR's likelihood estimation procedure. In all cases, the mosaic allele fraction f was set to one of [0.001, 0.01, 0.1, 0.2], the total number of reads N was set to one of [10, 25, 50, 75, 100, 1000], and stutter parameters were fixed at $\rho = 0.9, u = 0.02, d = 0.02$. We tested settings in which the true diploid

genotype was set either to $\langle 4, 4\rangle$ (homozygous) or $\langle 4, 6\rangle$ (heterozygous). The value of the mosaic allele C ranged from 6-10 repeats.

For each tested setting, we performed 200 simulations. Power was estimated as the percentage of simulation rounds for which prancSTR returned a significant P-value (P<0.05). Notably this captures relative power differences across settings but is not reflective of the absolute power in genome-wide analyses, in which a more stringent P-value threshold is required to account for multiple hypothesis testing. To evaluate false positive rates, we performed simulations with f set to 0 and similarly returned the percentage of simulation rounds with significant P-values.

A method for simulating error-prone next-generation sequencing reads at STRs

For our second simulation strategy, we developed a novel simulation framework, simTR, which simulates raw sequencing reads according to a specified coverage level and error model using user-defined repeat alleles. simTR is a wrapper built around ART (Huang *et al.*, 2012), an existing open source next generation sequencing read simulator. ART creates simulated reads that account for generic insertion and deletion mutations. However, stutter errors (additions or deletions of one or more repeat units introduced during PCR) characteristic of STRs are not specifically modeled. simTR adds to ART by incorporating stutter errors into the simulated reads, in addition to existing indel mutations. Stutter errors are incorporated based on the HipSTR error model described in Equation 2.

simTR takes as input a genome file (fasta format), the genomic coordinates of the target STR, and stutter parameters (u, d, and ρ). Users may also specify optional parameters to set the desired coverage, whether to generate paired-end vs. single-end reads, the mean and standard deviation of the sequencing fragment lengths, and the window size around the STR from which to simulate reads. It creates intermediate fasta files with separate entries to represent the different possible observed repeat lengths that could result from PCR stutter. It then invokes ART to simulate reads from the different fasta entries at rates proportional the expected proportion of each allele based on the input stutter parameters. Finally, it outputs simulated reads in fastq format which can be used for benchmarking downstream tools.

To evaluate the entire prancSTR pipeline starting from raw reads, we applied simTR to simulate reads at a target set of mSTRs under a range of settings. Simulated reads were aligned to a reference genome (hg38) using BWA MEM (Li, 2013) version 0.7.12-r1039. The resulting reads were used as input to HipSTR v0.6.1 for genotyping the target STRs using non-default options min-reads 5 and stutter-in to provide a file with simulated stutter error parameters. The VCF output by HipSTR was then used as input to prancSTR to estimate C and f.

We tested this pipeline on two example STR loci: (1) CSF1PO, a tetranucleotide $(ATCT)_n$ CODIS marker annotated by the National Institute of Standards and Technology (NIST) (`https://strbase-archive.nist.gov/str_CSF1PO.htm`); (hg38 chr5:150071324-150081375), and (2) a $(CGG)_n$ repeat in *CBL* (hg38 chr11:119206289-119206322). For CSF1PO the true diploid genotype was set to either $\langle 13, 11\rangle$ (heterozygous) or $\langle 13, 13\rangle$ (homozygous), and the mosaic allele was set to 15. For the *CBL* repeat the true diploid genotype was set to either $\langle 11, 14\rangle$ or $\langle 11, 11\rangle$, and the mosaic allele was set to either 9 or 10. We tested a range of values for N (10, 25, 50, 75, 100, 1000) and f (0, 0.01, 0.1, 0.2) and set stutter parameters $\rho = 0.9$, $u = d = 0.02$ to simulate 150bp paired-end reads. Example IGV (ttir *et al.*, 2013) screenshots for simulated reads at mSTRs are shown in **Supplementary Fig. 1**.

Obtaining estimated stutter parameters from real WGS datasets

We previously performed genome-wide STR genotyping using HipSTR on high-coverage PCR-free WGS for 3,202 individuals from the 1000 Genomes Project and 348 PCR+ samples from the H3Africa cohort (Jam *et al.*, 2023). Per-locus stutter parameters estimated by HipSTR were extracted from VCF files (INFO fields INFRAME_UP, INFRAME_DOWN, and INFRAME_PGEOM) for individuals from the Yoruban population (1000Genomes) and H3Africa cohorts separately using bcftools (Danecek *et al.*, 2021) v1.10.2.

Implementation

prancSTR and simTR are implemented in Python as an open source command line tool and are available as part of the TRTools (Mousavi *et al.*, 2021) package.

Validating mSTRs from NA12878 using PacBio HiFi long reads

HipSTR genotypes for NA12878 obtained previously (Jam *et al.*, 2023) were used as input to prancSTR to identify candidate mSTR sites. prancSTR output was initially filtered to include candidate mSTRs with: at least two reads supporting the identified mosaic allele C and read depth at least 10. To adjust for multiple hypothesis correction (one test per locus), we applied the Benjamini-Hochberg (Benjamini and Hochberg, 1995) method to identify mSTRs at a false discovery rate of 5%. In a second round of filtering we additionally removed mSTRs with $f > 0.3$ (indicating likely heterozygous sites), HipSTR quality score < 0.8, and mosaic support < 3 reads.

