Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press



Accurate typing of short tandem repeats from genome-wide sequencing data and its applications

Arkarachai Fungtammasan, Guruprasad Ananda, Suzanne Hile, et al.

Genome Res. published online March 30, 2015
Access the most recent version at doi:10.1101/gr.185892.114

P<P Published online March 30, 2015 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**



To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

ACCURATE TYPING OF SHORT TANDEM REPEATS FROM GENOME-WIDE SEQUENCING DATA AND ITS APPLICATIONS

Arkarachai Fungtammasan[1,2,3,7*], Guruprasad Ananda[1,3,4,7*#], Suzanne E Hile[5], Marcia Shu-Wei Su[2], Chen Sun[6], Robert Harris[2], Paul Medvedev[3,4,6,7$], Kristin Eckert[3,5$], Kateryna D. Makova[2,3,7$]

[1] Integrative Biosciences, Bioinformatics and Genomics Option, Pennsylvania State University, Pennsylvania, 16802 USA

[2] Department of Biology, Pennsylvania State University, Pennsylvania, 16802 USA

[3] Center for Medical Genomics, Pennsylvania State University, University Park, Pennsylvania 16802, USA

[4] Department of Biochemistry and Molecular Biology, Pennsylvania State University, Pennsylvania, 16802 USA

[5] Department of Pathology, The Jake Gittlen Laboratories for Cancer Research, The Pennsylvania State University College of Medicine, Hershey, Pennsylvania 17033

[6] Department of Computer Science and Engineering, Pennsylvania State University, Pennsylvania, 16802 USA

[7] The Genome Science Institute at the Huck Institutes of Life Sciences, Pennsylvania State University, Pennsylvania, 16802 USA

#Current address: The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06030, USA

$Corresponding authors

***These authors contributed equally**

Running title: Tandem Repeat Sequencing Error Profiles and Error Correction Models

Key words: Microsatellites, Tandem Repeats, Sequencing Errors, Illumina, PCR free, Error Correction Model, Minimum Depth, Genotyping

Supplemental material: 18 Supplemental Tables and 17 Supplemental Figures

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

ABSTRACT

Short Tandem Repeats (STRs) are implicated in dozens of human genetic diseases and contribute significantly to genome variation and instability. Yet profiling STRs from short-read sequencing data is challenging because of their high sequencing error rates. Here we developed STR-FM, Short Tandem Repeat profiling using Flank-based Mapping, a computational pipeline that can detect the full spectrum of STR alleles from short-read data, can adapt to emerging read-mapping algorithms, and can be applied to heterogeneous genetic samples (e.g., tumors, viruses, and genomes of organelles). We used STR-FM to study STR error rates and patterns in publicly available human, and in-house generated ultra-deep plasmid, sequencing datasets. We discovered that STRs sequenced with a PCR-free protocol have up to 9-fold fewer errors than those sequenced with a PCR-containing protocol. We constructed an error correction model for genotyping STRs that can distinguish heterozygous alleles containing STRs with consecutive repeat numbers. Applying our model and pipeline to Illumina sequencing data with 100-bp reads, we could confidently genotype several disease-related long trinucleotide STRs. Utilizing this pipeline, for the first time we determined the genome-wide STR germ-line mutation rate from a deeply sequenced human pedigree. Additionally, we built a tool that recommends minimal sequencing depth for accurate STR genotyping, depending on repeat length and sequencing read length. The required read depth increases with STR length and is lower for a PCR-free protocol. This suite of tools addresses the pressing challenges surrounding STR genotyping, and thus is of wide interest to researchers investigating disease-related STRs and STR evolution.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

