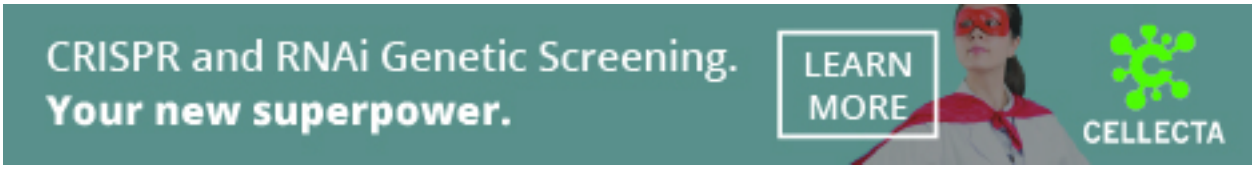
High-fidelity, large-scale targeted profiling of microsatellites

Caitlin A. Loh, Danielle A. Shields, Adam Schwing, et al.

Genome Res. 2024 34: 1008-1026 originally published online July 16, 2024
Access the most recent version at doi:10.1101/gr.278785.123



Method

High-fidelity, large-scale targeted profiling of microsatellites

Caitlin A. Loh,[1,2,3] Danielle A. Shields,[1,2,3] Adam Schwing,[1,2] and Gilad D. Evrony[1,2]

[1]*Center for Human Genetics and Genomics, New York University Grossman School of Medicine, New York, New York 10016, USA;* [2]*Department of Pediatrics, Department of Neuroscience & Physiology, Institute for Systems Genetics, Perlmutter Cancer Center, and Neuroscience Institute, New York University Grossman School of Medicine, New York, New York 10016, USA*

Microsatellites are highly mutable sequences that can serve as markers for relationships among individuals or cells within a population. The accuracy and resolution of reconstructing these relationships depends on the fidelity of microsatellite profiling and the number of microsatellites profiled. However, current methods for targeted profiling of microsatellites incur significant "stutter" artifacts that interfere with accurate genotyping, and sequencing costs preclude whole-genome microsatellite profiling of a large number of samples. We developed a novel method for accurate and cost-effective targeted profiling of a panel of more than 150,000 microsatellites per sample, along with a computational tool for designing large-scale microsatellite panels. Our method addresses the greatest challenge for microsatellite profiling—"stutter" artifacts—with a low-temperature hybridization capture that significantly reduces these artifacts. We also developed a computational tool for accurate genotyping of the resulting microsatellite sequencing data that uses an ensemble approach integrating three microsatellite genotyping tools, which we optimize by analysis of de novo microsatellite mutations in human trios. Altogether, our suite of experimental and computational tools enables high-fidelity, large-scale profiling of microsatellites, which may find utility in diverse applications such as lineage tracing, population genetics, ecology, and forensics.

[Supplemental material is available for this article.]

Microsatellites, also known as short tandem repeats, are genomic sequences composed of tandem repeats of short (1–6 bp) sequence motifs. Microsatellite loci vary widely in length and number of repeats, and there are more than one million microsatellite loci in the human genome (Ellegren 2004). The repetitive structure of microsatellites makes them highly mutable relative to other genomic sequences. The estimated de novo mutation rate of microsatellites is 1×10^{-4} to 1×10^{-3} per locus per generation, whereas the estimated de novo base substitution rate is 1.2×10^{-8} per base pair per generation (Sun et al. 2012; Kessler et al. 2020). Based on these de novo mutation rates, it is estimated that every cell division incurs several microsatellite mutations and that "silent" cell divisions with no microsatellite mutations are infrequent (Frumkin et al. 2005). This elevated mutability makes microsatellites highly variable within populations, and consequently, they are used extensively as markers of individuals members of a population in the fields of ecology (Selkoe and Toonen 2006; De Barba et al. 2017), forensics (Butler 2006; Moretti et al. 2016), and population genetics (Bruford and Wayne 1993; Putman and Carbone 2014). Microsatellites have also been utilized as markers of cells within organisms and tissues to reconstruct cell lineage histories (Reizel et al. 2012; Evrony et al. 2015; Wei and Zhang 2020).

The repetitive structure of microsatellites that makes them highly mutable in vivo and useful as lineage markers also has a downside: Microsatellites incur a higher frequency of in vitro artifacts than other genomic elements, which can confound accurate genotyping (Selkoe and Toonen 2006). Microsatellites mutate by polymerase slippage during replication, when transient denaturation and incorrect reannealing of the replicating strand leads to insertion or deletion of repeat units (Schlötterer 2000; Ellegren 2004; Bhargava and Fuentes 2010). This same process occurs during in vitro amplification of microsatellites, for example, in PCR. This produces an artifact pattern called "stutter" in which the final population of amplified molecules has a distribution of repeat unit counts with a peak at the true repeat unit count (i.e., the true genotype). Amplified microsatellites can be accurately analyzed by capillary electrophoresis despite this "stutter," because the highest signal peak representing the true genotype can be distinguished from adjacent "stutter" peaks that have lower signal (Wenz et al. 1998; Acquaviva et al. 2003; Vemireddy et al. 2007). However, the throughput of capillary electrophoresis is limited to small numbers of loci and samples (Hill et al. 2009; Guichoux et al. 2011).

High-throughput sequencing can profile many more microsatellites than capillary electrophoresis; however, the sequencing must have sufficient read depth, generally higher than that required for calling base substitutions, for the true genotype to be apparent among the stutter distribution (Willems et al. 2017). This requirement for higher than standard sequencing read depth is further exacerbated when using short-read sequencers, because only a fraction of reads will fully span each microsatellite to enable genotyping of its number of repeats. Stutter artifact and the read depth required for accurate genotyping can be mitigated by PCR-free sequencing (Fungtammasan et al. 2015); however, in some applications such as single-cell DNA sequencing and forensics, the amount of DNA available is low and requires amplification (Butler 2006; Evrony et al. 2015). Although high-depth whole-genome sequencing (WGS) of PCR-amplified libraries could achieve

[3]**These authors contributed equally to this work.**
Corresponding author: gilad.evrony@nyulangone.org
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278785.123.

34:1008–1026 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/24; www.genome.org

large-scale, high-fidelity profiling of microsatellites, this is not feasible to scale to hundreds of samples per condition or experiment. Here, we developed an integrated suite of computational and novel molecular methods for targeted sequencing and analysis of more than 150,000 microsatellite loci, achieving both cost-effective large-scale profiling of microsatellites and high fidelity. These tools may enable new applications of microsatellite profiling to a broad range of biological questions.

Results

Selection of microsatellite loci for profiling

We considered several factors when choosing microsatellite loci for targeted profiling. First, targeted profiling of microsatellites would benefit from capturing a panel of loci that are more likely to have mutations, because this would increase the power of resolving differences and lineage relationships among samples. Some microsatellite loci are less likely to be informative, because their mutation rates are relatively lower, such as loci with longer repeat motifs or lower repeat unit counts (Lai and Sun 2003; Ellegren 2004). Additionally, microsatellites must be in regions of the genome accessible to short reads, that is, with good mappability. Microsatellites in low-mappability regions, specifically loci with low-mappability (nonunique) flanking sequences, would yield sequencing reads that cannot be aligned well to the genome with short reads, thereby precluding analysis. Furthermore, loci with low mappability are poor targets for targeted profiling because the oligonucleotide probes used to target them would also hybridize to other undesired loci. For these reasons, it is essential that for any given microsatellite panel size, loci are selected that are most likely to be capturable, alignable to the genome, and informative. Although many computational tools exist to identify microsatellite loci across the genome (Lower et al. 2018), these lack the necessary informatic annotations and the ability to rapidly iterate across many analytic parameters to select such an optimized panel of microsatellite loci. We therefore developed a user-friendly, web-based app called Short Tandem Repeat Analysis and Target Identification (STRATIFY) that enabled us to design an optimized, large-scale panel of human microsatellites for targeted profiling (Fig. 1A).

First, we used the Tandem Repeats Finder (TRF) tool (Benson 1999) to annotate microsatellites in the human genome, and we optimized TRF's settings to maximize the number of annotated loci without introducing a large number of loci with highly imperfect motif repeats (Methods) (Supplemental Fig. S1A–D). We then annotated additional features of these microsatellite loci, including the following: (1) uninterrupted length (i.e., the length, in base pairs, of the longest span of perfect repeats of the microsatellite base motif without any indels or mismatches) and uninterrupted copy number (i.e., the maximum number of consecutive perfect repeats of the base motif without any indels or mismatches), as these parameters have been shown to contribute to the microsatellite mutation rate (Sun et al. 2012; Willems et al. 2016); (2) mappability (Karimzadeh et al. 2018) and GC content of the microsatellite and each of its 5′ and 3′ flanking sequences; (3) overlap between a given microsatellite and its neighboring microsatellite or the distance to the nearest microsatellite; (4) overlap with regions of the genome prone to alignment errors, for example, segmental duplications (Bailey et al. 2002) and the ENCODE Data Analysis Center Blacklist (Amemiya et al. 2019); (5) replication timing (Ryba et al. 2011; Marchal et al. 2018); and (6) estimated mutation rate based on a model taking into account the uninterrupted length and motif size of any given microsatellite (Gymrek et al. 2017).

We next built an interactive web application, STRATIFY, that allows filtering of the annotated microsatellites based on any of TRF's and our supplemental annotations (Supplemental Figs. S1E, S2). STRATIFY also dynamically updates plots of data for visualization and quality control (Supplemental Fig. S1F). This application was essential in allowing us to rapidly explore many combinations of parameters to select an optimal set of microsatellite loci for targeted profiling (Methods).

Probe design for hybridization capture of microsatellites

We chose a hybridization capture approach for targeted profiling of microsatellites, because it is scalable to hundreds of thousands of loci as evidenced by its use in whole-exome sequencing (Majewski et al. 2011). This contrasts with other amplicon-based targeted sequencing approaches, such as molecular inversion probes, that are typically limited to fewer than 20,000 loci (Mamanova et al. 2010), with the largest microsatellite panel to date profiling about 25,000 loci (Campbell et al. 2015; Wei and Zhang 2020; Tao et al. 2021). Scaling beyond the limits of amplicon-based approaches is important for microsatellite profiling, because the power to resolve relationships among samples or cells increases with the number of mutations and, consequently, with the number of microsatellite loci profiled (Gärke et al. 2012).

Importantly, hybridization capture of microsatellites cannot be performed efficiently with probes (oligonucleotides complementary to target sequences) directly targeting the microsatellite repeat sequences. First, the sequence of a microsatellite is often imperfect, with varying base substitutions and indels across loci and samples that would affect probe affinity. Second, probes designed to microsatellite motifs themselves cannot be targeted to specific loci and would therefore capture all loci with that motif, including those with nonideal features (e.g., low mappability). Additionally, the vast majority (97.6%) of microsatellites are shorter than the length of probes required by the temperature and stringency used to achieve specific genome-wide hybridization capture (55–120 bp) (Kruglyak et al. 1998; Samorodnitsky et al. 2015). Therefore, probes that both span the microsatellite sequence and extend into its flanks would have variable mismatches owing to variability in microsatellite length and sequence. For these reasons, we developed a different approach for capturing microsatellites that does not target the microsatellite sequence itself.

The sequences of the flanks adjacent to a microsatellite are much less likely to be variable across samples than the microsatellite itself, and they are also specific to each locus. We therefore decided to target our hybridization capture probes to the flanks of microsatellites. However, because standard capture probes are designed to directly overlap their target sequences, we first tested whether probes designed to microsatellite flanks would adequately capture reads fully spanning the adjacent microsatellite to enable genotyping. Specifically, we developed and evaluated three different probe design strategies for microsatellite flanks (Fig. 1B). In strategy 1, one flank of the microsatellite was targeted by a single probe. In strategy 2, both flanks of the microsatellite were targeted, each by a separate probe. In strategy 3, both flanks were targeted using a single probe: the left and right halves of the probe targeted the left and right flanks, respectively. In strategy 3, each flank binds to the probe, and the microsatellite sequence remains unhybridized as a single-stranded DNA "loop." The potential

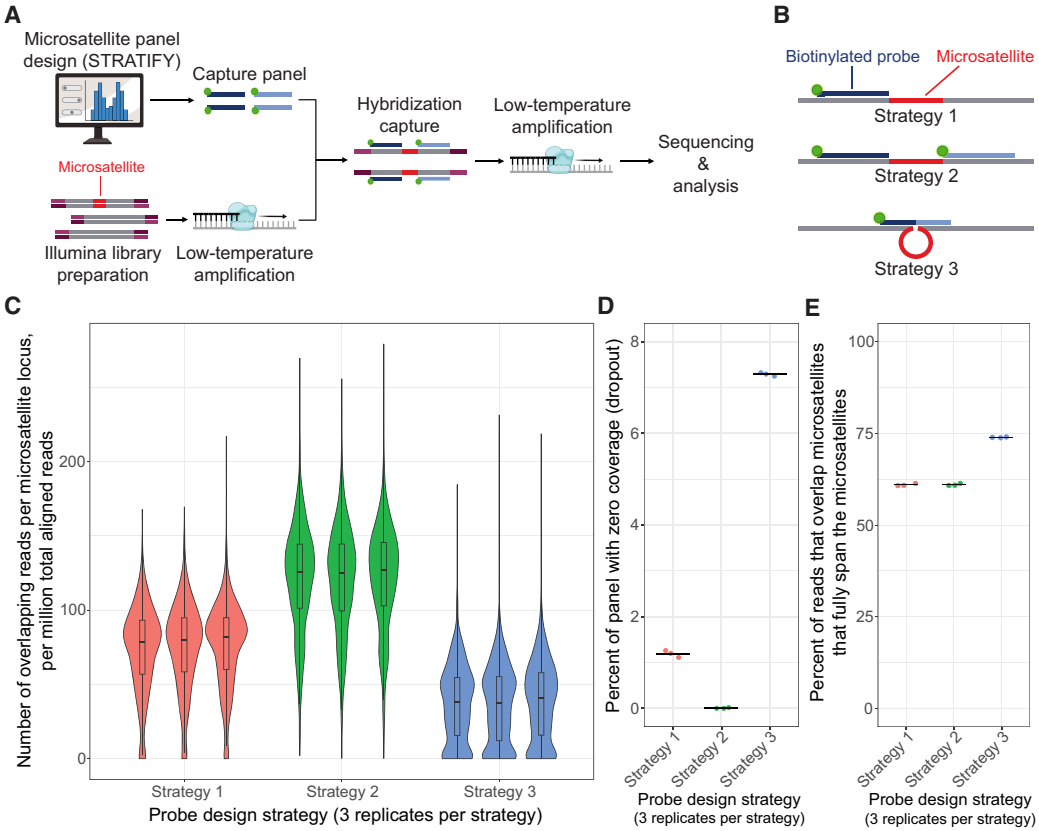


Figure 1. Overview of method and probe design strategy. (*A*) Schematic of microsatellite panel design, library preparation, hybridization capture, sequencing, and analysis. (*B*) The three probe design strategies tested in the pilot capture panel. (*C*) Distribution of coverage across targeted microsatellite loci for each probe design strategy. Normalized coverage is plotted as the [number of microsatellite-spanning reads at the locus]/[number of million aligned reads in the sample]. Box-and-whiskers show the first quartile, median, and third quartile of the distributions. Whiskers show 1.5× the interquartile range. (*D*) Fraction of loci in the panel with zero coverage (i.e., "dropout") for each probe design strategy. Mean value is represented by a black line. (*E*) Percentage of reads that overlap microsatellites that also fully span the microsatellites, for each probe design strategy. Calculated as [total number of reads fully spanning the targeted microsatellites]/[total number of reads overlapping targeted microsatellites by at least 1 bp] × 100. Mean value is represented by a black line. Note that *C*–*E* show experimental samples profiled using standard library amplification temperature.

advantage, if successful, of this unconventional probe strategy compared with strategies 1 and 2 would be a higher probability of capturing DNA fragments that fully span the microsatellite, which is required for genotyping. We subsequently used STRATIFY to obtain a set of loci that passed our design filters and that could be captured with all three design strategies (Methods) (Supplemental Fig. S3; Supplemental Table S1). We designed this first pilot panel to target only microsatellites with 2–4 bp motif lengths, because these loci are abundant and more mutable than those with longer motif lengths (Lai and Sun 2003). For each of the three strategies, we targeted 3333 random loci from this set of loci.