Aligned reads (BAM) for NA12878 based on PacBio HiFi long reads were obtained from Genome In A Bottle (GIAB dataset). We used the haplotag (HP) field to partition the BAM into separate files containing reads for each haplotype. We used a modified version of HipSTR (https://github.com/gymrek-lab/LongSTR) to perform targeted genotyping of candidate mSTRs identified in NA12878 using short read data. This version of HipSTR was modified to support long reads and run with non-default parameters min-reads 10, output-filters, max-str-len 10000, min-sum-qual -1e18, and skip-assembly. We extracted the MALLREADS field from the HipSTR VCF file to examine support for each allele in PacBio reads for each haplotype.

Characterizing mSTRs in the 1000 Genomes Project

We applied prancSTR to identify candidate mSTRs in 1000 Genomes Project samples based on previously obtained HipSTR calls (Jam *et al.*, 2023) and using the identical procedure and filtering as was applied to NA12878 above. These calls had already been filtered to exclude loci with call rate less than 75%, loci with genotypes not matching Hardy-Weinberg expectation (p<1e-06), and loci overlapping segmental duplications in the human genome.

Samples with outlier numbers of mSTRs were identified as those with mSTR counts more than two standard deviations above the mean across all individuals in each population. WGS sequencing coverage and EBV coverage for each sample was obtained from the 1000 Genomes Project website: `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/1000G_2504_high_coverage.sequence.index` and `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info/20130606_sample_info.txt`.

Results

Benchmarking prancSTR using simulated data

To evaluate prancSTR, we performed simulations using two strategies (**Methods**). First, to evaluate our likelihood maximization procedure, we simulated vectors of observed repeat counts in each read aligned to a locus ($\vec{R}$) according to the baseline model described in **Methods**. In this case, we assumed the germline (diploid) genotype $\langle A, B\rangle$ is known, and use the ground truth values of A and B as well as the simulated read vectors as input to the maximum likelihood estimation of mosaic allele (C) and mosaic fraction (f). Second, to evaluate our end to end pipeline starting from raw reads, we used simTR to simulate reads for mSTRs under a range of conditions, which were used as input to HipSTR to infer the germline genotype and compute read vectors. HipSTR results were used as input for mosaicism detection.

We first evaluated prancSTR under the null setting of $f = 0$ to determine how often we falsely detect a significant mSTR. P-values returned by prancSTR are well-calibrated, following the expected uniform distribution in this case (**Supplementary Fig. 2A-B**). As expected, at a P-value threshold of 0.05, prancSTR falsely identifies approximately 5% of null simulation rounds as significant mSTRs (**Supplementary Fig. 2C-D**).

Next, we simulated mSTRs under a range of values for coverage and mosaic allele fraction and for cases in which the germline genotype is either homozygous or heterozygous. Using both simulation strategies, estimated values of the mosaic allele fraction $\hat{f}$ are highly consistent with simulated values (**Fig. 1B-C**, **Supplementary Figs. 3-4**). In cases that are underpowered ($f < 0.02$ and/or coverage $10\times$), prancSTRtends to slightly but consistently overestimate the mosaic allele fraction. In practice, these cases are unlikely to reach genome-wide significance.

As expected, power to detect mSTRs increases as a function of f and sequencing coverage in all simulation settings (**Fig. 1D-E, Supplementary Figs. 3-4**) with near perfect power at P<0.05 to detect mSTRs with $f > 0.1$ at loci with at least $50\times$ coverage. In both simulation strategies, power is higher when the germline genotype is heterozygous vs. homozygous. This difference is more pronounced in results based on simTR simulations. In that case, this bias is partially explained by genotyping errors. We observed that cases where the simulated germline genotype is homozygous but the mosaic fraction is high are consistently misidentified by HipSTR as heterozygous sites, and therefore cannot be identified by prancSTR as mSTRs. We additionally evaluated the impact of the mosaic allele size on power. We observed that power increases with the absolute difference in length of the mosaic allele compared to the nearest germline allele (**Supplementary Fig. 6**). This is expected, since larger differences in size make it easier to distinguish true mosaic alleles from errors.

Finally, we evaluated the impact of sequencing errors at STRs on the ability to detect mSTRs from simulated read vectors under varying stutter model parameters meant to capture typical error rates in PCR+ ($\sim$10% of reads) vs. PCR-free ($\sim$1% of reads) data (**Supplementary Figs. 6-8**). As expected, with high stutter error rate, power is reduced in cases of low coverage and low mosaic fraction, and estimates of C and f show greater variability. This suggests mSTR detection will perform poorly on PCR+ short read data, where stutter error rates may often exceed expected mosaic fractions.

Detecting mSTRs in a deeply sequenced human sample

We applied prancSTRto detect genome-wide mSTRs from high-coverage PCR-free short read whole genome sequencing (WGS) from the highly characterized NA12878 sample. WGS was derived from a lymphoblastoid cell line (LCL), and therefore identified mSTRs likely consist of a combination of true somatic mutations that existed before sample collection as well as mutations that have accumulated during cell line passages. After applying prancSTRand performing minimal filtering to remove low quality calls (**Methods**), we identified 1,219 candidate autosomal mSTRs (adjusted P<0.05). Of candidate mSTRs identified above, 1,130 (92%) occurred at homopolymer loci.

To evaluate these mSTRs, we compared to an orthogonal dataset of haplotagged Pacbio HiFi long reads (mean coverage $\sim$30$\times$) available for the same sample (**Methods**). Notably, although Pacbio HiFi shows high accuracy at most regions, they have elevated error rates at homopolymers (Wenger *et al.*, 2019), suggesting repeat counts obtained from Pacbio reads at those loci may not serve as an accurate ground truth dataset. Additionally, we noticed that inferred stutter error rates are highest at homopolymer STRs (**Supplementary Fig. 7**). Therefore, results below are reported separately for non-homopolymer vs. homopolymer STRs.