INTRODUCTION

Short tandem repeats (STRs) of 1-6 base pairs per motif constitute approximately 3% of the human genome (Lander 2001). Due to the high incidence of polymerase slippage at STRs (Levinson and Gutman 1987; Abdulovic et al. 2011; Baptiste and Eckert 2012) these repeats have elevated germline mutation and polymorphism rates. After a certain threshold length, STRs are termed microsatellites (Kelkar et al. 2010; Ananda et al. 2013). The high level of polymorphism makes microsatellites attractive markers for population and conservation genetics studies (Jarne and Lagoda 1996; Sunnucks 2000; Wan et al. 2004; Kim and Sappington 2013) and for identifying individuals in forensics (Hagelberg et al. 1991; Chambers et al. 2014). Many STRs are involved in gene regulation and protein function (Li et al. 2004), with ~17% of human genes containing STRs in their open reading frames (Gemayel et al. 2010). While long microsatellites have attracted much attention, length alterations even within relatively short repeat tracts are sometimes associated with disease (Li et al. 2004). For instance, differences in the number of repeats at the $(TG)_{10-13}(T)_{5-9}$ STR located within the splicing branch/acceptor site of the *CFTR* gene (exon 9) can affect in-frame exon skipping and as a result, can influence the severity of cystic fibrosis (Cuppens et al. 1990; Chu et al. 1991). The purity of STRs (the degree to which the perfect STR sequence remains uninterrupted) also has a functional effect. Interrupted STRs have lower mutation rates (Ananda et al. 2014), and this can diminish disease risk. For instance, approximately 6% of Ashkenazi Jews have an A to T mutation in the *APC* gene (encoding for a tumor suppressor) that alters an interrupted STR $(A)_3T(A)_4$ into a perfect $(A)_8$ (Laken et al. 1997). This increases the probability of somatic frameshift mutation within the STR, leading to *APC* protein inactivation. As a result, Ashkenazi Jews have a higher colorectal cancer risk (Gryfe et al. 1999). Since even small changes in STR length and purity can have functional effects, accurate STR profiling is crucial.

Despite the importance of STRs in evolution and disease, their accurate genotyping from Next Generation Sequencing (NGS) data has been challenging (see (Treangen and Salzberg 2012) for a review). Sequencing library construction frequently includes Polymerase Chain Reaction (PCR) steps during which a polymerase might undergo slippage at STRs, leading to amplicons that differ in length due to expansion and contraction of repeat units (Ellegren 2004; Wang et al. 2011). Additionally, base calling by NGS instruments at repetitive regions is frequently imprecise. These factors result in high sequencing errors at homopolymer runs produced by the 454 and Illumina instruments (Balzer et al. 2010; Albers et al. 2011).

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

From a bioinformatics perspective, if STR-containing reads are mapped in their entirety, some reads cannot be mapped because of high mismatch/indel penalties associated with STR lengths different than those at the corresponding positions in the reference genome. This obscures accurate estimation of allele frequency and underestimates the real level of STR variation in the genome. To alleviate this problem, a short-read alignment approach using non-repetitive flanks of STR-containing reads has been proposed recently (lobSTR, (Gymrek et al. 2012)). This tool has fast running time and takes into account PCR stutter noise during the genotyping stage. However, the entropy scanning implemented by lobSTR to detect STRs has low sensitivity for mononucleotide STRs and short STRs (less than 25 bp), which constitute a large proportion of STRs in the genome. Additionally, the allele frequency at STRs for *genetically heterogeneous samples*, for which a simple 1:1 ratio in allele frequency present in heterozygous diploids is not expected (e.g., for tumors, viral populations, and organelles), cannot be determined. Furthermore, lobSTR uses a fixed (embedded in the program) mapping algorithm. Novel short-read mapping and STR detection algorithms (Pellegrini et al. 2010; Lim et al. 2013) are constantly being developed; an STR-profiling tool that can be customized to incorporate emerging mapping algorithms is needed.

The recently released PCR-free Illumina library preparation protocol (hereafter called ‘PCR-”) is expected to improve STR genotyping accuracy. The direct advantage of limiting PCR steps during NGS is the increased uniformity of the sequencing depth (Kozarewa et al. 2009). Also, this protocol eliminates duplicate reads that obscure allele frequency profiling for heterogeneous genetic samples. Importantly, the degree to which the accuracy of calling STR alleles is improved using the PCR-free protocol has not been evaluated previously. Moreover, massive amounts of data have already been generated by the NGS technology with the PCR-*containing* library preparation protocol (hereafter called ‘PCR+’), and some such data cannot be regenerated due to the scarcity of samples and/or time- and cost-constraints. Therefore, universal methods are urgently needed that can evaluate and correct STR errors generated by NGS technology (both PCR- and PCR+), and accommodate evolving protocols and sequencing techniques.