We performed hybridization capture with this pilot panel on three replicates of control genomic DNA (sample NA12877) (Supplemental Table S2) and evaluated the performance of each probe design strategy. As an initial metric of capture quality, we evaluated the distribution of coverage by reads fully spanning each microsatellite locus (Fig. 1C). Strategy 2 (both flanks targeted, each by a separate probe) had a significantly higher distribution of coverage than strategy 1 (one probe targeting one flank) and strategy 3 (one probe targeting both flanks), and strategy 3 had the lowest distribution of coverage. In strategy 2, only 0.005% of loci on average had complete dropout (zero coverage), whereas strategies 1 and 3 had 1.2% and 7.3% complete dropout, respectively (Fig. 1D; Supplemental Fig. S5A). The distribution across loci of the ratio of microsatellite coverage to flank coverage was also significantly higher for strategy 2 than for strategy 1 (Supplemental Fig. S4), consistent with targeting of both flanks more efficiently capturing DNA fragments containing the microsatellite. Notably, this ratio was highest for strategy 3, suggesting that when a locus is captured well by this design approach, a greater proportion of captured DNA fragments contain the microsatellite (Supplemental Fig. S4). However, this positive metric for strategy 3 is tempered by this design strategy's significantly worse absolute coverage (Fig. 1C–D), indicating that its absolute capture efficiency is lower than that of the other probe design strategies. Strategy 2 also performed best in terms of AT dropout, a measure of how undercovered AT-rich regions are compared with expectations based on the reference genome (Zhou et al. 2021), and fold 80 base penalty, a measure of capture uniformity (Supplemental Fig. S5B–D).

Overall, these analyses indicate that strategy 2, in which both flanks are targeted each by a separate probe, achieves the best capture performance for microsatellites, with nearly all loci captured with high average coverage. Although strategy 1 also performed well, especially in sequencing uniformity, its overall performance was lower than strategy 2. Strategy 3 had the worst performance,

with higher dropout and lower uniformity; whereas it achieved our goal of increasing the proportion of reads fully spanning the microsatellite (74% of reads overlapping targeted microsatellites by at least 1 bp also fully spanned the microsatellites, compared with 61% for both strategies 1 and 2) (Fig. 1E), this was counteracted by decreased overall capture efficiency (Fig. 1C,D; Supplemental Fig. S5A–D). We therefore chose to proceed with strategy 2 whenever a locus allowed design of probes to both flanks (based on design filters) and strategy 1 for loci that have only one flank passing our design filters. This combined approach using strategies 1 and 2 significantly increases the number of loci available for capture, because only one flank needs to pass our filters (Supplemental Fig. S6), and most loci are still captured efficiently with only one probe (Fig. 1C,D; Supplemental Fig. S5A–D).

Low-temperature amplification reduces microsatellite stutter artifacts

One of the most challenging aspects of microsatellite sequencing is the production of stutter artifacts during in vitro amplification (Ellegren 2004; Fungtammasan et al. 2015). In vivo, errors in microsatellite replication are corrected by mismatch repair machinery (Ellegren 2004), which is absent from in vitro amplification. In vitro amplification-induced stutter can obscure the true genotype of a microsatellite locus and confound downstream analyses. Although targeted hybridization capture can significantly improve the scalability of profiling microsatellites, it creates the additional challenge of requiring two in vitro PCR amplifications, before and after the hybridization capture. Therefore, we explored methods for reducing stutter artifact during in vitro amplification that could be applied to genomic libraries.

Stutter events that occur in early PCR cycles contribute more to the final stutter distribution. To mitigate this issue, we performed linear amplification cycles at the beginning of the PCR reaction that would prevent exponential expansion of stutter events. However, solely using linear amplification in our PCR protocol would not provide enough product for downstream steps of the protocol. We therefore decided to begin our protocol with three cycles of linear amplification and then add the second primer to initiate exponential PCR for the remainder of the reaction. To reduce the possibility of stutter events during this phase, we optimized our protocol to use the fewest possible number of amplification cycles required to achieve our desired DNA yield. We were able to reduce our cycle count to only three linear and two exponential cycles for the prehybridization PCR and to three linear and eight exponential cycles for the posthybridization PCR (Methods) (Fig. 2A).

We also suspected that temperature was one of the most important factors for reducing the frequency of stutter events, based on two prior studies that observed significantly decreased stutter when using low-temperature or isothermal amplification protocols (Hite et al. 1996; Daunay et al. 2019). In standard PCR, extension is typically conducted between 55°C and 72°C. At these elevated temperatures, the synthesized and template strands are more likely to denature and "slip" during reannealing, causing insertion or deletion of repeat units (i.e., stutter). If the temperature is lowered during extension, the microsatellite strands may be less likely to denature and cause stutter events. Hite et al. (1996) applied low-temperature amplification to one microsatellite locus by adding a thermolabile polymerase at each cycle with extension at 37°C, and they observed reduced stutter bands in gel electrophoresis compared with amplification at high temperature with thermostable polymerase. Daunay et al. (2019) applied recombinase polymerase amplification (RPA), an isothermal amplification method, and observed a similar reduction in stutter by capillary electrophoresis and amplicon sequencing, even when the reaction was multiplexed for three loci. However, low-temperature amplification has not yet been applied to large-scale microsatellite genomic libraries.

Here, we further developed the above low-temperature PCR method for large-scale microsatellite libraries (Fig. 2A,B). In our method, a thermolabile polymerase is added at each cycle of amplification, with extension occurring at 34°C. Denaturation in each cycle occurs at 98°C followed by rapid cooling in a water bath to minimize extension by any residual polymerase activity that survives the heat denaturation (Hite et al. 1996). The combination of this low-temperature annealing and extension, linear amplification, and fewer amplification cycles should significantly lower the amount of stutter observed in genomic sequencing libraries.

To test the ability of our low-temperature amplification method to reduce stutter in a genomic library, we profiled three technical replicates of NA12877 with the above pilot panel using low temperature for both the pre- and posthybridization capture amplification steps. Note that the standard-temperature pilot panel samples of the prior section and these low-temperature pilot panel samples were profiled identically, except for temperature, in the same experiment on the same day. We then compared the stutter levels of standard-temperature and low-temperature samples for loci on Chromosome X, because in male samples like NA12877, Chromosome X loci are single-allele loci and any read not matching the genotype reflects stutter. In contrast, stutter cannot be readily measured in biallelic loci (i.e., autosomal loci) because the stutter distribution of one allele may overlap the true peak of the other allele.

We found that low-temperature amplification significantly reduced the fraction of loci with high levels of stutter (Fig. 2C; Supplemental Table S3), as measured by the fraction of reads supporting the called genotypes (variant allele fraction [VAF]; higher VAF = lower stutter). Across all technical replicates of both temperature conditions, an average of 432 (range: 429–437) Chromosome X loci were genotyped. In the low-temperature samples, an average of 0.08% of genotyped loci had a VAF less than 0.75, whereas in the standard-temperature samples, an average of 19.6% of genotyped loci had a VAF less than 0.75. All loci with a VAF less than 0.75 had a motif length of 2 bp. In both standard- and low-temperature conditions, all 3 and 4 bp loci had VAFs of 0.85 or higher, although the low-temperature samples still had less overall stutter than the standard-temperature samples (Supplemental Table S3). Of note, 93% of stutter reads were shorter than the called allele (Supplemental Table S3), consistent with studies showing that PCR stutter preferentially decreases rather than increases microsatellite length (Ellegren 2004). Furthermore, the increased stutter seen in loci with a 2 bp motif length is consistent with prior studies that found a higher probability of microsatellite mutation for shorter motif lengths (Kruglyak et al. 1998; Ellegren 2004; Bhargava and Fuentes 2010; Fungtammasan et al. 2015).

We also observed that for each motif length, low-temperature amplification achieved lower stutter levels than standard-temperature amplification across all total microsatellite lengths (Fig. 2D). Within each motif length, longer microsatellites (i.e., loci with more repeat units) also generally had higher stutter levels—again consistent with prior studies (Fig. 2D; Ellegren 2004; Bhargava and Fuentes 2010; Sun et al. 2012). Together, these

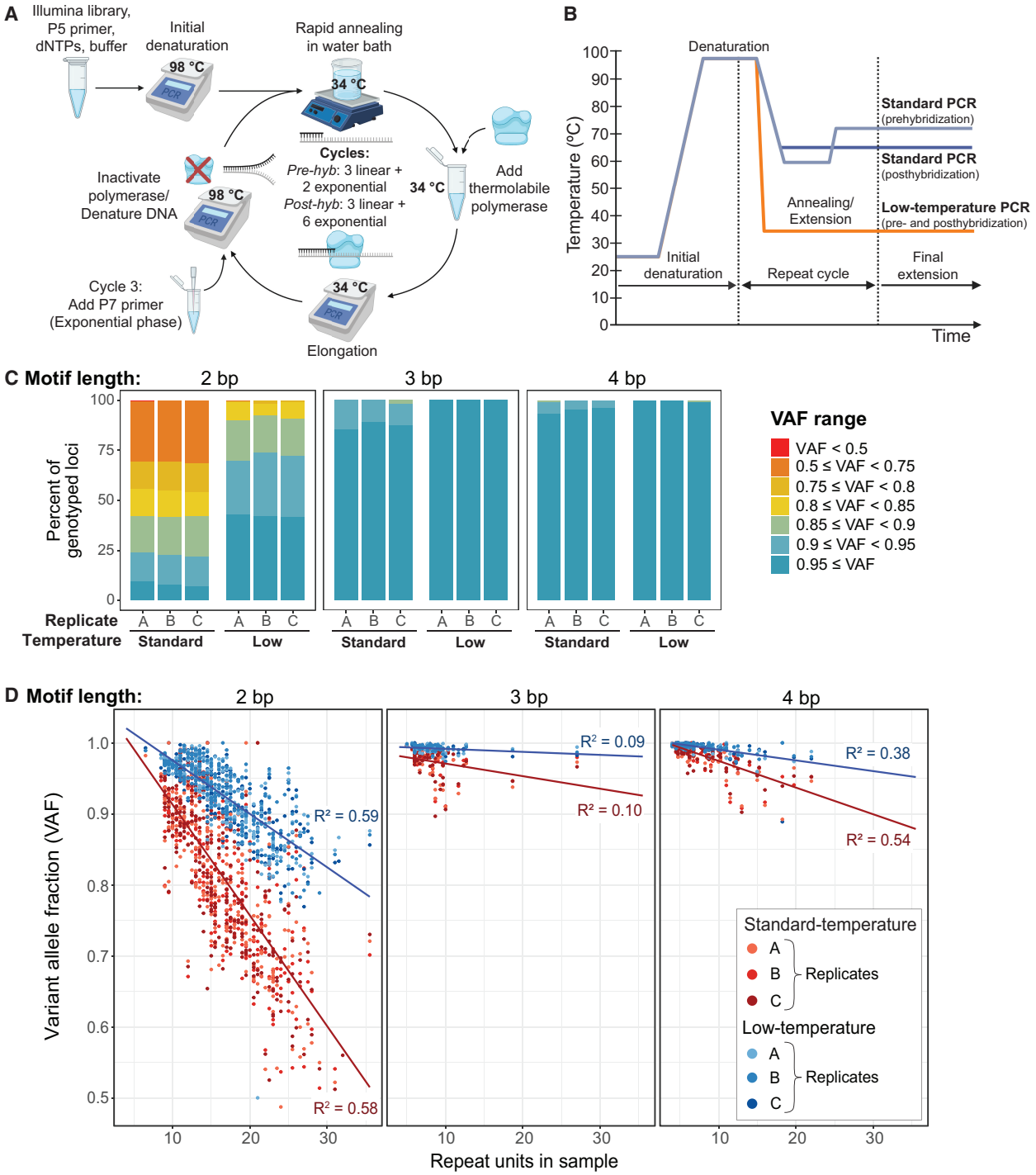


Figure 2. Impact of amplification temperature on microsatellite stutter. (*A*) Schematic of low-temperature PCR of sequencing libraries that significantly reduces microsatellite stutter artifact. (Pre-hyb) Prehybridization PCR; (Post-hyb) posthybridization PCR. The number of cycles specified is for the large-scale capture panel (Methods). (*B*) Schematic graph of temperatures during standard and low-temperature PCR protocols. Prehybridization and posthybridization standard PCRs use different thermal cycling protocols. In both standard- and low-temperature PCR, the prehybridization protocol consisted of three linear and two exponential PCR cycles, and the posthybridization protocol consisted of three linear and five exponential PCR cycles. (*C*) Fraction of genotyped loci on male Chromosome X microsatellite loci in three replicates of sample NA12877 (male) in different bins of variant allele fraction (VAF). Only Chromosome X loci from male individuals are included in the analysis to provide accurate stutter estimates, which is not feasible for biallelic loci. Loci are grouped by motif length, because motif length correlates with mutability. VAF = [number of reads supporting the allele genotyped by HipSTR]/[total number of reads at the locus calculated as the sum of the HipSTR MALLREADS field]. A VAF < 1 indicates stutter reads at the locus. Low-temperature amplification significantly decreased the fraction of loci with high levels of stutter (i.e., with low VAF), most noticeably for 2 bp motifs. (*D*) VAF of genotyped Chromosome X loci (as calculated in *C*) versus the number of repeat units called in the sample in three standard-temperature (red) and three low-temperature (blue) amplification replicates of individual NA12877. Red line indicates linear regression for VAF values in the standard-temperature condition; blue line, linear regression for VAF values in the low-temperature condition.

findings show that temperature is the most important factor identified to date in reducing stutter artifact during microsatellite amplification. Additionally, there were no significant differences in the distribution of coverage across loci or the quality of hybridization capture between standard- and low-temperature amplification (Supplemental Fig. S5A–D). Overall, our low-temperature amplification method achieves, for the first time, a significant reduction in stutter artifacts for all types of microsatellites in highly multiplexed, large-scale genome sequencing libraries. This represents an important step toward large-scale, accurate genotyping of microsatellites.