We reasoned that true mosaic alleles with sufficiently high variant allele fractions should be observed in both datasets, and that the mosaic allele should typically only occur on long reads from one of the two haplotypes at a locus (**Fig. 2A**). On the other hand, inferred mosaic alleles that are actually due to stutter or other error sources might be found on both haplotypes. Of the mSTRs identified above, we deemed 40 (273) corresponding to 45% (24%) of candidate non-homopolymer (homopolymer) mSTRs to have sufficient Pacbio HiFi coverage (at least 10 reads per haplotype) to attempt validation.

For each candidate mSTR, we examined the percentage of long reads from each haplotype supporting the inferred mosaic allele (C) (**Fig. 2B-C**) and classified calls into three categories. Category I, corresponding to 67% (31%) of non-homopolymers (homopolymers), consists of calls for which C is only identified in HiFi reads from a single haplotype, representing likely true positives. For these mSTRs, variant allele fractions estimated from short reads are strongly correlated with those observed in the HiFi reads (Pearson r=0.76, two-sided P=0.001 for non-homopolymers and r=0.55, P=5.11e-6 for homopolymers; **Fig. 2D-E**). Category II, corresponding to 7% (59%) of non-homopolymers (homopolymers), consists of calls for which C is supported by at least one HiFi read from each haplotype, representing likely false positive calls. Category III, corresponding to 26% (10%) of non-homopolymers (homopolymers), consists of calls for which C is not supported by long reads on either haplotype. This could indicate an incorrect mSTR call, but could also originate from insufficient coverage at mSTRs with low variant allele fractions. Upon further inspection of mSTRs in Categories II and III, we determined the majority had either low mosaic read support in the short read data, occurred at loci with low genotype quality, or had very high mosaic allele support suggesting miscalled heterozygous sites. After applying an additional round of filtering based on these metrics (**Methods**), 18 (117) non-homopolymer (homopolymer) mSTRs remained of which 83% (49%) were classified as Category I.

We further examined read support on each haplotype at remaining candidate mSTRs (**Supplementary Fig. 9**). This revealed that the majority of homopolymer mSTRs identified as likely false positives occurred at loci for which the germline genotype was called as homozygous and long reads from both haplotypes supported multiple different alleles, suggesting reads at these loci are error prone. We additionally observed across all loci that the majority of validated high-confidence mSTRs occur at STRs for which the germline genotype is heterozygous. This is consistent with our simulation results, in which true mosaic alleles with high mosaic allele fraction occurring at homozygous sites are incorrectly genotyped as heterozygous and therefore systematically missing from our mSTR callset. On the other hand, those with low allele fraction are unlikely to be detected at genome-wide significance.