Some efforts have been made to evaluate errors generated by NGS at STRs. For instance, errors at STRs sequenced with the PCR+ protocol vary with repeat number and motif size (Luo et al. 2012). However, an explicit quantitation of various sources of STR-related sequencing

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

errors has been lacking, which hinders an unambiguous estimation of STR mutational properties. Indeed, as both mutation and sequencing error rates increase with STR length (Kelkar et al. 2008; Luo et al. 2012; Highnam et al. 2013), one cannot confidently decipher mutation rates without accounting for sequencing error rates. Recently, a tool to guide genotyping of STRs using informed error profiles from inbred *Drosophila* lines (RepeatSeq), has been released (Highnam et al. 2013). This tool utilizes reads mapped by other programs, such as BWA (Li and Durbin 2009), Bowtie (Langmead et al. 2009), etc., and predicts the most probable genotype at a locus based on STR motif, length, and base quality. However, RepeatSeq uses the whole-read mapping approach which introduces a bias towards the STR length in the reference genome (Gymrek et al. 2012) and thus might obscure the true STR variation spectrum. Such biases can be accounted for by an error correction model based on the STR flank-based method.

To profile the full spectrum of STR lengths in the human and other genomes, and to correct for NGS-associated STR errors, we developed STR-FM (Short Tandem Repeat profiling using a Flank-based Mapping approach), a flexible pipeline for detecting and genotyping STRs from short-read sequencing data. Our pipeline can detect STRs of any length, including short ones (as short as only two repeats), includes an error-correcting module, and can incorporate any NGS mapping algorithm with paired-end mapping capability, making it adaptable to new mapping methods as they become available. Applying this pipeline, we asked the following questions. *First*, what are the rates and patterns of sequencing errors associated with STRs of different motif sizes (mono-, di-, tri-, and tetranucleotides), motif compositions, and repeat numbers? These were contrasted between publicly available genome-wide data sets sequenced with PCR+ and PCR- protocols, and validated with in-house generated, ultra-deep sequencing of plasmids harboring individual STR sequences. *Second*, do technical errors have different patterns from true STR mutations? *Third*, based on the detailed knowledge of the error profiles, what is the minimum sequencing depth required for producing reliable STR genotypes for PCR+ and PCR- protocols? As a result, we provide the scientific community with STR-FM, a reproducible and versatile pipeline for genotyping STRs that incorporates an error correction model. To illustrate the utility of STR-FM, we applied it to the completely sequenced human genomes from the Platinum Genomes Project (Ajay et al. 2011) and determined human genome-wide germ-line mutation rates at STRs.

RESULTS

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

The STR-FM pipeline. We designed the STR-FM pipeline as a collection of tools in Galaxy (Giardine et al. 2005; Goecks et al. 2010; Blankenberg et al. 2010, 2014), providing integration with various mapping algorithms and customization for a variety of applications. Either single-end or paired-end sequencing data can be utilized; for paired-end read data, each read is treated separately. The core of the pipeline consists of the following three procedures (Fig. S1). First, STR-FM runs a short-read STR detection tool using a string comparison algorithm (see Methods for details). The algorithm can detect exact (pure, or uninterrupted) STRs (mono- through hexa-nucleotide STRs greater than or equal to two repeats), incomplete motifs (e.g. ATATATA), interrupted STRs (e.g. AAAATAAAAA), or multiple STRs in a read. Reads that do not have sufficient upstream or downstream sequences flanking the STRs are discarded (we used a threshold of 20 bp on each side of an STR). Next, each read is split into two 'pseudo-reads', containing the upstream and downstream flanks surrounding the STR. These are mapped to the reference genome using a standard paired-end read-mapping algorithm (e.g. BWA (Li and Durbin 2009, 2010), Bowtie (Langmead et al. 2009), or Bowtie2 (Langmead and Salzberg 2012)), treating each pair of flanking sequences as a *faux* paired-end read. Finally, STR-FM runs a profiler tool, which groups all reads with STRs that are mapped to the same location in the reference genome. As a result, an array of all STR lengths from the reads mapping to a particular STR-containing locus is generated. The number of reads that completely cover an STR and its flanks is referred to as the *informative sequencing depth*. See Supplemental Text for details about running time and sensitivity of detection (Tables S1-S5; Fig S2).