Large-scale profiling of microsatellites

After developing an optimal probe design strategy and a low-temperature amplification method for reducing stutter with a small-scale pilot panel of loci, we next proceeded to scale our method to a much larger number of microsatellite loci. A larger number of loci in the capture panel would better resolve differences among individuals or cells by increasing the number of detected mutations. Using our optimized approach for probe design (targeting both microsatellite flanks when possible, and only one flank when not possible) and an optimized set of filters, we produced a large-scale panel of 154,188 microsatellite loci (Supplemental Fig. S7A,B; Supplemental Table S4). Similar to the pilot panel, this panel included microsatellites with 2–4 bp motif lengths (147,656 loci) because of their relatively increased mutability compared with penta- and hexa-nucleotide loci (Lai and Sun 2003). The panel also included 6532 poly(A) loci, which have even higher mutation rates than do loci with 2–4 bp motifs (Ellegren 2004). Because poly(A) loci incur more significant stutter artifacts that make them challenging to genotype, especially in diploid chromosomes in which the signals of two alleles need to be deconvoluted, we only included poly(A) loci on Chromosome X (Supplemental Fig. S7B). We reasoned these may be amenable to genotype at least in male samples containing only a single Chromosome X. At the same time, our low-temperature amplification method with reduced stutter motivated the possibility that poly(A) loci may be feasible to genotype even in females. The size of this panel (28 Mb total probe length) is on the order of a whole-exome capture panel (~40 Mb) (Zhou et al. 2021), making it the largest-scale targeted profiling of microsatellites to date.

We applied this large-scale panel to five family trios (father, mother, child) (Fig. 3A; Supplemental Table S5), because Mendelian discordance analyses can be used for high-fidelity assessment of our method's accuracy as well as for optimization of analytic parameters. DNA from three trios was derived from lymphoblastoid cell lines (LCLs), and DNA from the other two trios was derived from blood samples. Either three or six samples were multiplexed per hybridization capture, and we sequenced the resulting libraries to an average of 873 total reads per locus (i.e., the total number of reads divided by the number of targeted loci) and an average of 218 reads fully spanning the microsatellite per locus. The distribution of coverage across loci was consistent across samples, indicating that our capture panel and low-temperature amplification are reliable and reproducible methods even at a very large scale (Fig. 3B). An average of 93.6% of reads were on-target (i.e., aligned to the targeted microsatellites), and dropout was very low (average of 1.7% of loci with zero coverage) (Supplemental Fig. S8A,B; Supplemental Table S5). AT dropout averaged 8.1%, similar to the pilot panels for probe design strategies 1 and 2 (Supplemental Figs. S5C, S8C). Finally, the average fold 80 base penalty was 1.6, which was similar to the fold 80 base penalty for strategy 1 in the pilot panel (Supplemental Figs. S5D, S8D).

We also assessed the level of molecular duplication in the sequencing data. Duplicate sequencing reads of the same original input DNA molecule may differ because of stutter artifacts. If molecular duplication levels were very high (e.g., more than 20 duplicates per molecule), this could enable creation of a high-fidelity consensus sequence for each original DNA molecule. However, achieving this level of duplication per molecule would require more than an order of magnitude more sequencing per sample, which is not scalable. In the absence of this, ideally there would be minimal duplication of molecules so that sequencing reads provide independent observations of a larger number of input DNA molecules. To assess the level of molecular duplication, we sequenced unique molecular identifiers (UMIs) that are present in our library adapters and found that nearly all molecules (98.1%) had only one or two molecular copies in the final sequencing data (Supplemental Fig. S9A,B). This very low molecular duplication rate indicates that our capture is highly efficient and that our optimized low number of PCR cycles maintains high library complexity. It also suggests that in the future, it may be feasible to multiplex more than six samples per hybridization capture for even better cost efficacy.

As a final evaluation of our large-scale panel's performance in terms of stutter artifact, we analyzed, for each motif length, the fraction of loci genotyped in different VAF ranges (per above, higher VAF indicates lower stutter) (Supplemental Fig. S10). As for the pilot panel, this analysis examined only Chromosome X loci (eight of the trio individuals were male) to obtain accurate estimates of stutter. For 2–4 bp loci, we saw similar stutter levels as in the pilot panel, confirming that the size of the panel does not impact the ability of our low-temperature amplification method to reduce stutter (Supplemental Fig. S10). Similar to the pilot panel, the fraction of loci with low VAF values decreased with increasing motif length (Supplemental Fig. S10). Poly(A) loci had more stutter than 2–4 bp motif loci, but most genotyped poly(A) loci (84.6%) still had a VAF $\geq$ 0.75 (Supplemental Fig. S10).

Ensemble computational pipeline for genotyping microsatellites

While we achieve efficient capture of more than 150,000 microsatellite loci with low levels of stutter, calling microsatellite genotypes from sequencing data requires specialized approaches (Gymrek et al. 2012; Treangen and Salzberg 2012). Indeed, microsatellite stutter artifacts and the increased error in calling insertions and deletions in sequencing data have led several groups to develop tools specifically for genotyping microsatellites (Treangen and Salzberg 2012; Fang et al. 2014; Raz et al. 2019). Preliminary application of a few of these tools—HipSTR (Willems et al. 2017), GangSTR (Mousavi et al. 2019), and ExpansionHunter (Dolzhenko et al. 2019)—to our data revealed complementary strengths and weaknesses. For example, although HipSTR produces accurate calls as seen in prior studies (Halman and Oshlack 2020; Oketch et al. 2024), it does not output a genotype for many loci (22% of loci on average in our samples) owing to HipSTR being unable to perform local realignment of the flanks. On the other hand, GangSTR and ExpansionHunter can call genotypes of loci that HipSTR does not call, but their accuracy is lower, especially at lower read depths (Halman and Oshlack 2020; Oketch et al. 2024). We therefore developed a computational pipeline, Short Tandem Repeat Ensemble Analysis Method (STREAM) (for schematic, see Fig. 3C; for detailed schematic, see Supplemental

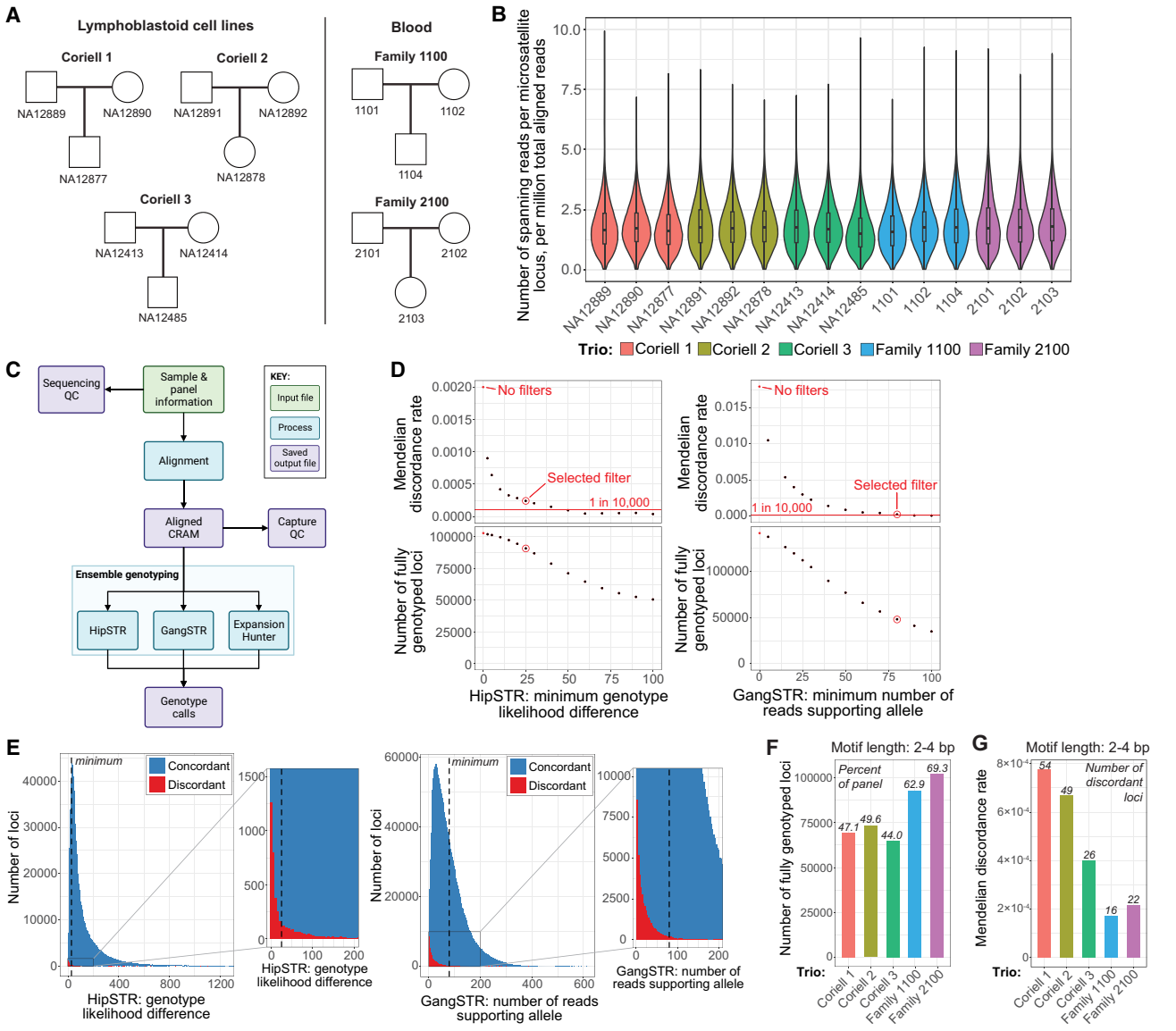


Figure 3. Large-scale microsatellite capture, genotyping pipeline STREAM, and Mendelian discordance analysis. (*A*) Five family trios profiled with our large-scale microsatellite panel. (*B*) Distribution of coverage across targeted microsatellite loci for the large-scale panel. Trios are arranged by trio in the following order: father, mother, child. Normalized coverage is plotted as the [number of microsatellite-spanning reads at the locus]/[number of million aligned reads in the sample]. Box-and-whisker plot shows first quartile, median, and third quartile of distribution. Whiskers show 1.5× the interquartile range. (*C*) Simplified schematic of STREAM, our microsatellite ensemble genotyping pipeline that optimally integrates calls from HipSTR, GangSTR, and ExpansionHunter. (QC) Quality control. A detailed schematic of the pipeline is shown in Supplemental Figure S9, C and D. (*D*) Representative plots of filter parameter optimizations for STREAM based on Mendelian discordance rates and the number of loci genotyped at different thresholds. Data are shown for Family 1100. Red dot indicates values without quality filtering; red circle, threshold value chosen for final filtering settings; and red line, approximate expected microsatellite de novo mutation rate (i.e., Mendelian discordance rate) based on prior studies (Weber and Wong 1993; Huang et al. 2002; Sun et al. 2012; Kristmundsdottir et al. 2023). Note that individual parameter filters do not achieve the expected discordance rate, and the final filter settings utilize an optimized combination of filters to achieve the expected discordance rate (Methods). For plots of other filter parameter optimizations, see Supplemental Figure S11. (*E*) Overlayed histograms of parameter values across all loci for two of the filtering parameters we optimized. Histograms are generated from data from all 15 samples captured with the large-scale panel. Plots include only fully genotyped loci, as only these loci have concordance calls. Concordant calls are blue, and discordant calls are red. The zoom plot shows the distribution of discordant calls near the final filtering threshold. Dashed line indicates the chosen threshold in our final filter settings. For similar plots of other filter optimization parameters, see Supplemental Figure S12. (*F*) Number of fully genotyped 2–4 bp motif length loci for each family trio, that is, loci with genotypes called for all three members of the trio. Italics indicate the fraction of the capture panel that is fully genotyped. (*G*) Mendelian discordance rate of fully genotyped loci with 2–4 bp motif lengths in each family trio. Mendelian discordance rate = [number of discordant loci]/[number of fully genotyped loci]. Italics indicate the number of discordant 2–4 bp motif length loci.

Fig. S9C,D), that combines calls from all three of these tools to genotype as many loci as possible while maintaining high call quality. Additionally, our trio experimental design provided a very low-level true signal of de novo mutations (discordant child–parent calls) that allowed us to optimize our pipeline's parameters based on expected de novo mutation rates from prior trio studies, which was then followed by our own experimental validation of candidate de novo mutations called by our pipeline.

STREAM first aligns reads and then runs HipSTR, GangSTR, and ExpansionHunter. Then, for each locus, STREAM chooses one of the three callers to provide the genotypes for all samples. This ensures that the genotypes of each locus can be compared across samples, because different callers may output different calls given the same data. The algorithm for choosing a caller for each locus and the order of caller priority was based on systematic optimization of filter parameter values to produce the expected Mendelian discordance rates per prior de novo mutation studies (about one in 10,000 loci) (Weber and Wong 1993; Huang et al. 2002; Sun et al. 2012; Kristmundsdottir et al. 2023) while maximizing the number of "fully genotyped" loci (i.e., loci genotyped in all members of the trio; Methods). We used Family 1100 for this initial optimization because its blood samples would be expected to have a lower discordance rate than LCLs, which may accumulate microsatellite mutations during cell culture. Family 1100 was also sequenced more deeply than the LCL trios and more evenly across trio members than Family 2100 (Supplemental Table S5).