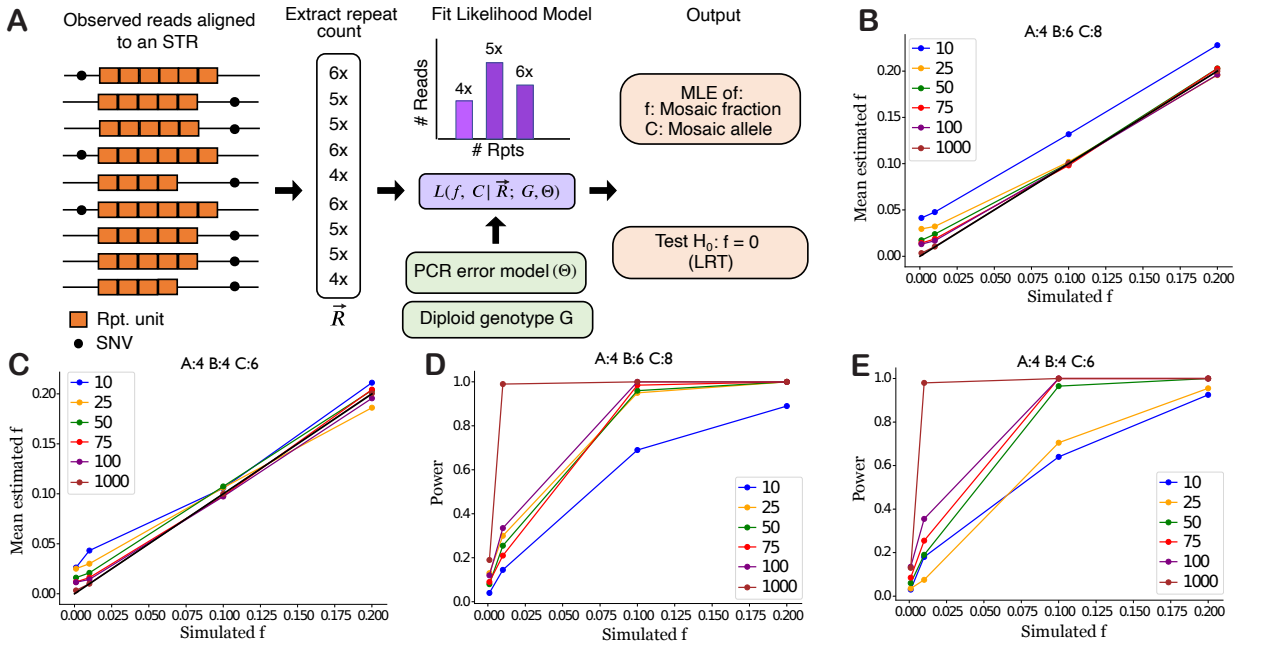

Fig. 1. prancSTR overview and validation. (A) Overview of the prancSTR method. The copy numbers observed in each read aligned to a target STR are extracted to a vector $\vec{R}$, from which prancSTR obtains maximum likelihood estimates for the mosaic allele (C) and mosaic allele fraction (f), and a P-value testing $H_0 : f = 0$. (B-C) Simulated vs. estimated values of f. We simulated mSTRs under a range of coverage levels and values for f for cases in which the germline genotype is heterozygous (B) or homozygous (C). Dots represent the mean estimated f value from 200 simulations. The black line denotes the x=y diagonal. (D-E) Power to detect mSTRs. Power is computed as the percent of simulations for which $P < 0.05$. For B-E, lines denote different coverage levels, where coverage gives the total number of reads spanning the STR of interest. Simulated values for A, B, and C are denoted at the top of each panel. Panels here are based on simulated read vectors $\vec{R}$. Similar results for simulations based on raw reads are shown in Supplementary Figs. 3-4.

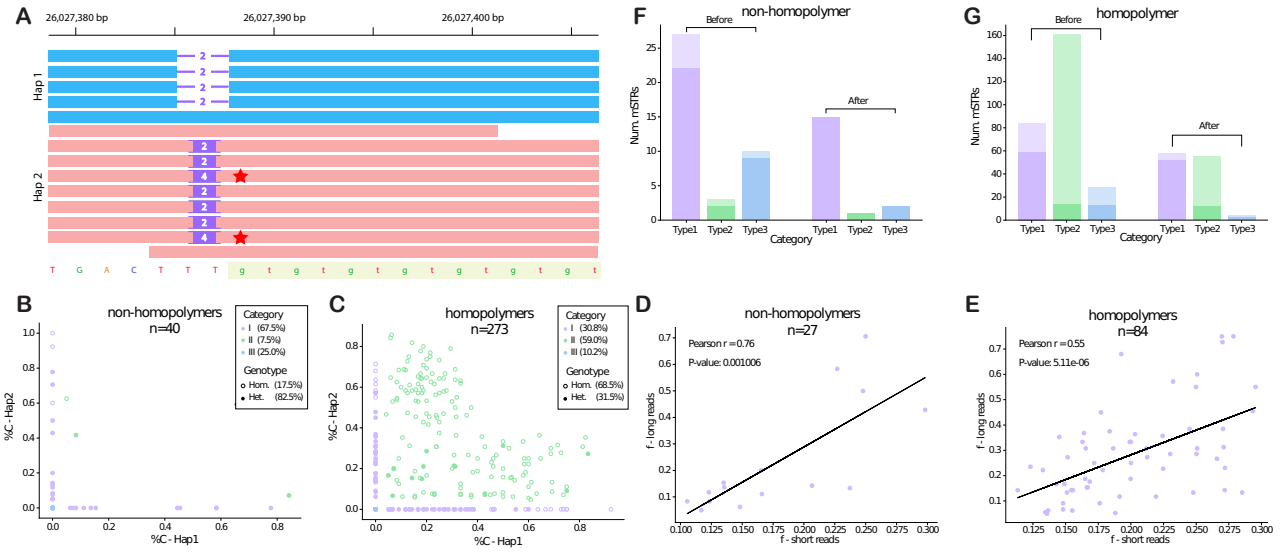

Fig. 2. Validating candidate mSTRs identified in NA12878 (HG001) with Pacbio HiFi reads. (A) Schematic representation of mosaicism validation with long reads. Pacbio HiFi reads are haplotagged as belonging to either of the two haplotypes in a sample. The highlighted region denotes the STR region. Bars indicate deletions and purple rectangles indicate insertions compared to the reference genome. Mosaic alleles (red star) are typically expected to occur on only one of the two haplotypes. (B-C) Mosaic allele support on long reads from each haplotype. The x-axis and y-axis show the percentage of reads on each haplotype matching the mosaic allele for non-homopolymer (B) and homopolymer (C) loci. mSTRs were classified as Category I (purple; mosaic support on a single haplotype, likely true positives), Category II (green; mosaic support on both haplotypes, likely false positives), and Category III (blue; no mosaic support, undetermined). (D-E) Number of mSTRs before and after filtering. Bars show the number of mSTRs for non-homopolymers (D) and homopolymers (E) in each category before and after filtering on mosaic read support, mosaic allele fraction, and genotype quality. (F-G) Correlation of mosaic fraction between short and long reads. Comparisons of estimated allele fractions (for the mosaic allele identified in short reads) from short reads (x-axis) vs. Pacbio reads (y-axis) are shown for non-homopolymer (F) and homopolymer (G) mSTRs. Black lines denote the best fit line.