Assessment of error profiles at STRs for PCR+ and PCR- library preparation protocols. The NGS errors at STRs were evaluated using two data sets. First, we utilized the PCR+ DNA sequencing data generated on a HiSeq instrument (150x depth) for a human male (SAMN00716185) sequenced as a part of the iPOP study (Chen et al. 2012). We combined the blood and saliva samples, as they had similar error profiles (see below). Second, we utilized the PCR- sequencing data generated on a HiSeq instrument (245x depth) for a human male blood sample (NA12882) sequenced as a part of the Platinum Genome Project (Ajay et al. 2011). The PCR- data were down-sampled to obtain the same number of filtered STRs (for each motif size) as present in the PCR+ dataset for fair comparison, i.e. to obtain the 150x depth. These two datasets are comparable as both were generated on a HiSeq instrument using paired-end 100-bp reads. Therefore, the differences in their STR error profiles are expected to reflect the contribution of the PCR step during library preparation. To avoid heterozygous sites, we limited

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

our analysis to the non-pseudoautosomal regions of Chromosome X. As this chromosome is hemizygous in males, any STR variability within an allele should result primarily from PCR and sequencing errors. We analyzed mono-, di-, tri-, and tetranucleotide STRs, which have lengths of at least 5, 6, 9, and 12 bp, respectively (the numbers of STRs analyzed are listed in Table S6; fewer loci were detected in the regions occupied by *Alu* and L1 repetitive elements). Only uninterrupted (perfect) STRs were considered. A stringent procedure was used to reduce bioinformatics errors (e.g., we removed non-uniquely mapped reads, reads with improper orientation between pairs, etc.; see Methods for details). For each STR locus in the reference genome's Chromosome X, we collected all mapped STR-containing reads using STR-FM and assigned the most frequent STR length as the major allele length for that particular locus. All STR lengths that differed from the major allele length were considered to represent sequencing errors. As the rate of sequencing errors correlates with STR length (Luo et al. 2012; Highnam et al. 2013), we grouped STR loci by the major allele length and evaluated their error profiles.

The PCR+ data had significantly higher error rates than the PCR- data for both mono- and dinucleotide STRs (Fig. 1A), except for very short ones (repeat number ≤5). For mononucleotide STRs, the difference in error rates increases until about 11 bp, after which it plateaus at an ~2.5-fold difference. For dinucleotide STRs, the difference increases until about 16 bp, plateauing at an ~3.6-fold difference. For most tri- and tetra-nucleotide loci studied (Fig. S3), the PCR+ data also had higher error rates than the PCR- data, however the confidence intervals of the two curves frequently overlapped. Blood and saliva PCR+ data had similar error profiles (Fig. S4). Our overall results did not change considerably depending on whether we used the complete (245x) or downsampled (150x) PCR- data (Fig S5); therefore, in the subsequent analyses we used the error rates from the complete (245x) PCR- data set, because it produces more narrow confidence intervals.

We observed several trends that were similar for both PCR+ and PCR- data. For all four repeat motif sizes, the error rates increased exponentially with repeat length and approached a plateau at longer repeats (Figs. 1A, S3, S5), corroborating previous studies (Albers et al. 2011; Montgomery et al. 2013; Gymrek et al. 2012). For a fixed repeat number, error rates were highest for mononucleotide, intermediate for dinucleotide and trinucleotide, and lowest for tetranucleotide STRs (Figs. 1A, S3, S5). We also classified sequencing errors into four categories: (1) shorter than the major allele length by one repeat (1-unit-del); (2) shorter by more than one repeat (>1-unit-del); (3) longer by one repeat (1-unit-ins); or (4) longer by more

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

than one repeat (>1-unit-ins). The '1-unit-del' was the most common error category for all STRs (Figs. 1B, 1C, S6, S7). For mono- and di-nucleotide STRs, the '1-unit-del' was followed (in the order of diminishing prevalence) by '1-unit-ins', '>1-unit-del', and '>1-unit-ins' (Figs. 1B, 1C, S6A, S6B). For the other STRs, no further conclusions could be drawn due to limited number of sites and overlapping confidence intervals (Figs. S6, S7).

We also studied the effect of motif composition on the error rate. For dinucleotide STRs at least 14 bp (7 units) long, significantly higher error rates were observed at $(AT/TA)_n$ repeats than at $(AC/GT)_n$, $(AG/CT)_n$, and $(GC/CG)_n$ repeats, in both PCR+ and PCR- data (Figs. S8B, S8D). For mononucleotide STRs, no significant differences in error rates between (A/T) and (G/C) repeats were observed for the lengths studied (Figs. S8A, S8C). For tri- and tetranucleotide STRs, the number of loci was insufficient to evaluate the influence of motif composition on error rates.

Somatic mutations also might contribute to the STR variability in the two data sets analyzed here; however, both the PCR+ and PCR- data should be affected by them equally. Moreover, the somatic mutation rates of dinucleotide STRs estimated in a previous study (Hile et al. 2000; Baptiste and Eckert 2012) are 100-1000-fold lower than the error rates estimated here (for the same repeat number). Therefore, the contribution of somatic mutations to STR error rates estimation here is negligible.