We began this filter optimization by first testing a range of threshold values independently for each filter parameter (Fig. 3D; Supplemental Fig. S11A–J). This identified parameters that help lower the Mendelian discordance rate, such as mean quality, the difference in genotype likelihood between the called genotype and the next most likely genotype, and the number of reads supporting the called alleles. Next, we tested different combinations of these filters and caller preference orders until we achieved the expected low-level Mendelian discordance rate while maximizing the number of "fully genotyped" loci. Subsequently, we plotted the distribution of concordant and discordant calls across all trios to fine-tune the final filtering thresholds (Fig. 3E; Supplemental Fig. S12A–I). This filtering was performed separately for loci with 2–4 bp motif lengths and poly(A) loci because the higher stutter of poly(A) loci requires more stringent filtering thresholds. Additionally, for loci with 2–4 bp motif lengths, we monitored the balance of each motif length in the final set of genotypes versus the full panel to ensure that our filtering does not significantly bias the genotyped motif lengths (Supplemental Fig. S13A).

Our final optimized pipeline produced an average of 55% and 57% fully genotyped loci for 2–4 bp motif loci and poly(A) loci, respectively (Fig. 3F; Supplemental Fig. S13B). Three trios (Coriell 1, Coriell 2, and Coriell 3) had fewer fully genotyped loci (47% and 51% on average for 2–4 bp motif and poly(A) loci, respectively) compared with the other two trios (66% and 66% on average, respectively; Family 1100 and Family 2100) owing to the former trios' average 31% lower sequencing depth (Supplemental Table S5). This suggests that higher read depth increases the number of fully genotyped loci.

Importantly, for loci with 2–4 bp motif lengths, STREAM measured an average Mendelian discordance rate of 1.93×10^{-4} for trios whose DNA derived from blood (range: 1.72×10^{-4}–2.15×10^{-4}) and 6.15×10^{-4} for the trios whose DNA derived from LCLs (range: 4.00×10^{-4}–7.77×10^{-4}) (Fig. 3G). These results are consistent with expected microsatellite de novo mutation rates (Weber and Wong 1993; Huang et al. 2002; Sun et al. 2012) and the accumulation of additional mutations during divergence of the trio LCLs in culture. Loci with a motif length of 2 bp had the highest discordance rate in all trios, consistent with the higher mutability of 2 bp motifs compared with 3 and 4 bp motifs (Supplemental Fig. S13D; Ellegren 2004). Additionally, as expected for the more mutable poly(A) loci, we found they had an 8.1-fold higher discordance rate on average relative to 2–4 bp motif loci across all trios (Fig. 3G; Supplemental Fig. S13C). Ensemble calling, when compared with each caller individually, outputs more fully genotyped loci and lowers or maintains Mendelian discordance rates for all motif lengths (Supplemental Fig. S13E,F). Overall, these results support STREAM's ability to call microsatellites with high fidelity across a large panel of microsatellites.

To further demonstrate the specificity of STREAM, we validated discordant loci called in the trios Coriell 3, Family 1100, and Family 2100 using capillary electrophoresis (Supplemental Table S6). Capillary electrophoresis validated 97% (59 of 61) of loci with 2–4 bp motifs that were called as Mendelian discordant by STREAM. Three additional 2–4 bp motif loci were excluded from the validation, because upon manual review of sequencing data, we found that these apparent Mendelian discordant calls were caused by a heterozygous deletion in one of the parents that was inherited by the child. The two loci that did not validate by capillary electrophoresis were the result of miscalls by genotyping tools, because manual review of sequencing data was concordant with capillary electrophoresis. Poly(A) loci are more difficult to validate by capillary electrophoresis, owing to their broad stutter patterns. Nevertheless, we validated 12 of 13 poly(A) loci called as Mendelian discordant by STREAM whose capillary electrophoresis results were interpretable (six additional loci were excluded from analysis owing to broad stutter patterns that were not interpretable). These validation results demonstrate the high specificity of our method to detect rare microsatellite mutations.

Because total read depth is an important component for cost-effective scaling of our method to many samples, we further evaluated the dependence of our method's performance on read depth. We subsampled the data for each member of Family 1100 from 30× to 1013× average reads per locus (1013× is the average read depth of sample 1101, the lowest coverage sample from Family 1100). We found that the Mendelian discordance rate remained stable regardless of sequencing depth, but the number of fully genotyped loci was highly dependent on sequencing depth (Supplemental Fig. S14A). At 200× average coverage per locus, only one-quarter of loci are fully genotyped compared with the almost two-thirds of loci that are fully genotyped at full read depth (1013×), which may explain why discordant loci are not observed below 200× coverage, as they may not have been genotyped or passed our quality filters. These results emphasize the importance of high sequencing depths to genotype most loci and further highlight the importance of our targeted approach that makes these sequencing depths feasible.

To evaluate the cost-effectiveness of our method compared with PCR-free WGS, we also analyzed high-read-depth PCR-free WGS data from a family trio (Zook et al. 2016). We subsampled each family member's data from 30× to 309× average coverage (309× is the average read depth of the child's sample, the lowest coverage sample in the trio). We analyzed the trio using either the loci targeted by our large-scale capture panel or a larger genome-wide set of 563,770 microsatellite loci and calculated the cost per thousand fully genotyped loci (Methods) (Supplemental Fig. S14A,B; Supplemental Note). When analyzing only the loci from our large-scale panel, our method was significantly more cost effective than WGS at all coverage levels necessary to achieve detection of Mendelian discordant loci. When analyzing the genome-wide set of microsatellite loci, WGS achieves a comparable cost per 1000 fully genotyped loci (at a read depth required for reliable detection of Mendelian discordant loci) with an approximately 3.2-fold higher number of fully genotyped loci, albeit with a 3.8-fold higher cost per sample (Supplemental Fig. S14A). Furthermore, our targeted capture method becomes increasingly

more cost effective as average read depth increases, which may be needed for some applications to maximize the number of samples genotyped at each locus (Supplemental Fig. S14A,B).

Discussion

Microsatellites are the most mutable genomic elements, and their consequent variation has enabled diverse applications in population genetics, forensics, and lineage tracing. Here, we introduce four tools and innovations that will increase the scalability and resolution of microsatellite profiling: (1) STRATIFY, a tool for designing large-scale microsatellite capture panels; (2) a novel method for the largest-scale targeted capture of microsatellite loci, to date; (3) a method for low-temperature amplification of large-scale microsatellite libraries that significantly reduces stutter artifact; and (4) STREAM, an ensemble genotyping pipeline integrating three microsatellite callers for more comprehensive and accurate genotyping.

Current microsatellite profiling methods are limited by their scalability. Capillary electrophoresis can only process a small number of loci, whereas WGS can profile microsatellites genome-wide, but the read depths required for accurate genotyping are not currently feasible to scale to hundreds of samples per experiment. One of the most important advantages of our method is its ability to readily scale to more than 150,000 loci, comparable in size to a whole-exome panel, which is important for maximizing the number of captured mutations for meaningful biological analyses. Our method captures significantly more loci than other recently reported methods, the largest of which captures about 12,000 loci per sample (Wei and Zhang 2020; Tao et al. 2021). Additionally, we can capture loci with diverse motifs from across the genome, in contrast to methods that focus on specific motifs (Wang et al. 2022) or that utilize restriction enzymes to fragment DNA, which limits the number of targetable loci (Wei and Zhang 2020).

Stutter has long been one of the most challenging problems in microsatellite studies. Prior methods have achieved lower stutter for small numbers of loci (Hite et al. 1996; Daunay et al. 2019) or for specific microsatellite motifs such as "G" homopolymers (Wang et al. 2022). Notably, our low-temperature amplification method achieves, for the first time, significant reduction of stutter in a large, complex sequencing library across many types of microsatellite motifs and motif lengths.

Additionally, emerging sequencing technologies may further reduce stutter. We compared stutter profiles of PCR-free WGS sequenced by the recently released Pacific Biosciences (PacBio) Onso platform, which uses sequencing by binding, and by Illumina sequencing, which uses sequencing by synthesis. In both healthy individual 1104 and in an individual with constitutional mismatch repair deficiency syndrome (CMMRD), we found significantly lower stutter in poly(A) microsatellites in Onso compared with Illumina sequencing (Supplemental Fig. S15). Because poly(A) microsatellites are the most abundant and mutable microsatellite motif, this finding could lead to new applications of microsatellite profiling. Additionally, the individual with CMMRD exhibited significantly elevated stutter compared with that of the healthy individual, supporting the potential of WGS as a potential diagnostic tool for CMMRD (Supplemental Fig. S15; Chung et al. 2023).

Concomitant with our novel molecular methods, we introduce two analytic tools, STRATIFY and STREAM, that facilitate the design and analysis, respectively, of large-scale microsatellite panels. STRATIFY allows for rapid iteration of different parameter thresholds for facile curation of panels, while STREAM combines the strengths of three microsatellite genotyping tools while mitigating their individual weaknesses to call genotypes accurately. During preparation of this paper, a new tool called EnsembleTR was published that also utilizes HipSTR, GangSTR, and ExpansionHunter for ensemble genotyping of microsatellites (Ziaei Jam et al. 2023). STREAM and EnsembleTR differ in their analysis process. STREAM utilizes more fine-grained optimization of filter settings for each individual caller and then chooses the final genotype per an optimized order of caller preference, whereas EnsembleTR performs lenient filtering for each individual caller and then chooses the final genotype based on quality scores summed across individual callers. EnsembleTR's approach requires it to reconcile differences in allele representations of each locus across the individual callers, whereas STREAM selects the final genotype directly from individual callers. To illustrate the differences between the two tools, we processed PCR-free 100× Illumina WGS data with both tools and found that for all motif lengths, EnsembleTR outputs more fully genotyped loci, whereas STREAM produces a lower Mendelian discordance rate (Supplemental Fig. S16A,B). These differences are most likely owing to the different filtering threshold parameters chosen by each group, with EnsembleTR prioritizing sensitivity and STREAM prioritizing specificity. Despite these differences, both tools highlight the advantage of an ensemble approach for maximizing genotyping accuracy for microsatellites and the number of loci genotyped.

Homopolymers are often excluded from microsatellite profiling methods owing to their significant stutter artifacts. Here, we were able to successfully capture and genotype poly(A) homopolymer loci owing to the lower level of stutter achieved by our low-temperature amplification. Our success in genotyping poly(A) loci on Chromosome X in females (in Family 2100, 51% of loci were fully genotyped with a Mendelian discordance rate of 6.1×10^{-4}) suggests that we could extend profiling of poly(A) loci to all chromosomes. Given that poly(A) repeats are the most abundant and most mutable type of microsatellite in the genome, including more poly(A) loci in future panels would increase the number of mutations observed per sample and further improve the resolution of analyses.

We analyzed the cost-effectiveness of our method compared with PCR-free WGS and found that our method is more cost effective than WGS at high read depths, or when a smaller number of loci need to be profiled for the biological application (Supplemental Fig. S14A,B; Supplemental Note). Therefore, the most cost-effective method depends on two key considerations: (1) the number of genomic loci required by the biological application and (2) the importance of maximizing the fraction of samples genotyped at each locus; for example, lineage tracing applications would benefit from a higher read depth that maximizes the number of samples genotyped at each locus. Samples with baseline stutter that require a higher read depth for accurate genotyping, such as amplified single-cell genomes, may also benefit from the higher cost-effectiveness of our method at higher read depths. Additionally, for applications requiring fewer loci (e.g., molecular markers for ecology), our method is far more cost effective than WGS. Panel design is also an important consideration. Although in this study we selected loci with more "perfect" motif repeats to maximize the probability of capturing mutations that are useful for most applications, the Mendelian discordance rates we measured may overestimate genome-wide rates. At the same time, as sequencing costs continue to decrease, high-depth PCR-free WGS has benefits in terms of ease of sample preparation, uniformity of genomic

coverage, and profiling of more microsatellites per sample, such that it is likely to become the preferred method for applications requiring profiling of hundreds of thousands of loci.

Though we were able to reduce the level of stutter in our final libraries, we still observe a low-level stutter signal that confounds some genotype calls, and our filtering to remove these false calls eliminates the corresponding loci from the final analysis. Further reduction of stutter through temperature modification is unlikely. A recent study reduced stutter by targeted mutagenesis (Wang et al. 2022) but is limited to "G" homopolymer microsatellites that would not be compatible with the scale of capture we have achieved. Thus, stutter continues to be a challenge despite the improvements we have demonstrated, motivating further research to address it, perhaps by novel polymerases (Yamanoi et al. 2021) or accessory proteins during library amplification.

One limitation of our method is that the low-temperature amplification requires the manual transfer of samples between temperature conditions and the addition of enzyme each cycle. Automation of this amplification process is feasible, either through liquid-handling robots or with microfluidics (Ahrberg et al. 2016), and this would lower the barrier to implementing this protocol while also enabling higher throughput. Another limitation of our method is that STREAM, our ensemble microsatellite genotyping tool, is limited by the accuracy and genotyping capabilities of the three microsatellite callers it utilizes. For example, HipSTR does not genotype some loci owing to failure in local realignment. Our open-source pipeline can be readily adapted to incorporate new microsatellite genotyping tools, while our trio sequencing data sets could then be used to refine these tools' new parameters.

The ability of our method to profile more than 150,000 microsatellites across many samples with high fidelity may help elucidate previously opaque relationships among individuals in a population or among cells using mutational processes that are ubiquitous in all organisms and cells. Capture of the high endogenous levels of microsatellite mutations across many loci will allow for reconstruction of these relationships at a higher resolution than would be achievable with other markers that mutate less frequently. This higher resolution across a large number of samples will be useful for diverse applications, such as single-cell lineage tracing of healthy tissues and tumors (Wei and Zhang 2020; Tao et al. 2021) and ecology studies seeking to study relationships among many individuals in a population (De Barba et al. 2017). Our method also opens new opportunities to study non-model organisms without the need for engineered molecular markers, which is especially important for retrospective studies of human tissues.

Methods

Microsatellite panel design tool: STRATIFY

Identifying microsatellite loci in the genome

We identified microsatellite loci in the hg38 human reference genome with TRF version 4.09.1 (Benson 1999). We selected TRF because it provides flexibility in tuning microsatellite annotation thresholds. TRF operates by k-tuple matching of short strings (termed "probes") across the reference sequence. TRF uses three sets of parameters to identify microsatellites: alignment parameters, conservation parameters, and a selection parameter. The alignment parameters include matching weight, mismatch penalty (MP), and indel penalty (IP). Using these, an alignment score is calculated for a candidate microsatellite of length n relative to a perfect microsatellite motif of length m by summing the matching weights and subtracting the MPs and IPs for each position. Candidate loci with an alignment score exceeding a user-defined threshold are output by TRF. This minimum alignment score also indirectly sets the minimum length of annotated microsatellites because a perfect microsatellite of length n would have an alignment score of $n \times$ matching weight.

The conservation parameters of TRF specify the minimum acceptable average percent of identity (match probability) and the maximum acceptable average percent of insertions and deletions (indel probability) between all adjacent copies of a motif in a candidate microsatellite. The selection parameter of TRF specifies the maximum period size (i.e., maximum length of the motif).