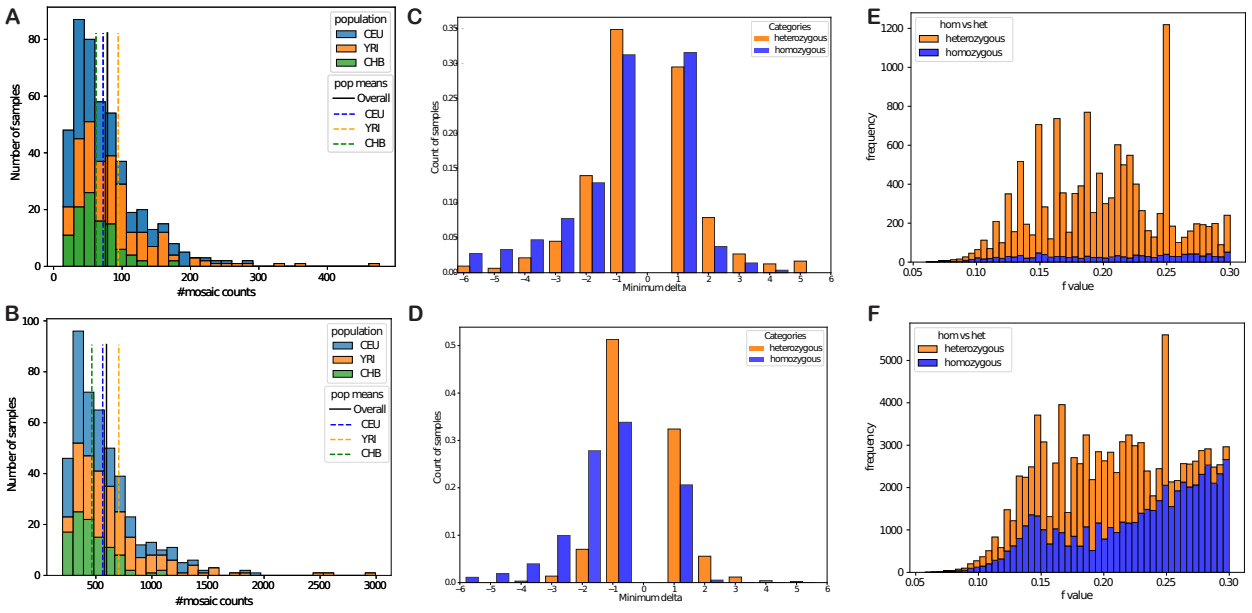

Fig. 3. mSTR trends across populations. (A-B) Distribution of the number of mSTRs across different populations. The x-axis gives the number of mSTRs for a given population and the y-axis gives the count. Data is shown for non-homopolymers (A) and homopolymers (B). Dashed colored lines (CEU=blue, YRI=orange, CHB=green) give population-specific means and the black line denotes the overall mean. (C-D) Distribution of mSTR mutation sizes. The x-axis represents the mutation size, computed as the difference between the mosaic allele length and the closest germline allele. Positive mutation sizes indicate insertions and negative sizes indicate deletions. Data is shown for non-homopolymers (C) and homopolymers (D) and is for CEU only. Other populations showed similar trends. Blue represents homozygous loci and orange represents the heterozygous loci. (E-F) Distribution of mosaic allele fraction (f) across mSTRs. Data is shown for non-homopolymers (E) and homopolymers (F) and is for CEU only. Bars are colored to denote the number of mSTRs occurring at homozygous (blue) vs. heterozygous (orange) sites.

Population-wide characterization of mSTRs

We next applied prancSTR to characterize population-wide trends of STR mosaicism. We focused on WGS derived from LCLs for individuals from the CEU (Northern Europeans from Utah; n=179), YRI (Yorubans from Nigeria; n=178), and CHB (Han Chinese; n=103) populations. After filtering (**Methods**), we identified an average of 76 (577) non-homopolymer (homopolymer) mSTRs per cell line (**Fig. 3A-B**). As observed for NA12878, homopolymer mSTRs far outnumber non-homopolymers, and the majority of non-homopolymer mSTRs identified occur at loci for which the germline genotype is heterozygous (**Supplementary Fig. 10**). This trend is consistent across all populations analyzed.

We noticed substantial variation in mSTR counts across cell lines. The number of homopolymer and non-homopolymer mSTRs per cell line are highly correlated (**Supplementary Fig. 11**), with the correlation strongest when considering mSTR calls at germline heterozygous sites (Pearson r=0.97, two-sided P=9.6e-110 in CEU). We also identified 13, 8, and 4 cell lines from CEU, YRI, and CHB with outlier mSTR counts (**Methods**) for both homopolymer and non-homopolymer mSTRs. Overall, these results suggest certain cell lines have higher rates of STR instability, either due to genetic or environmental factors. Variation in mSTR counts across cell lines is not significantly correlated with the number of sites considered or EBV virus count (two-sided P$\geq$ 0.05), and is only modestly correlated with sequencing coverage (Pearson r=0.21, two-sided P=0.039 for non-homopolymers and r=0.17, P=0.10) (**Supplementary Fig. 12**). Passage numbers for these cell lines was not available at the time of writing, and so the impact of cell culture history, which is likely to play a role in mutation counts, could not be assessed.

We next investigated the distribution of the sizes of mosaic STR mutations. The majority of events (60.2% and 68.7% for non-homopolymer and homopolymer mSTRs) result in insertions or deletions of a single repeat unit (**Fig. 3C-D**), although larger step sizes were observed. Mutation sizes are larger on average for mutations at STRs with homozygous vs. heterozygous germline genotypes and show an overall bias toward deletions vs. contractions. A similar deletion bias has been observed for somatic mutations at STRs in cancer (Fujimoto *et al.*, 2020). However, both bias described above are far more pronounced at homopolymer loci, suggesting these biases my arise in part from erroneous mSTR calls (**Supplementary Fig. 13**). Indeed, inferred stutter error rates suggest deletion errors are more common than insertions (**Supplementary Figs. 7-8**), and large mutation step sizes at homozygous sites may reflect true heterozygous sites that were incorrectly genotyped.

Finally, we examined the distribution of variant allele fractions (f) for detected mSTRs (**Fig. 3E-F, Supplementary Fig. 14**). In all cases, f distributions show peaks around 0.15-0.20, consistent with the range where we expect to have sufficient power (**Fig. 1D-E**), whereas true mosaic sites with higher f values are likely to be indistinguishable from heterozygous sites. Further, homopolymer mSTRs with high f values nearly all occur at homozygous sites, and the observed deletion bias is strongest overall for sites with high f values, indicating mSTRs with $f > 0.2$ may be enriched for false positive calls. Overall, in combination with long read validation analysis performed above, our results suggest mSTRs identified at heterozygous sites at moderate f values are robust, whereas accurate identification of mosaicism at homopolymers or for sites with high mosaic allele fractions is challenging with short read data.

Discussion

Here we presented prancSTR, a method for genome-wide detection of somatic mosaicism at STRs from high throughput sequencing datasets. prancSTR can accurately identify mSTRs without the need for a matched control sample. It has highest power to detect mSTRs with mosaic allele fractions of approximately 10-20% in PCR-free datasets with 30-50× coverage, but could detect reproducible mSTR sites with mosaic allele fractions as low as 7%. We applied prancSTR to identify mSTRs using PCR-free short read data for NA12878. Validation with orthogonal long read (Pacbio Hifi) data supported 83% and 49% of high-confidence mSTR calls at non-homopolymers and homopolymers, respectively, at sites with sufficient long read coverage. Application of prancSTR to population-scale short read WGS for the 1000 Genomes derived from lymphoblastoid cell lines identified hundreds of mSTRs per cell line with broadly consistent mSTR patterns across populations.

prancSTR is a versatile tool that can be used to detect mosaicism in a variety of settings, including PCR-free or PCR+ short read sequencing, as long as accurate stutter error parameters are available. It can also be applied as-is to Pacbio Hifi datasets, which are becoming increasingly widely available. prancSTR as well as the read simulation method developed here (simTR) have been packaged into our existing toolkit, TRTools (Mousavi *et al.*, 2021), enabling easy integration with other TR analysis tools. It is currently compatible with STR genotypes output by HipSTR (Willems *et al.*, 2017), but could be easily modified to work downstream of other STR genotypers provided they output diploid genotypes and read support for each observed allele.

Application of prancSTR genome-wide to WGS from 460 cell lines revealed interesting patterns of mSTRs. Our results broadly suggest mSTRs identified from short reads at non-homopolymers and at sites with germline heterozygous genotypes are most reliable, whereas homopolymers remain particularly challenging. Overall, we found an average of 76 and 577 non-homopolymer and homopolymer mSTRs per cell line, corresponding to mutation rates of XX and XX mutations per STR per sample. These rates are broadly similar to those measured in non-MSI tumors (Fujimoto *et al.*, 2020). Intriguingly, we identified multiple cell lines from each population with outlier mutation counts, and found strong correlation between the number of mSTRs at homopolymers vs. non-homopolymers. This suggests cell lines have higher rates of STR instability than others, and that these trends are present across a broad set of loci. Although we could not determine the source of this variation, but hypothesize it could be due to differences in cell line passage history, in which cell lines that have been maintained for more passages accumulate more somatic variation. Alternatively, individual-level variation in mutation patterns could arise from germline factors such as mutations in DNA repair genes, which we have observed previously in mice (Maksimov *et al.*, 2023). Profiling somatic variation in larger sample sizes is likely needed to identify similar effects in humans.

prancSTR currently faces multiple limitations. First, it relies on an upstream genotyper (here, HipSTR) to provide accurate germline genotype calls as input. We identified several scenarios where germline genotype calls may be problematic. First, in cases where a mosaic allele is present at high frequency, it may be indistinguishable from a germline allele and incorrectly genotyped as heterozygous, causing mosaicism to be missed. Second, particularly at loci with high stutter error rates or low coverage, a truly heterozygous site may be incorrectly genotyped as homozygous, causing prancSTR to incorrectly identify the second germline allele as mosaicism. As a result, mSTRs identified at heterozygous sites are likely more reliable. We anticipate these challenges will be largely alleviated by haplotagged long reads, which will make distinguishing heterozygous vs. homozygous sites easier. Second, prancSTR currently focuses on identifying mSTRs with a single high frequency mosaic allele. While this is likely to capture mosaic events at shorter STRs, longer repeats such as the Huntington's Disease locus where mosaicism is known to play a role in disease pathogenesis tend to show a broad range of mosaic allele lengths (Swami *et al.*, 2009) and will require extensions to the current model to detect. Third, similar to mosaicism detection tools for other variant types, prancSTR is limited by the coverage of current datasets, which is insufficient to detect most mosaic events below 5% frequency.

Overall, prancSTR can serve as a valuable method to characterize somatic mosaicism at STRs in a range of settings, including in healthy individuals or in disease settings such as microsatellite instability in cancer or neurological diseases where mosaicism is known to play a key role. Profiling mosaicism at population-scale from the large number of existing WGS datasets may also give insight into inherited factors driving differences in mSTRs patterns across individuals. We envision future extensions of this framework can allow for directly incorporating phase information from haplotagged reads or quantifying mosaicism at highly unstable repeats such as long Huntington's alleles, which will further improve our ability to characterize STR mosaicism and its role in human health.

Acknowledgements

We thank Alon Goren, Vineet Bafna, and Jonathan Margoliash for helpful discussions about the manuscript.

Funding

This work was supported in part by the National Institutes of Health [Grant No. 1R01HG010149].