To determine settings for TRF, we reviewed settings used by 13 prior studies (Benson 1999; Gardner et al. 2002; Denœud and Vergnaud 2004; Boby et al. 2005; Domaniç and Preparata 2007; Leclercq et al. 2007; De Grassi and Ciccarelli 2009; Ryba et al. 2011; Willems et al. 2016; Bilinski et al. 2017; Gymrek et al. 2017; Karimzadeh et al. 2018; Marchal et al. 2018) and the settings recommended by the developer of TRF (Benson 1999). MPs and IPs varied most among surveyed studies. To find the best parameters for identifying as many microsatellites as possible while excluding loci with highly imperfect motif repeats, we tested TRF with 36 combinations of MP and IP values, from least stringent (MP and IP both = 3) to most stringent (MP and IP both = 20). This analysis showed the following: (1) IP tends to have a stronger effect on the overall percent of match (a measure of the "perfection") of a microsatellite than MP (Supplemental Fig. S1B,C), (2) higher MPs and IPs yield shorter and more perfect microsatellites (Supplemental Fig. S1B,C), and (3) very high MP and IP values (MP = 20 and IP = 20) do not yield major improvements in microsatellite quality (i.e., more perfect repeats) over strict, but not as high, penalty values (e.g., MP = 7 and IP = 7), and they exclude many microsatellites that may still have elevated mutation rates despite having some imperfect repeats (Supplemental Fig. S1B,C). Additionally, more lenient parameter values did not significantly change the length or motif distribution of the pool of returned microsatellites, and the frequencies of different motifs out of all annotated loci were stable across parameter values (Supplemental Fig. S1C,D).

Based on these analyses, our final TRF parameters were as follows: matching weight = 2 (value used in all surveyed studies), MP = 7 and IP = 7 (values recommended by the TRF manual), match probability = 80 and indel probability = 10 (values used in nearly all surveyed studies), minimum alignment score = 36, and maximum period length = 6. Note that this minimum alignment score sets the minimum total microsatellite length to 18 bp and the minimum number of repeats of three for loci with the maximum period size of six.

Microsatellite feature annotations

TRF annotates each microsatellite with several features: chromosome, start/end position, period size, copy number (i.e., the number of repeats of the microsatellite motif), consensus motif (i.e., the best-fitting motif of the microsatellite as identified by TRF), consensus size (i.e., the length of the consensus motif, which may or may not be equal to the period size), percent match, percent indel, alignment score (i.e., the sum of the matching weight, MP, and IP), % A/C/G/T, entropy (i.e., the balance of the base composition of the microsatellite; a microsatellite with 25% A, 25% C, 25% G, and 25% T has maximal entropy), and microsatellite sequence.

We also added many additional annotations to these data as follows. We derived the microsatellite motif from the TRF consensus motif, because for some loci, the TRF consensus motif was

>7 bp, although the maximum period size was set to six (this has to do with how TRF works, as it defines consensus motif as the motif that maximizes the alignment score of the called microsatellite, rather than simply using the motif that the program used to call the microsatellite in the first place). We excluded microsatellites with a consensus size of more than seven, and for consensus motifs with a size of seven, we compared it to all possible 6 bp motifs using the adist function in R and changed it to the most similar motif. We did this because our definition of a microsatellite included a maximum motif size of 6 bp. After defining the microsatellite motif, we added an annotation for motif family, which is the motif after collapsing all circular permutations (e.g., ACG, CGA, GAC) and reverse complements of the microsatellite motif and its permutations (e.g., the reverse complement of ACG would be CGT, and then its circular permutations would be CGT, GTC, TCG) into a single representative motif family (e.g., the "ACG" motif family includes microsatellites with any of the following motifs: ACG, CGA, GAC, CGT, GTC, and TCG). We also excluded microsatellites that have another microsatellite overlapping >90% of its span. Next, we added annotations for uninterrupted length (i.e., the length, in base pairs, of the longest span of perfect repeats of the microsatellite base motif without any indels or mismatches) and uninterrupted copy number (i.e., the maximum number of consecutive perfect repeats of the base motif without any indels or mismatches), as these parameters have been shown to positively correlate with the microsatellite mutation rate (Hutter et al. 2006; Gupta et al. 2007). We also annotated the 150 bp flanking sequences of the microsatellites, the overlap between each microsatellite and its neighboring microsatellite(s), the distance to the next nearest microsatellite in both 5′ and 3′ directions, and the percentage of GC content and mappability (using 24, 36, 50, and 100 *k*-mers) (Madsen et al. 2008) for each microsatellite and its 5′ and 3′ flanking regions. We further annotated the coverage depth (normalized to genome-wide average coverage) for each microsatellite and its flanking regions in a prior single-cell sequencing data set amplified by primary template-directed amplification (PTA; Bioskryb) (Gonzalez-Pena et al. 2021) in anticipation of potential future design of panels for profiling single-cell genomes that have undergone initial nonuniform single-cell genome amplification. We also added annotations for replication timing based on 4D nucleome data in GM12878 lymphoblastoid cells, as well as Repli-chip data of neural progenitor cells (Mizuta et al. 2004; Mousavi et al. 2019). Finally, we annotated estimated mutation rates calculated by a linear regression model we created with the "lm" function in R (with default settings) trained on mutation rate data of autosomal intergenic microsatellite loci from a prior study (Gymrek et al. 2017); the model specifically utilized the "uninterrupted length" and "motif size" annotations to predict mutation rate, because these best explained the variance in the test data set (Gymrek et al. 2017).

Building STRATIFY

We built a user-friendly, web-based app that allows users to filter microsatellites based on any combination of the above annotations (Supplemental Fig. S1E). It also dynamically updates several plots for data visualization and quality control (e.g., histogram of common microsatellite motifs by length and histogram of estimated mutation rates) (Supplemental Fig. S1F). Once the desired filters are selected, the microsatellite panel can be exported along with any desired annotations. The app utilizes the Shiny platform (https://shiny.posit.co/) along with the following packages for R statistical software version 4.2.3 (R Core Team 2023): memoise, shiny, shinyjs, shinyWidgets, shinybusy, shinyBS, shinyalert, dplyr, readr, data.table, ggplot2, and GenomicRanges. For a link to the full open-source code and instructions for setting up STRATIFY, see the Software Availability section below.

Probe panel design

Pilot panel

We used STRATIFY to design the pilot hybridization capture panel, testing three different probe design strategies: strategy 1 with only one flank of the microsatellite targeted by a 120 bp probe; strategy 2 with both flanks of the microsatellite targeted, each by a separate 120 bp probe; and strategy 3 with both flanks of the microsatellite targeted by a single 120 bp probe (half the probe complementary to the 5′ flank and the other half complementary to the 3′ flank). Using STRATIFY's adjustable filters, we iterated systematically through combinations of parameters to find filtering thresholds for each strategy that removed unsuitable loci and left at least 100,000 loci for possible targeting (Supplemental Fig. S6). Once we determined the optimal filtering thresholds for each strategy separately, we combined these filters to accommodate all three design strategies.

The strictness of the filtering for the combination of strategies was largely determined by strategy 3, because both flanks needed to pass the thresholds and because we needed to use a stricter flank mappability filter because only 60 bp from each flank would be in the probe design and because shorter sequences typically have lower mappability scores (Supplemental Figs. S3, S6). The full set of requirements for loci in the pilot panel was as follows: (1) total microsatellite length ≤100 bp, (2) period size between 2 and 4 bp, (3) percent match ≥90%, (4) percent indel ≤10%, (5) uninterrupted copy number of four or more, (6) distance to the nearest microsatellite in both directions ≥100 bp, (7) GC content in both the microsatellite and its flanks ≤70%, (8) 50 *k*-mer mappability score in both 90 bp flanks ≥0.95 (note that the 50 *k*-mer mappability was used to ensure uniqueness of the 60 bp flanking regions targeted by strategy 3), (9) single-cell PTA normalized coverage (see above) between 0.4 and two in both flanks to target more uniformly amplified regions that may be useful in future single-cell applications, (10) neither 90 bp flank overlaps either the 1000 Genomes Project pilot accessibility mask (regions that are challenging to resolve by short-read sequencing) (The 1000 Genomes Project Consortium 2015) or the ENCODE Data Analysis Center (DAC) blacklist regions (regions that are prone to artifacts and erroneous signal in short-read sequencing) (Amemiya et al. 2019), and (11) neither 90 bp flank or the microsatellite overlap annotated segmental duplications (Supplemental Fig. S3; Bailey et al. 2002). It was found that 102,456 loci passed these filters and were submitted to Twist Bioscience for probe design and synthesis.

Twist Bioscience applied proprietary quality filters to the full set of loci and assembled a list of approximately 83,000 loci that passed Twist's filters in both flanks. From these loci, we randomly assigned 3333 loci to each design strategy, for a total of 9999 loci in the panel (Supplemental Table S1). The probes were then designed as described for each strategy and delivered as a single, combined panel with each probe present at an equal concentration.

Large-scale panel

We used STRATIFY to design the large-scale hybridization capture panel, which targeted microsatellites with either strategy 1 (only one flank passes filters) or strategy 2 (when both flanks pass filters). The full set of requirements for loci in the large-scale panel was as follows: (1) period size between 2 and 4 bp, (2) percent match ≥90%, (3) percent indel ≤10%, (4) uninterrupted copy number of four or more, (5) GC content in both the microsatellite and its flanks ≤70%, (6) distance to the nearest microsatellite ≥100 bp

for at least one direction, (7) 100 *k*-mer mappability score in at least one of the two 90 bp flanks ≥0.95, (8) single-cell PTA normalized coverage between 0.4 and two in both flanks, (9) maximum of one 90 bp flank overlaps either the 1000 Genomes Project pilot accessibility mask (The 1000 Genomes Project Consortium 2015) or the ENCODE Data Analysis Center Blacklist Regions (Amemiya et al. 2019), (10) at least one 90 bp flank and the microsatellite itself do not overlap annotated segmental duplications (Bailey et al. 2002), and (11) at least one 90 bp flanks and the microsatellite itself do not overlap pseudoautosomal regions (Supplemental Fig. S7A). Importantly, we required for each microsatellite locus that the mappability filters, region filters, and distance to the nearest microsatellite filter all pass in the same flank (for at least one flank). It was found that 173,996 loci passed these filters. Note that the "percent match" and "percent indel" filters were chosen to retain loci with more "pure" repeat sequences, because these loci are more highly mutable, which increases the probability of capturing mutations among samples in an experiment (Sun et al. 2012; Willems et al. 2014). Additionally, for this large-scale panel, we removed the filter for maximum microsatellite length, because the number of repeat units in any sample may vary from the reference genome and because only 3.2% of microsatellites are longer than the 100 bp length limit used for the pilot panel.

We separately selected poly(A) loci on Chromosome X using the same filters as for the above large-scale panel of 2–4 bp microsatellites, but with the following filters removed: (1) percent match and percent indel, which are likely to be lower given the mutability of poly(A) loci; (2) uninterrupted copy number, because poly(A) loci are very mutable and therefore unlikely to match uninterrupted copy number in the reference genome; (3) GC content of the microsatellite, because poly(A) loci have a GC content ∼0%; and (4) GC content of the flanks, in order to retain more loci. It was found that 9163 poly(A) loci passed these filters (Supplemental Fig. S7B).

We submitted the 173,996 loci with 2–4 bp motifs and the 9163 poly(A) loci to Twist Biosciences, who applied final proprietary filters for probe design and synthesis. Twist was able to design probes for 147,656 loci with 2–4 bp motifs and for 6532 poly(A) loci by targeting at least one flank (Supplemental Table S4). Of these, 74,635 loci with 2–4 bp motifs and 1205 poly(A) loci were targeted by probes in both flanks (strategy 2), and the rest were targeted by a probe in only one flank (strategy 1).

Sample sources

Genomic DNA from LCLs was obtained from the Coriell Institute. Genomic DNA for Family 1100, Family 2100, and an individual with congenital mismatch repair deficiency syndrome (CMMRD; subject ID BTM-1400; biallelic *PMS2* loss-of-function mutations) was extracted with the MagAttract HMW DNA kit (Qiagen) from whole-blood samples of individuals enrolled in human subject research protocols approved by the New York University Grossman School of Medicine Institutional Review Board.

Library preparation for targeted microsatellite profiling

Genomic DNA was fragmented on an LE220 instrument (Covaris) using a 96 microTUBE plate (Covaris) in a volume of 55 µL with a buffer of 10 mM Tris (pH 8). Instrument settings were 450 W peak incident factor, 20% duty factor, 1000 cycles per burst, and a total treatment time of 200 sec. Fragmented DNA was profiled on a High Sensitivity D1000 ScreenTape TapeStation assay (Agilent). Fragmented DNA ranged in size from 50 bp to 500 bp.

Next, we performed a nick ligation reaction to seal nicks caused by fragmentation, which may otherwise allow library preparation polymerases to initiate synthesis of new DNA strands with stutter artifacts. Nick ligation was performed in a 32 µL reaction containing 500 ng of fragmented genomic DNA, 3.73 µL NEBNext Ultra II End Prep Reaction Buffer (NEB), 26 µM NAD+, and 0.5 U/µL *Escherichia coli* DNA Ligase (NEB). The reaction was incubated for 30 min at 16°C. For samples with lower concentrations (i.e., blood samples), the nick ligation reaction was scaled up to 65 µL to accommodate a larger volume of fragmented DNA.

The end repair and adaptor ligation steps of library preparation were performed with the NEBNext Ultra II DNA library prep kit for Illumina (NEB) either at the standard total reaction volume of the manufacturer's protocol or in a reaction volume scaled to 117% of the manufacturer's protocol for samples for which the nick-ligated DNA had low concentration. The standard volume (60 µL) end repair reaction was performed by combining nick-ligated DNA, 3.27 µL NEBNext Ultra II End Prep Reaction Buffer (NEB), 3 µL NEBNext Ultra II End Prep Enzyme Mix (NEB), and nuclease-free water (NFW). The reaction was incubated for 30 min at 20°C and then for 30 min at 65°C. Subsequently, the standard volume (93.5 µL) adaptor ligation was performed by combining 2.5 µL of 15 µM xGen UDI-UMI adaptors (IDT), 30 µL NEBNext Ultra II Ligation Master Mix, and 1 µL of NEBNext Ligation Enhancer and then incubating for 15 min at 20°C. The xGen UDI-UMI adapters contain both unique dual indexes (UDIs) for identifying the sample and UMIs for identifying copies of the same molecule.