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Breuss, M. W., Yang, X., Schlachetzki, J. C. M., Antaki, D., Lana, A. J., Xu, X., Chung, C., Chai, G., Stanley, V., Song, Q., Newmeyer, T. F., Nguyen, A., O'Brien, S., Hoeksema, M. A., Cao, B., Nott, A., McEvoy-Venneri, J., Pasillas, M. P., Barton, S. T., Copeland, B. R., Nahas, S., Van Der Kraan, L., Ding, Y., Glass, C. K., Gleeson, J. G., Gleeson, J. G., Breuss, M. W., Yang, X., Antaki, D., Chung, C., Averbuj, D., Courchesne, E., Ball, L. L., Roy, S., Weinberger, D., Jaffe, A., Paquola, A., Erwin, J., Shin, J., McConnell, M., Straub, R., Narurkar, R., Mathern, G., Walsh, C. A., Lee, A., Huang, A. Y., D'Gama, A., Dias, C., Maury, E., Ganz, J., Lodato, M., Miller, M., Li, P., Rodin, R., Borges-Monroy, R., Hill, R., Bizzotto, S., Khoshkhoo, S., Kim, S., Zhou, Z., Park, P. J., Barton, A., Galor, A., Chu, C., Bohrson, C., Gulhan, D., Lim, E., Lim, E., Melloni, G., Cortes, I., Lee, J., Luquette, J., Yang, L., Sherman, M., Coulter, M., Kwon, M., Lee, S., Lee, S., Viswanadham, V., Dou, Y., Chess, A. J., Jones, A., Rosenbluh, C., Akbarian, S., Langmead, B., Thorpe, J., Cho, S., Abyzov, A., Bae, T., Jang, Y., Wang, Y., Molitor, C., Peters, M., Gage, F. H., Wang, M., Reed, P., Linker, S., Urban, A., Zhou, B., Pattni, R., Zhu, X., Amero, A. S., Juan, D., Povolotskaya, I., Lobon, I., Moruno, M. S., Perez, R. G., Marques-Bonet, T., Soriano, E., Moran, J. V., Sun, C., Flasch, D. A., Frisbie, T. J., Kopera, H. C., Kidd, J. M., Moldovan, J. B., Kwan, K. Y., Mills, R. E., Emery, S. B., Zhou, W., Zhao, X., Ratan, A., Vaccarino, F. M., Cherskov, A., Jourdon, A., Fasching, L., Sestan, N., Pochareddy, S., and Scuder, S. (2022). Somatic mosaicism reveals clonal distributions of neocortical development. *Nature*, **604**(7907), 689–696.

Cleaver, J. E. (1969). Xeroderma pigmentosum: a human disease in which an initial stage of DNA repair is defective. *Proc Natl Acad Sci U S A*, **63**(2), 428–435.

Cohen, M. M. (1993). Proteus syndrome: clinical evidence for somatic mosaicism and selective review. *Am J Med Genet*, **47**(5), 645–652.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, **10**(2).

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H. P., van Es, M. A., McLaughlin, R., Sproviero, W., Shatunov, A., Jones, A., Al Khleifat, A., Pittman, A., Morgan, S., Hardiman, O., Al-Chalabi, A., Shaw, C., Smith, B., Neo, E. J., Morrison, K., Shaw, P. J., Reeves, C., Winterkorn, L., Wexler, N. S., Housman, D. E., Ng, C. W., Li, A. L., Taft, R. J., van den Berg, L. H., Bentley, D. R., Veldink, J. H., and Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res*, **27**(11), 1895–1903.

Dou, Y., Gold, H. D., Luquette, L. J., and Park, P. J. (2018). Detecting Somatic Mutations in Normal Cells. *Trends Genet*, **34**(7), 545–557.

Dou, Y., Kwon, M., Rodin, R. E., s Ciriano, I., Doan, R., Luquette, L. J., Galor, A., Bohrson, C., Walsh, C. A., and Park, P. J. (2020). Accurate detection of mosaic variants in sequencing data without matched controls. *Nat Biotechnol*, **38**(3), 314–319.

Fernández, L. C., Torres, M., and Real, F. X. (2016). Somatic mosaicism: on the road to cancer. *Nat Rev Cancer*, **16**(1), 43–55.

Fujimoto, A., Fujita, M., Hasegawa, T., Wong, J. H., Maejima, K., Oku-Sasaki, A., Nakano, K., Shiraishi, Y., Miyano, S., Yamamoto, G., Akagi, K., Imoto, S., and Nakagawa, H. (2020). Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res*, **30**(3), 334–346.

Glessner, J. T., Chang, X., Liu, Y., Li, J., Khan, M., Wei, Z., Sleiman, P. M. A., and Hakonarson, H. (2021). MONTAGE: a new tool for high-throughput detection of mosaic copy number variation. *BMC Genomics*, **22**(1), 133.

Hause, R. J., Pritchard, C. C., Shendure, J., and Salipante, S. J. (2016). Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med*, **22**(11), 1342–1350.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

Jam, H. Z., Li, Y., DeVito, R., Mousavi, N., Ma, N., Lujumba, I., Adam, Y., Maksimov, M., Huang, B., Dolzhenko, E., Qiu, Y., Kakembo, F. E., Joseph, H., Onyido, B., Adeyemi, J., Bakhtiari, M., Park, J., Javadzadeh, S., Jjingo, D., Adebiyi, E., Bafna, V., and Gymrek, M. (2023). A deep population reference panel of tandem repeat variation. *bioRxiv*.

Kim, T. M., Laird, P. W., and Park, P. J. (2013). The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell*, **155**(4), 858–868.