After ligation, a double-sided size selection was performed with SPRI beads (solid-phase reversible immobilization; made by washing 1 mL Sera-Mag SpeedBead carboxylate-modified [E3] magnetic particles [Cytiva] and resuspending the beads in 50 mL of 18% PEG-8000, 1.75 M NaCl, 10 mM Tris at pH 8, 1 mM EDTA, 0.044% Tween 20) (Abascal et al. 2021). For the first bead addition, a ratio of 0.4× bead volume to sample volume was used, and the supernatant was retained. For the second bead addition, a bead volume of 0.25× the original volume (before the first bead addition) was used, and the sample was eluted in 22 µL of 10 mM Tris (pH 8.0).

Prehybridization library amplification

Prehybridization low-temperature amplification

Prehybridization low-temperature amplification was performed in a 50 µL reaction volume containing 20 µL ligated library, 25 mM Tris (pH 7.5), 40 mM NaCl, 6 mM $MgCl_2$, 1 mM DTT, 2 µM P5 primer (for sequence, see the section Library Amplification Primers), and 1 mM dNTP mix. Before the reaction, Sequenase 2.0 DNA polymerase (Thermo Fisher Scientific) was diluted in a buffer containing 25 mM Tris (pH 7.5), 40 mM NaCl, and 1 mM DTT to a final polymerase concentration of 0.67 U/µL.

Libraries underwent initial denaturation on a thermal cycler for 30 sec at 98°C and then were left on the thermal cycler for the denaturation step of the first cycle for another 20 sec. Samples were then immediately transferred to a water bath for annealing for 20 sec at 34°C. Excess water was quickly blotted with a Kimwipe, and the samples were briefly spun down in a desktop centrifuge to remove any remaining water. Samples were then rapidly transferred to a thermal cycler set to 34°C for extension.

Immediately after transfer to the extension thermal cycler, 0.75 µL of diluted polymerase (0.5 U) was added to the reaction; the sample was pipette mixed; and the tube was recapped with new caps because the seal created by the caps weakens after heat exposure, which can allow water to enter the reaction during annealing in the water bath. Extension continued for 90 sec after the polymerase was added. In the third cycle, after the extension incubation was completed, 1 µL of 100 µM P7 primer (for

sequence, see section Library Amplification Primers) was added to the reaction and pipette-mixed to initiate the exponential phase of the reaction. After extension, the samples were transferred back to the 98°C thermal cycler to complete the cycle (Fig. 2A,B). The above cycling process was repeated for three linear cycles and two exponential cycles, for a total of five cycles. After the last cycle's extension step, samples were left on the 34°C thermal cycler for a final 5 min extension. Libraries were then purified with a 1× SPRI bead cleanup with two 80% ethanol washes and eluted in 22 µL of 10 mM Tris (pH 8.0). Final concentrations were measured with the Qubit 1× dsDNA HS assay kit (Thermo Fisher Scientific), and the libraries were profiled with a TapeStation HS D1000 ScreenTape assay (Agilent).

Prehybridization standard-temperature amplification

Prehybridization standard-temperature library amplification was performed in a 50 µL PCR reaction containing 20 µL ligated library, 1× NEBNext Ultra II Q5 Master Mix (NEB), and 2 µM P5 primer. Thermal cycling was conducted as follows: initial denaturation for 30 sec at 98°C, denaturation for 10 sec at 98°C, annealing and extension for 75 sec at 65°C, and final extension for 5 min at 65°C (Fig. 2B). P7 primer was added after the extension step of the third cycle as in the low-temperature amplification to start the exponential cycles. The cycling process was repeated for three linear cycles and two exponential cycles, as was performed for the low-temperature prehybridization amplification. The samples were purified, quantified, and profiled as described for the low-temperature amplification.

Library amplification primers

Library amplification primers were ordered as HPLC-purified oligonucleotides from Integrated DNA Technologies (IDT).

P5-PTx2: AATGATACGGCGACCACCGA*G*A
P7-PTx2: CAAGCAGAAGACGGCATACG*A*G
*Phosphorothioate bonds

Hybridization capture

Hybridization capture was performed with the Standard Hybridization Reagent Kit v2 (Twist Bioscience) and biotinylated 120 bp probes delivered at a concentration of 0.35 fmol/probe (Twist Bioscience) per the manufacturer's protocol. In pilot panel hybridizations, 375 ng of each library was input into a four-plex hybridization. In large-scale panel hybridizations, 350 ng (Coriell 1, Coriell 2, Coriell 3 trios) or 250 ng (Family 1100 and Family 2100 trios) of each library was input into the hybridizations. Each Coriell trio was hybridized in a separate three-plex hybridization, and Family 1100 and Family 2100 trios were multiplexed together in a single hybridization.

Posthybridization library amplification

Posthybridization low-temperature amplification

Posthybridization low-temperature amplification was performed in a 50 µL reaction volume containing 22.5 µL of the streptavidin bead slurry from hybridization, 1× NEBuffer 2 (NEB), 0.5 µM P5 primer, and 1 mM dNTP mix. Klenow fragment (3′→5′ exo-; NEB) was diluted in 1× NEBuffer 2 to a final concentration of 2 U/µL. Thermal cycling and polymerase addition were conducted as in the prehybridization low-temperature amplification. Note that posthybridization amplification utilized diluted Klenow fragment polymerase rather than the Sequenase 2.0 polymerase used in the prehybridization amplification. After the extension step of the third cycle, 1 µL of 25 µM P7 primer was added to the reaction and pipette-mixed to start the exponential phase of the reaction (Fig. 2A,B). The cycling process was repeated for three linear cycles and eight exponential cycles for the pilot panel, and three linear cycles and six exponential cycles for the large-scale panel. The samples were purified with a 1× DNA purification bead (Twist Bioscience) cleanup with two 80% ethanol washes and eluted in 25 µL of 10 mM Tris (pH 8.0). Because the 3′→5′ exo- Klenow fragment leaves a 3′ nontemplated nucleotide at the ends of molecules (Clark et al. 1987), we removed these in a 35 µL reaction containing the purified library, 1× NEBuffer 2, 0.033 mM dNTP mix, and 1 U of DNA Polymerase I, Large (Klenow) fragment (NEB), incubated for 15 min at 25°C. The reaction was stopped by adding EDTA to a final concentration of 10 mM and purified with a 1× DNA purification bead (Twist) cleanup with two 80% ethanol washes and eluted in 32 µL of 10 mM Tris (pH 8.0).

Posthybridization standard-temperature amplification

Posthybridization standard-temperature library amplification was performed in a 50 µL PCR reaction containing 22.5 µL of the streptavidin binding bead slurry from hybridization, 0.5 µM P5 primer, and 1× Equinox Library Amp Master Mix (Twist Bioscience). The thermal cycling protocol was as follows: initial denaturation for 45 sec at 98°C, denaturation for 15 sec at 98°C, annealing for 30 sec at 60°C, extension for 30 sec at 72°C, and final extension for 60 sec at 72°C (Fig. 2B). Note that this posthybridization standard-temperature amplification thermal cycling differs from the prehybridization standard-temperature amplification, because the former uses a PCR master mix made by NEB, whereas the latter uses a master mix that is part of the hybridization reagents kit made by Twist Bioscience. P7 primer was added after the extension step of the third cycle as in the low-temperature amplification. The cycling process was repeated for three linear cycles and eight exponential cycles for the pilot panel, as was performed for the low-temperature posthybridization amplification. The libraries were purified, quantified, and profiled as in the low-temperature posthybridization amplification.

Library preparation for PCR-free WGS

Genomic DNA of individual 1104 and an individual with CMMRD were processed to make Illumina sequencing libraries using the TruSeq DNA PCR-free kit (Illumina), and they were processed to make PCR-free Onso sequencing libraries using the Onso DNA library prep kit (Pacific Biosciences).

Sequencing

All Illumina sequencing was performed on NovaSeq 6000 instruments with 150 bp paired-end reads at the University of California Irvine Genomics High-Throughput Facility, the University of California Berkeley QB3 Genomics, and Psomagen. For Illumina sequencing metrics for all samples in the study, see Supplemental Tables S2 and S5. Onso sequencing was performed by Pacific Biosciences.

Note that 150 bp read lengths were used because the subset of reads that fully spans microsatellites that are used for genotyping will always have >20 bp extending to at least one flank for microsatellite loci that are <110 bp in length, which accounts for nearly all (97.2%) microsatellites annotated by STRATIFY in the genome (Supplemental Fig. S1A). This read length therefore maximizes the probability of being able to uniquely map the read to a specific microsatellite locus. Additionally, note that genotyping of microsatellites does not necessarily require fully spanning reads to align uniquely to both flanks of the microsatellite by >20 bp. In situations in which a read is uniquely aligned to only one flank of

the microsatellite, the read must then extend into the other flank by a sufficient number of bases to identify the length of the microsatellite, but this extension does not need to be >20 bp. Each of the three microsatellite genotyping tools we utilize in STREAM—HipSTR (Willems et al. 2017), GangSTR (Mousavi et al. 2019), and ExpansionHunter (Dolzhenko et al. 2019)—incorporates read filters that the authors of those tools optimized to keep reads that confidently align to the microsatellite with sufficient flanking bases.

Computational pipeline for microsatellite genotyping

Molecular duplication analysis

Molecular duplication rates were calculated for samples from trios for which UMIs were sequenced (Family 1100 and Family 2100). The computational pipeline (for schematic, see Supplemental Fig. S9A) converted FASTQ sequencing files to unmapped BAMs with the fgbio v2.0.2 (https://github.com/fulcrumgenomics/fgbio) FastqToBam tool with the option --extract-umis-from-read-names to extract UMI sequences. The unmapped BAM file was then reverted to FASTQ format with the SAMtools (v1.14) fastq command (Danecek et al. 2021) and then aligned to the reference genome with BWA-MEM v0.7.17 (Li 2013) using the options "-p -K 150000000 -Y". The mapped BAM and the unmapped BAM were input into fgbio ZipperBams to add the UMI metadata to the mapped BAM, and the mapped BAM was subsequently sorted by the query name with the SAMtools (v1.14) sort command using the option "-n". After sorting, optical duplicates were removed (because these should not be included in the UMI metrics as they are not true molecular duplicates) with the Picard v2.27.4 MarkDuplicates tool (https://broadinstitute.github.io/picard/) with the options "REMOVE_DUPLICATES = false OPTICAL_DUPLICATE_PIXEL_DISTANCE = 2500 ASSUME_SORT_ORDER = queryname CLEAR_DT = false REMOVE_SEQUENCING_DUPLICATES = true READ_NAME_REGEX = [a-zA-Z0-9]+:[0-9]+:[a-zA-Z0-9]+:[0-9]:\([0-9]+\):\([0-9]+\):\([0-9]+\).*". The BAM file was then unmarked with the UnmarkDuplicates tool of the Genome Analysis Toolkit (GATK) v4.2.4.1 (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). Finally, the number of UMIs observed for each molecular duplicate family size (where family size is the number of molecules observed with the same UMI sequence) was calculated using fgbio GroupReadsByUmi with the options "--strategy Adjacency --edits 1 --family-size-histogram".

STREAM ensemble microsatellite genotyping pipeline

STREAM, our ensemble microsatellite genotyping pipeline, is run via the Nextflow v21.10.6 scientific workflow tool (for schematic, see Supplemental Fig. S9C,D; Di Tommaso et al. 2017). Reads were aligned to the human reference genome (hg38) with BWA-MEM v0.7.17 (Li 2013). Optical duplicates were removed with Picard MarkDuplicates with the options "REMOVE_DUPLICATES = false OPTICAL_DUPLICATE_PIXEL_DISTANCE = 2500 ASSUME_SORT_ORDER = queryname CLEAR_DT = false REMOVE_SEQUENCING_DUPLICATES = true READ_NAME_REGEX = [a-zA-Z0-9]+:[0-9]+:[a-zA-Z0-9]+:[0-9]:\([0-9]+\):\([0-9]+\):\([0-9]+\)*" (note that the read name regular expression was included to handle files with UMI sequences). The fraction of optical duplication was calculated with the equation (READ_PAIR_ OPTICAL_DUPLICATES × 2)/(UNPAIRED_READS_EXAMINED + READ_PAIRS_EXAMINED × 2), based on MarkDuplicates documentation. The resulting BAM file was unmarked using GATK UnmarkDuplicates (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013) and sorted with Picard SortSam with the parameter "SORT_ORDER = coordinate", followed by conversion to CRAM format with the SAMtools (v1.14) view command with the options "-F 2304 -C" (filters out nonprimary and supplemental alignments). Hybridization capture quality metrics, including percentage of on-target reads, percentage of zero-coverage loci, AT-dropout, and fold 80 base penalty, were calculated using Picard CollectHsMetrics.

The aligned reads were then input into three different microsatellite callers: HipSTR v0.7 (Willems et al. 2017), GangSTR v2.5.0 (Mousavi et al. 2019), and ExpansionHunter v5.0.0 (Dolzhenko et al. 2019). HipSTR was run on all samples jointly, because this allows generation of more accurate stutter models that improve accuracy (Willems et al. 2017), with the parameters "--min-reads 20 --max-str-len 150 --no-rmdup --output-filters", and the output VCF was split by sample using BCFtools (Danecek et al. 2021). For the temperature-comparison assay, HipSTR was also run with the option "--haploid-chrs chrX,chrY" because all of the replicates being profiled were male. GangSTR and Expansion Hunter do not support joint genotyping and were run on each sample individually. GangSTR was run with the parameters "--min-sample reads 20 --nonuniform --frrweight 0 --spanweight 0 --flankweight 0", and ExpansionHunter was run with default parameters. For both GangSTR and ExpansionHunter, each sample's sex was specified. Because ExpansionHunter does not output total read depth for loci, which is required for downstream analysis of ExpansionHunter data, the BEDTools coverage tool (Quinlan and Hall 2010) was used with the parameters "-sorted -f 1.0" to calculate the number of mapped reads that completely span the coordinates of the microsatellite loci (spanning reads). Relevant fields were then extracted from each caller's VCF output and converted to a tab-delimited format with BCFtools (Danecek et al. 2021).