King, D. A., Sifrim, A., Fitzgerald, T. W., Rahbari, R., Hobson, E., Homfray, T., Mansour, S., Mehta, S. G., Shehla, M., Tomkins, S. E., Vasudevan, P. C., and Hurles, M. E. (2017). Detection of structural mosaicism from targeted and whole-genome sequencing data. *Genome Res*, **27**(10), 1704–1714.

Kraft, D. (1988). *A Software Package for Sequential Quadratic Programming*. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht. Wiss. Berichtswesen d. DFVLR.

Kristmundsdottir, S., Eggertsson, H. P., Arnadottir, G. A., and Halldorsson, B. V. (2020). popSTR2 enables clinical and population-scale genotyping of microsatellites. *Bioinformatics*, **36**(7), 2269–2271.

Kurek, K. C., Luks, V. L., Ayturk, U. M., Alomari, A. I., Fishman, S. J., Spencer, S. A., Mulliken, J. B., Bowen, M. E., Yamamoto, G. L., Kozakewich, H. P., and Warman, M. L. (2012). Somatic mosaic activating mutations in PIK3CA cause CLOVES syndrome. *Am J Hum Genet*, **90**(6), 1108–1115.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., cker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.

Lynch, H. T., Lynch, P. M., Lanspa, S. J., Snyder, C. L., Lynch, J. F., and Boland, C. R. (2009). Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet*, **76**(1), 1–18.

Maksimov, M. O., Wu, C., Ashbrook, D. G., Villani, F., Colonna, V., Mousavi, N., Ma, N., Lu, L., Pritchard, J. K., Goren, A., Williams, R. W., Palmer, A. A., and Gymrek, M. (2023). in the propensity for genome-wide short tandem repeat expansions in mice. *Genome Res*, **33**(5), 689–702.

Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., and Gymrek, M. (2021). TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics*, **37**(5), 731–733.

Raz, O., Biezuner, T., Spiro, A., Amir, S., Milo, L., Titelman, A., Onn, A., Chapal-Ilani, N., Tao, L., Marx, T., Feige, U., and Shapiro, E. (2019). Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Res*, **47**(5), 2436–2445.

Ruggieri, M. and Huson, S. M. (2001). The clinical and diagnostic implications of mosaicism in the neurofibromatoses. *Neurology*, **56**(11), 1433–1443.

Salipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H., and Pritchard, C. C. (2014). Microsatellite instability detection by next generation sequencing. *Clin Chem*, **60**(9), 1192–1199.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, **458**(7239), 719–724.

Sun, J. X., Helgason, A., Masson, G., ttir, S. S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., and Stefansson, K. (2012). A direct characterization of human mutation based on microsatellites. *Nat Genet*, **44**(10), 1161–1165.

Swami, M., Hendricks, A. E., Gillis, T., Massood, T., Mysore, J., Myers, R. H., and Wheeler, V. C. (2009). Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum Mol Genet*, **18**(16), 3039–3047.

ttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, **14**(2), 178–192.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., pfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., Marschall, T., Sedlazeck, F. J., Zook, J. M., Li, H., Koren, S., Carroll, A., Rank, D. R., and Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, **37**(10), 1155–1162.

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*, **14**(6), 590–592.

Yang, X., Xu, X., Breuss, M. W., Antaki, D., Ball, L. L., Chung, C., Shen, J., Li, C., George, R. D., Wang, Y., Bae, T., Cheng, Y., Abyzov, A., Wei, L., Alexandrov, L. B., Sebat, J. L., Gleeson, J. G., Averbuj, D., Roy, S., Courchesne, E., Huang, A. Y., D'Gama, A., Dias, C., Walsh, C. A., Ganz, J., Lodato, M., Miller, M., Li, P., Rodin, R., Hill, R., Bizzotto, S., Khoshkhoo, S., Zhou, Z., Lee, A., Barton, A., Galor, A., Chu, C., Bohrson, C., Gulhan, D., Maury, E., Lim, E., Lim, E., Melloni, G., Cortes, I., Lee, J., Luquette, J., Yang, L., Sherman, M., Coulter, M., Kwon, M., Park, P. J., Borges-Monroy, R., Lee, S., Kim, S., Lee, S., Viswanadham, V., Dou, Y., Chess, A. J., Jones, A., Rosenbluh, C., Akbarian, S., Langmead, B., Thorpe, J., Cho, S., Jaffe, A., Paquola, A., Weinberger, D., Erwin, J., Shin, J., McConnell, M., Straub, R., Narurkar, R., Jang, Y., Molitor, C., Peters, M., Gage, F. H., Wang, M., Reed, P., Linker, S., Urban, A., Zhou, B., Zhu, X., Amero, A. S., Juan, D., Povolotskaya, I., Lobon, I., Moruno, M. S., Perez, R. G., Marques-Bonet, T., Soriano, E., Mathern, G., Flasch, D., Frisbie, T., Kopera, H., Kidd, J., Moldovan, J., Moran, J. V., Kwan, K., Mills, R., Emery, S., Zhou, W., Zhao, X., Ratan, A., Jourdon, A., Vaccarino, F. M., Fasching, L., Sestan, N., Pochareddy, S., and Scuderi, S. (2023). Control-independent mosaic single nucleotide variant detection with DeepMosaic. *Nat Biotechnol*, **41**(6), 870–877.

Youssoufian, H. and Pyeritz, R. E. (2002). Mechanisms and consequences of somatic mosaicism in humans. *Nat Rev Genet*, **3**(10), 748–758.