Genotype information from each caller was further processed with an R (R Core Team 2023) script (requiring the packages dplyr [https://dplyr.tidyverse.org/] and tidyr [https://tidyr.tidyverse.org/]) as follows:

1. The capture panel information, sample information, and data from HipSTR, GangSTR, ExpansionHunter, and BEDTools coverage analyses are loaded and joined into a single table.
2. ExpansionHunter calls for which the BEDTools spanning read coverage is zero are excluded from further analysis.
3. HipSTR diploid calls for Chromosome X loci in male samples are converted to haploid calls by keeping only the allele with the most supporting reads per the MALLREADS field.
4. For each caller, we calculate the difference in repeat units of the called genotype from the reference genome.
5. For HipSTR only, we calculate the fraction of total reads (DP) with stutter (DSTUTTER) and the fraction of total reads with flank indels (DFLANKINDEL).
6. For HipSTR, we sum the reads reported in the MALLREADS field to obtain the total number of reads used to call the genotype at each locus.
7. For GangSTR, we sum the reads reported in the ENCLREADS field to obtain the total number of spanning reads reported by GangSTR.
8. The VAF of each allele is calculated by dividing the number of reads supporting that allele by the total number of reads at the locus, as follows:
    a. For HipSTR, read classes are not reported separately, so we use the number of reads in the MALLREADS field that support the called allele as the numerator and the total number of reads reported in MALLREADS as the denominator.
    b. For GangSTR, we use the number of reads in the ENCLREADS field that support the called allele as the numerator and the total number of reads reported in ENCLREADS as the

denominator. We use ENCLREADS because this reports only the spanning reads.

c. For ExpansionHunter, we use the number of reads in the ADSP field that support the called allele as the numerator. We use ADSP because this field reports only the spanning reads identified by ExpansionHunter. ExpansionHunter does not report a total read depth, so we use the number of spanning reads counted by BEDTools coverage as the denominator (see above), which results in some VAF values greater than one because the BEDTools coverage read count is not exactly equivalent to the reads counted by ExpansionHunter.

9. Data from all samples are joined into a single table.

At this stage, each locus has three genotype calls, one from each caller. For each locus, we need to choose one caller whose genotypes will be used as the final genotypes for all samples. We want to use only one caller for all samples because the reported genotype can differ between callers, precluding concordance analysis between samples if genotypes from different callers are used. The first step in choosing a caller is to filter the reported genotypes based on a variety of quality parameters (Supplemental Figs. S9D, S11). The pipeline uses a YAML configuration file to store the values for each filtering threshold and an order of preference for which caller's genotypes to use. First, sample-level filters are applied to each caller's genotype calls for each sample. The sample-level filters common to all callers are (1) minimum total reads (MALLREADS field for HipSTR, ENCLREADS for GangSTR, and BEDTools spanning read count for ExpansionHunter), (2) minimum number of reads supporting each allele (same fields as total reads except for ExpansionHunter, which uses the ADSP field), and (3) minimum VAF. ExpansionHunter has both a minimum and maximum VAF threshold owing to the previously explained issue with the VAF denominator. HipSTR also has additional sample-level filters: (1) minimum quality score, (2) maximum fraction of reads containing stutter, (3) maximum fraction of reads containing flank indels, and (4) minimum likelihood difference between the reported genotype and the next best genotype (GLDIFF). GangSTR also has an additional sample-level filter of minimum quality score. ExpansionHunter also has an additional sample-level filter of minimum average locus coverage. For many filters, both alleles are required to pass the filters, so the pipeline accounts for the sex of the sample when applying filters to sex chromosomes.

Next, the pipeline applies locus-level filters to check for systemic genotyping issues across multiple samples for each locus. The locus-level filters used for this step are (1) minimum mean quality score across samples (HipSTR and GangSTR only), (2) minimum mean total read depth across samples (using the same fields as in the sample-level filters), and (3) minimum fraction of samples passing the sample-level filters.

After the sample-level and locus-level filters have been applied, the final genotype call is chosen from one of the three callers. The caller used for each locus is chosen based on the caller order of preference specified in the YAML configuration file and whether the locus has passed the locus-level filters for that caller. If the locus does not pass locus-level filters for any caller, no call is reported. Then, at each locus, the pipeline checks if the sample passed the sample-level filters. If the sample passed filtering, the genotype from the caller chosen for that locus is entered into the final genotype column; if the sample did not pass filtering at that locus; no genotype is entered.

For this analysis, we chose the combination of filtering parameters and optimized their thresholds (for details of all parameters and thresholds, see below) (Supplemental Figs. S11, S12) by testing a range of parameter combinations and thresholds and calculating the Mendelian discordance rate of the resulting genotypes, that is, the fraction of loci at which the genotypes of the parents and offspring do not match patterns of Mendelian inheritance. If the genotype of the offspring does match Mendelian inheritance patterns, that locus is considered concordant. A true discordant locus indicates the presence of a de novo mutation in the offspring, so the rate of discordance should match the microsatellite de novo mutation rate, which prior studies have estimated to be about 1×10^{-4} mutations per locus per generation (Weber and Wong 1993; Huang et al. 2002; Sun et al. 2012; Kristmundsdottir et al. 2023). In addition to optimizing our filters to achieve this discordance rate, we also optimized them to maximize the number of loci that were genotyped (i.e., that passed filters). Optimization was performed using Family 1100 and confirmed on the other trios. Mendelian discordance calculations were only performed on loci that were genotyped in all three members of the family trio ("fully genotyped loci"). For trios with a male child, loci on Chromosome X were considered fully genotyped if only the mother and son were genotyped because the father did not pass on an X Chromosome.

We implemented different filtering parameters and thresholds for 2–4 bp motif loci and poly(A) loci owing to the latter's higher stutter levels. For 2–4 bp motif loci, we used a caller preference order of (1) HipSTR, (2) GangSTR, and (3) ExpansionHunter. For HipSTR, we used the following filter parameters: (1) VAF ≥ 0.2, (2) fraction of stutter reads ≤ 0.13, (3) fraction of reads with flank indels ≤ 0.13, (4) genotype likelihood difference of 25 or more, and (5) mean quality score ≥ 0.95. For GangSTR, we used the following filter parameters: (1) number of reads supporting the allele of 80 or more for both alleles and (2) mean quality score ≥ 1.0. For ExpansionHunter, we used the following filter parameters: (1) $0.5 \leq$ VAF ≤ 2.5 and (2) number of reads supporting the allele of 60 or more for both alleles.

We separately optimized filtering parameters and thresholds for poly(A) loci owing to their overall lower-quality genotype calls and higher stutter levels. We used slightly different filters depending on whether the trio had a male or a female child because we observed that trios with female children tended to have higher discordance rates, likely because genotyping of heterozygous Chromosome X loci is more challenging than genotyping hemizygous Chromosome X loci. For all trios regardless of the sex of the child, we used a caller preference order of (1) GangSTR and (2) HipSTR. We did not use ExpansionHunter for poly(A) loci calls because we could not find a combination of filters that achieved the expected Mendelian discordance rate. For trios with male children, we used the following filter parameters for GangSTR: (1) number of reads supporting each allele of ten or more and (2) mean quality score ≥ 0.5. For HipSTR, we used the following filter parameters: (1) VAF ≥ 0.2, (2) fraction of reads with flank indels ≤ 0.13, (3) genotype likelihood difference of one or more, and (4) mean quality score ≥ 0.85. For trios with female children, we adjusted the GangSTR filters by increasing the number of reads supporting each allele to 15 or more and adding a filter for VAF ≥ 0.4, and the HipSTR VAF filter was changed to ≥ 0.3.

Hybridization capture subsampling analysis

We performed an analysis to determine how average read depth per locus affects the results of our hybridization capture genotyping. Specifically, we compared the number of fully genotyped loci and the Mendelian discordance rate at different subsampled average read depths per locus, in which average read depth per locus is the total number of aligned paired-end sequencing reads divided by the number of loci in the panel. This subsampling analysis was performed on the large-scale panel data of Family

1100 because, across all trios, its lowest coverage sample had the highest number of aligned reads. We subsampled FASTQ files for each member of the trio using seqtk sample v1.3 (Li 2012) with the options "-2 -s100" to 30, 60, 120, 200, 400, 600, 800, 900, and 1013 average reads per locus. The highest subsampling level of 1013 average reads per locus corresponds to the original sample size of individual 1101, which is the sample in the trio with the lowest average reads per locus. For this subsampling level, we subsampled the other samples in the trio to the number of reads in the sample of individual 1101. For lower coverage levels, we multiplied the desired average reads per locus by the number of loci and subsampled to that number of reads. Each set of subsampled FASTQ files was processed with our STREAM pipeline along with the other four trios, whose sequencing levels had not been adjusted; we ran all samples together for consistency of HipSTR joint genotyping and genotype filtering. Finally, the Mendelian discordance rate and the number of fully genotyped loci were calculated for each subsampling level.

WGS subsampling analysis

We performed a subsampling analysis of high-depth PCR-free Illumina WGS on a family trio (HG002 = child, HG003 = father, HG004 = mother) from the NIST Genome in a Bottle (GIAB) project. We downloaded the FASTQ data from the NIST GIAB (https://github.com/genome-in-a-bottle/giab_data_indexes). The reads were subsampled as in the hybridization capture subsampling analysis at 30×, 60×, 100×, 150×, 200×, 250×, and 309× average coverage. The highest subsampling level is the average coverage of sample HG002, which is the sample with the lowest average coverage in the trio.

Because these are WGS data, we used STRATIFY to create a genome-wide list of microsatellites for analysis, and because we no longer needed to consider capture probes in our filtering, we used the following minimal filters: (1) distance to the nearest microsatellite ≥75 bp in both flanks, (2) 100 *k*-mer mappability score in both 90 bp flanks ≥0.95, (3) exclusion of microsatellites for which at least one 90 bp flank overlaps either the 1000 Genomes Project pilot accessibility mask (The 1000 Genomes Project Consortium 2015) or the ENCODE Data Analysis Center Blacklist Regions (Amemiya et al. 2019), and (4) exclusion of microsatellites for which at least one 90 bp flank and/or the microsatellite itself overlaps annotated segmental duplications (Bailey et al. 2002) or pseudoautosomal regions. We further required for each microsatellite locus that the mappability filters, region filters, and distance to the nearest microsatellite filter all pass in the same flank (for at least one flank). It was found that 563,770 loci passed these filters, 395,725 of which have 2–6 bp motifs and 168,045 of which have a poly(A) motif.

We processed the high-depth sequencing data with stage 1 of STREAM (Supplemental Fig. S9C) using either the loci from the large-scale panel or the loci from the genome-wide list described above. We then reoptimized the quality filtering parameters for the 309× WGS data to obtain the expected Mendelian discordance rate for the large-scale panel. The filtering thresholds used were identical to the thresholds used for the hybridization capture large-scale panel analysis described above, except for the following filters for 2–6 bp motifs: (1) fraction of stutter reads ≤0.1, because we observed slightly lower levels of stutter in the PCR-free data, and (2) turning off the filter for fraction of reads with flank indels, because we observed only a handful of loci above this threshold. The filtering thresholds used for poly(A) loci were identical to those used for the hybridization capture large-scale panel analysis for trios with male children described above. Lower subsampling levels were then analyzed with STREAM using these filtering thresholds. The same filtering thresholds were used to analyze the WGS data both for the large-scale panel and the genome-wide list of loci. Finally, the Mendelian discordance rate and the number of fully genotyped loci were calculated for each subsampling level and for each set of loci.

EnsembleTR analysis

EnsembleTR (Ziaei Jam et al. 2023) was performed on high-depth PCR-free Illumina WGS using VCF input files generated by HipSTR, GangSTR, and Expansion Hunter. These VCF files were generated by STREAM during analysis of the 100× subsampled WGS data with the genome-wide list of loci described previously. EnsembleTR does not process haploid genotypes, so "chrX" loci were removed from the VCFs using BCFtools (Danecek et al. 2021). Individual sample VCFs for ExpansionHunter and GangSTR were merged for each caller using mergeSTR v6.0.1 (Mousavi et al. 2021), whereas HipSTR was already run jointly on all samples in STREAM and did not require VCF merging. The merged GangSTR VCF was filtered using dumpSTR v6.0.1 (Mousavi et al. 2021) with the options "--gangstr-filter-span bound-only --gangstr-filter-badCI". Subsequently, we ran EnsembleTR v1.0.0 on the joint VCFs of all three callers. EnsembleTR's VCF output was sorted, and relevant fields were converted to a tab-delimited format with BCFtools (Danecek et al. 2021). The VCF output contained two records for 38,774 of the loci (7.2% of the loci), one using the HipSTR genotypes and another using genotypes from Expansion Hunter or GangSTR, owing to differences in the repeat motifs called by each tool. We excluded these loci from downstream analyses.

Tools for data analysis and plots

We used the following data analysis tools and packages to analyze and plot our data: R packages—dplyr (https://dplyr.tidyverse.org/), tidyr (https://tidyr.tidyverse.org/), ggplot2 (Wickham 2016), gghighlight (https://yutannihilation.github.io/gghighlight/), ggbeeswarm (https://github.com/eclarke/ggbeeswarm), ggforce (https://ggforce.data-imaginist.com/), and wesanderson (https://github.com/karthik/wesanderson); other analysis tools—BED Tools (Quinlan and Hall 2010), UCSCtools bedGraphToBigWig (Kent et al. 2010), and BioRender (https:// www.biorender.com).

Capillary electrophoresis validation

We used capillary electrophoresis to validate discordant genotype calls in three family trios: Coriell 3, Family 1100, and Family 2100. For each locus called as discordant by STREAM, we designed a primer pair with one primer marked with fluorescein (6-FAM) at its 5′ end (for primer sequences, see Supplemental Table S6). PCR was performed in a 20 µL reaction containing 1× Colorless GoTaq Flexi buffer (Promega), 1.5 mM $MgCl_2$, 0.2 mM dNTP mix, 0.25 uM each primer, and 0.5 U GoTaq G2 Flexi DNA polymerase (Promega). Thermal cycling was performed for 40 cycles with the following protocol: initial denaturation for 2 min at 95°C, denaturation for 30 sec at 95°C, annealing for 30 sec at 58°C, extension for 45 sec at 72°C, and final extension for 5 min at 72°C. Capillary electrophoresis was performed by the University of Arizona Genetics Core. The analysis software Peak Scanner (Thermo Fisher Scientific Connect Platform) was used to determine the amplicon length for each locus and sample. Inferred repeat unit counts from the fragment analysis assay were then compared to the genotypes output by STREAM.

Software availability

The source code for STRATIFY is available at GitHub (https://github.com/evronylab/STRATIFY). The source code for STREAM is available at GitHub (https://github.com/evronylab/STREAM). The source codes for these tools are also available as Supplemental Code (note that some large required files are only available in the above GitHub links).

Data access

All raw sequencing data generated in this study have been submitted to the NCBI database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) under accession number phs003679.

Competing interest statement

Sequencing of the two Onso libraries was performed by Pacific Biosciences. G.D.E. owns equity in Illumina and Pacific Biosciences. All other authors declare no competing interests.

Acknowledgments

This work was supported by grants from the Sontag Foundation, the McKnight Endowment Fund for Neuroscience, and the National Institutes of Health Common Fund (DP5OD028158).

Author contributions: C.A.L., D.A.S., and G.D.E. conceived the technology. C.A.L. designed the experiments with input from G.D.E. C.A.L. performed the experiments. D.A.S. developed STRATIFY. C.A.L. developed STREAM and analyzed the data with input from D.A.S., A.S., and G.D.E. C.A.L. wrote the manuscript with input from D.A.S. and G.D.E.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, Russell AJC, Alcantara RE, Baez-Ortega A, Wang Y, et al. 2021. Somatic mutation landscapes at single-molecule resolution. *Nature* **593:** 405–410. doi:10.1038/s41586-021-03477-4

Acquaviva C, Duval M, Mirebeau D, Bertin R, Cavé H. 2003. Quantitative analysis of chimerism after allogeneic stem cell transplantation by PCR amplification of microsatellite markers and capillary electrophoresis with fluorescence detection: the Paris–Robert Debré experience. *Leukemia* **17:** 241–246. doi:10.1038/sj.leu.2402762

Ahrberg CD, Manz A, Chung BG. 2016. Polymerase chain reaction in microfluidic devices. *Lab Chip* **16:** 3866–3884. doi:10.1039/C6LC00984K

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep* **9:** 9354. doi:10.1038/s41598-019-45839-z

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003–1007. doi:10.1126/science.1072047

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573–580. doi:10.1093/nar/27.2.573

Bhargava A, Fuentes FF. 2010. Mutational dynamics of microsatellites. *Mol Biotechnol* **44:** 250–266. doi:10.1007/s12033-009-9230-4

Bilinski P, Han Y, Hufford MB, Lorant A, Zhang P, Estep MC, Jiang J, Ross-Ibarra J. 2017. Genomic abundance is not predictive of tandem repeat localization in grass genomes. *PLoS One* **12:** e0177896. doi:10.1371/journal.pone.0177896

Boby T, Patch AM, Aves SJ. 2005. TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics* **21:** 811–816. doi:10.1093/bioinformatics/bti059

Bruford MW, Wayne RK. 1993. Microsatellites and their application to population genetic studies. *Curr Opin Genet Dev* **3:** 939–943. doi:10.1016/0959-437X(93)90017-J

Butler JM. 2006. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci* **51:** 253–265. doi:10.1111/j.1556-4029.2006.00046.x

Campbell NR, Harmon SA, Narum SR. 2015. Genotyping-in-thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Mol Ecol Resour* **15:** 855–867. doi:10.1111/1755-0998.12357

Chung J, Negm L, Bianchi V, Stengs L, Das A, Liu ZA, Sudhaman S, Aronson M, Brunga L, Edwards M, et al. 2023. Genomic microsatellite signatures identify germline mismatch repair deficiency and risk of cancer onset. *J Clin Oncol* **41:** 766–777. doi:10.1200/JCO.21.02873

Clark JM, Joyce CM, Beardsley GP. 1987. Novel blunt-end addition reactions catalyzed by DNA polymerase I of *Escherichia coli*. *J Mol Biol* **198:** 123–127. doi:10.1016/0022-2836(87)90462-1

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

Daunay A, Duval A, Baudrin LG, Buhard O, Renault V, Deleuze J-F, How-Kit A. 2019. Low temperature isothermal amplification of microsatellites drastically reduces stutter artifact formation and improves microsatellite instability detection in cancer. *Nucleic Acids Res* **47:** e141. doi:10.1093/nar/gkz811

De Barba M, Miquel C, Lobréaux S, Quenette PY, Swenson JE, Taberlet P. 2017. High-throughput microsatellite genotyping in ecology: improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Mol Ecol Resour* **17:** 492–507. doi:10.1111/1755-0998.12594

De Grassi A, Ciccarelli FD. 2009. Tandem repeats modify the structure of human genes hosted in segmental duplications. *Genome Biol* **10:** R137. doi:10.1186/gb-2009-10-12-r137

Denœud F, Vergnaud G. 2004. Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource. *BMC Bioinformatics* **5:** 4. doi:10.1186/1471-2105-5-4

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498. doi:10.1038/ng.806

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35:** 316–319. doi:10.1038/nbt.3820

Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, Emig-Agius D, Gross A, Narzisi G, Bowman B, et al. 2019. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35:** 4754–4756. doi:10.1093/bioinformatics/btz431

Domaniç NO, Preparata FP. 2007. A novel approach to the detection of genomic approximate tandem repeats in the Levenshtein metric. *J Comput Biol* **14:** 873–891. doi:10.1089/cmb.2007.0018

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5:** 435–445. doi:10.1038/nrg1348

Evrony GD, Lee E, Mehta Bhaven K, Benjamini Y, Johnson Robert M, Cai X, Yang L, Haseley P, Lehmann Hillel S, Park Peter J, et al. 2015. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85:** 49–59. doi:10.1016/j.neuron.2014.12.028

Fang H, Wu Y, Narzisi G, Orawe JA, Barrón LTJ, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, Lyon GJ. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med* **6:** 89. doi:10.1186/s13073-014-0089-z

Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. 2005. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput Biol* **1:** e50. doi:10.1371/journal.pcbi.0010050

Fungtammasan A, Ananda G, Hile SE, Su MS, Sun C, Harris R, Medvedev P, Eckert K, Makova KD. 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res* **25:** 736–749. doi:10.1101/gr.185892.114

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498–511. doi:10.1038/nature01097

Gärke C, Ytournel F, Bed'hom B, Gut I, Lathrop M, Weigend S, Simianer H. 2012. Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Anim Genet* **43:** 419–428. doi:10.1111/j.1365-2052.2011.02284.x

Gonzalez-Pena V, Natarajan S, Xia Y, Klein D, Carter R, Pang Y, Shaner B, Annu K, Putnam D, Chen W, et al. 2021. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc Natl Acad Sci* **118:** e2024176118. doi:10.1073/pnas.2024176118

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, et al. 2011. Current trends

in microsatellite genotyping. *Mol Ecol Resour* **11:** 591–611. doi:10.1111/j.1755-0998.2011.03014.x

Gupta R, Sarthi D, Mittal A, Singh K. 2007. A novel signal processing measure to identify exact and inexact tandem repeat patterns in DNA sequences. *EURASIP J Bioinform Syst Biol* **2007:** 43596. doi:10.1155/2007/43596

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res* **22:** 1154–1162. doi:10.1101/gr.135780.111

Gymrek M, Willems T, Reich D, Erlich Y. 2017. Interpreting short tandem repeat variations in humans using mutational constraint. *Nat Genet* **49:** 1495–1501. doi:10.1038/ng.3952

Halman A, Oshlack A. 2020. Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. *F1000Res* **9:** 200. doi:10.12688/f1000research.22639.1

Hill CR, Butler JM, Vallone PM. 2009. A 26plex autosomal STR assay to aid human identity testing. *J Forensic Sci* **54:** 1008–1015. doi:10.1111/j.1556-4029.2009.01110.x

Hite JM, Eckert KA, Cheng KC. 1996. Factors affecting fidelity of DNA synthesis during PCR amplification of d(C-A)n•d(G-T)n microsatellite repeats. *Nucleic Acids Res* **24:** 2429–2434. doi:10.1093/nar/24.12.2429

Huang Q-Y, Xu F-H, Shen H, Deng H-Y, Liu Y-J, Liu Y-Z, Li J-L, Recker RR, Deng H-W. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. *Am J Hum Genet* **70:** 625–634. doi:10.1086/338997

Hutter B, Helms V, Paulsen M. 2006. Tandem repeats in the CpG islands of imprinted genes. *Genomics* **88:** 323–332. doi:10.1016/j.ygeno.2006.03.019

Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. 2018. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* **46:** e120. doi:10.1093/nar/gky677

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26:** 2204–2207. doi:10.1093/bioinformatics/btq351

Kessler MD, Loesch DP, Perry JA, Heard-Costa NL, Taliun D, Cade BE, Wang H, Daya M, Ziniti J, Datta S, et al. 2020. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc Natl Acad Sci* **117:** 2560–2569. doi:10.1073/pnas.1902766117

Kristmundsdottir S, Jonsson H, Hardarson MT, Palsson G, Beyter D, Eggertsson HP, Gylfason A, Sveinbjornsson G, Holley G, Stefansson OA, et al. 2023. Sequence variants affecting the genome-wide rate of germline microsatellite mutations. *Nat Commun* **14:** 3855. doi:10.1038/s41467-023-39547-6

Kruglyak S, Durrett RT, Schug MD, Aquadro CF. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci* **95:** 10774–10778. doi:10.1073/pnas.95.18.10774

Lai Y, Sun F. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol Biol Evol* **20:** 2123–2131. doi:10.1093/molbev/msg228

Leclercq S, Rivals E, Jarne P. 2007. Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics* **8:** 125. doi:10.1186/1471-2105-8-125

Li H. 2012. seqtk toolkit for processing sequences in FASTA/Q formats. *GitHub* **767:** 69. https://github.com/lh3/seqtk

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN].

Lower SS, McGurk MP, Clark AG, Barbash DA. 2018. Satellite DNA evolution: old ideas, new approaches. *Curr Opin Genet Dev* **49:** 70–78. doi:10.1016/j.gde.2018.03.003

Madsen BE, Villesen P, Wiuf C. 2008. Short tandem repeats in human exons: a target for disease mutations. *BMC Genomics* **9:** 410. doi:10.1186/1471-2164-9-410

Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. 2011. What can exome sequencing do for you? *J Med Genet* **48:** 580–589. doi:10.1136/jmedgenet-2011-100223

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7:** 111–118. doi:10.1038/nmeth.1419

Marchal C, Sasaki T, Vera D, Wilson K, Sima J, Rivera-Mulia JC, Trevilla-García C, Nogues C, Nafie E, Gilbert DM. 2018. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repliseq. *Nat Protoc* **13:** 819–839. doi:10.1038/nprot.2017.148

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303. doi:10.1101/gr.107524.110

Mizuta S, Munakata H, Aimaiti A, Oosawa K, Shimizu T. 2004. Evaluation of the color-coding method for searching tandem repeats in prokaryotic genomes. *Chem-Bio Inf J* **4:** 133–141. doi:10.1273/cbij.4.133

Moretti TR, Moreno LI, Smerick JB, Pignone ML, Hizon R, Buckleton JS, Bright J-A, Onorato AJ. 2016. Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. *Forensic Sci Int Genet* **25:** 175–181. doi:10.1016/j.fsigen.2016.07.022

Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. 2019. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res* **47:** e90. doi:10.1093/nar/gkz501

Mousavi N, Margoliash J, Pusarla N, Saini S, Yanicky R, Gymrek M. 2021. TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **37:** 731–733. doi:10.1093/bioinformatics/btaa736

Oketch JW, Wain LV, Hollox EJ. 2024. A comparison of software for analysis of rare and common short tandem repeat (STR) variation using human genome sequences from clinical and population-based samples. *PLoS One* **19:** e0300545. doi:10.1371/journal.pone.0300545

Putman AI, Carbone I. 2014. Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecol Evol* **4:** 4399–4428. doi:10.1002/ece3.1305

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Raz O, Biezuner T, Spiro A, Amir S, Milo L, Titelman A, Onn A, Chapal-Ilani N, Tao L, Marx T, et al. 2019. Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. *Nucleic Acids Res* **47:** 2436–2445. doi:10.1093/nar/gky1318

R Core Team. 2023. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Reizel Y, Itzkovitz S, Adar R, Elbaz J, Jinich A, Chapal-Ilani N, Maruvka YE, Nevo N, Marx Z, Horovitz I, et al. 2012. Cell lineage analysis of the mammalian female germline. *PLoS Genet* **8:** e1002477. doi:10.1371/journal.pgen.1002477

Ryba T, Battaglia D, Pope BD, Hiratani I, Gilbert DM. 2011. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nat Protoc* **6:** 870–895. doi:10.1038/nprot.2011.328

Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, et al. 2015. Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. *Hum Mutat* **36:** 903–914. doi:10.1002/humu.22825

Schlötterer C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109:** 365–371. doi:10.1007/s004120000089

Selkoe KA, Toonen RJ. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecol Lett* **9:** 615–629. doi:10.1111/j.1461-0248.2006.00889.x

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet* **44:** 1161–1165. doi:10.1038/ng.2398

Tao L, Raz O, Marx Z, Ghosh MS, Huber S, Greindl-Junghans J, Biezuner T, Amir S, Milo L, Adar R, et al. 2021. Retrospective cell lineage reconstruction in humans by using short tandem repeats. *Cell Rep Methods* **1:** None. doi:10.1016/j.crmeth.2021.100054

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13:** 36–46. doi:10.1038/nrg3117

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43:** 11.10.11–11.10.33. doi:10.1002/0471250953.bi1110s43

Vemireddy LR, Archak S, Nagaraju J. 2007. Capillary electrophoresis is essential for microsatellite marker based detection and quantification of adulteration of basmati rice (*Oryza sativa*). *J Agric Food Chem* **55:** 8112–8117. doi:10.1021/jf0714517

Wang Z, Moffitt AB, Andrews P, Wigler M, Levy D. 2022. Accurate measurement of microsatellite length by disrupting its tandem repeat structure. *Nucleic Acids Res* **50:** e116. doi:10.1093/nar/gkac723

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Hum Mol Genet* **2:** 1123–1128. doi:10.1093/hmg/2.8.1123

Wei CJ, Zhang K. 2020. RETrace: simultaneous retrospective lineage tracing and methylation profiling of single cells. *Genome Res* **30:** 602–610. doi:10.1101/gr.255851.119

Wenz H-M, Robertson JM, Menchen S, Oaks F, Demorest DM, Scheibler D, Rosenblum BB, Wike C, Gilbert DA, Efcavitch JW. 1998. High-precision genotyping by denaturing capillary electrophoresis. *Genome Res* **8:** 69–80. doi:10.1101/gr.8.1.69

Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. Springer-Verlag, New York.

Willems T, Gymrek M, Highnam G, Consortium TGP, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res* **24:** 1894–1904. doi:10.1101/gr.177774.114

Willems T, Gymrek M, Poznik GD, Tyler-Smith C, Erlich Y. 2016. Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *Am J Hum Genet* **98:** 919–933. doi:10.1016/j.ajhg.2016.04.001

Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. 2017. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* **14:** 590–592. doi:10.1038/nmeth.4267

Yamanoi E, Sakurada M, Ueno Y. 2021. Low stutter ratio by SuperFi polymerase. *Forensic Sci Int Rep* **3:** 100201. doi:10.1016/j.fsir.2021.100201

Zhou J, Zhang M, Li X, Wang Z, Pan D, Shi Y. 2021. Performance comparison of four types of target enrichment baits for exome DNA sequencing. *Hereditas* **158:** 10. doi:10.1186/s41065-021-00171-3

Ziaei Jam H, Li Y, DeVito R, Mousavi N, Ma N, Lujumba I, Adam Y, Maksimov M, Huang B, Dolzhenko E, et al. 2023. A deep population reference panel of tandem repeat variation. *Nat Commun* **14:** 6711. doi:10.1038/s41467-023-42278-3

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025. doi:10.1038/sdata.2016.25

Received November 28, 2023; accepted in revised form July 11, 2024.