Confirmation of STR error patterns via ultra-deep sequencing of plasmids. The genome-wide PCR+ and PCR- data sets utilized above were generated in two different sequencing facilities, and this might have contributed to the discrepancy in their error rates, e.g. due to batch effect. We therefore sought to confirm our findings with an additional data set that we generated entirely in our own laboratory. Namely, using both PCR+ and PCR- protocols, we sequenced STRs cloned in a plasmid vector. All data were generated in our facility using the same DNA samples, personnel, and equipment. We sequenced 11 plasmid libraries containing STR inserts (Tables S7-S8) and a control vector plasmid library without an STR insert. We generated ~100,000x depth on a MiSeq instrument, which has error rates comparable to those of HiSeq (Quail 2012). In addition to the inserts, all plasmid vector sequences contained 49, 14, and one mono-, di-, and trinucleotide STRs, respectively (Table S8). Our plasmid analyses confirmed that PCR+ data had significantly higher STR error rates than PCR- data for mono- and dinucleotide STRs (Fig. 1D). The error rates between PCR+ and PCR- data were comparable for short STRs; however, for both mono- and dinucleotide STRs that were at least

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

seven repeats long, the PCR+ data had consistently higher STR rates than PCR- data -- by ~4.2- and ~9.3-fold for mono- and dinucleotide STRs, respectively. Our plasmid analysis also confirms that the '1-unit-del' is the most common category of errors (Fig S9).

Statistical model for genotyping STRs from NGS data. The error profiles estimated in this study can be used to correct for TR errors in several applications, including genotyping of diploid samples (distinguishing between homozygotes and heterozygotes) and of heterogeneous genetic samples (e.g., tumors, viral populations, and DNA-containing organelles). To achieve this, we developed an error correction model that takes STR features (repeat number and motif identity) into account (see Methods for details). Our error correction model can utilize error profiles (error rates depending on repeat numbers and motif identity) provided by the user (e.g. for novel sequencing technology), and thus is not limited to those generated by this study. The model determines the three most frequent STR lengths for an individual locus, and, based on the STR error array for this motif and repeat number, calculates the probability of generating this profile by a homozygote (all three possible homozygotes are evaluated) vs. a heterozygote (all three possible heterozygotes are evaluated). For instance, the model is able to decipher the true genotype $(A)_9/(A)_{11}$, even in the presence of $(A)_{10}$ variant calls. It determines the most probable heterozygote and homozygote alleles at a locus from the NGS data, and reports the log odds ratio between their probabilities, which can be interpreted as a confidence of genotyping (the higher this value is from 0, the more confidence we have in this genotype). For example, if for an STR locus sequenced at depth 9x we observed three, three, and three reads with $(A)_9$, $(A)_{10}$, and $(A)_{11}$, respectively, the most probable heterozygote is $(A)_9(A)_{11}$, and the most probable homozygote is $(A)_{10}$. According to the model, the log odds ratio for $(A)_9(A)_{11}$ compared to $(A)_{10}$ is 8.64×10^4, strongly favoring a heterozygote in this case.

Accuracy of the genotyping model in diploid samples. To test the performance of our error correction model, we generated a pseudo-diploid data set. For each class of STRs, we selected loci on the human male X Chromosome (from the same datasets that we used to evaluate the STR error profiles of the PCR+ and PCR- data above) that are supported by at least five reads. Then we generated 10,000 sets of 'homozygous' loci by randomly selecting two loci with the same motif and repeat number and combining their reads. Similarly, 10,000 'heterozygous' loci were generated by randomly selecting two loci with the same motif but different repeat numbers and combining their reads. We focused on the repeat numbers for which we could estimate error profiles with high confidence: length ranges of 6-15 bp, 8-23 bp, 12-20 bp, and 16-23 bp

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

for mono-, di-, tri-, and tetranucleotide STRs, respectively (see Methods and Tables S9-S10). Here we used error profiles from the human X Chromosome data as they include wider STR length range and higher resolution compared with the plasmid data.

The genotyping model correctly determined the genotype for 98%-100% of both homozygotes and heterozygotes (Table 1). For STRs with motif size of at least two, our model's accuracy was at least 99.68%. Mononucleotide STRs were slightly more difficult to genotype correctly than the other STR classes (1-2% lower accuracy) and were more difficult to genotype with PCR+ data than with PCR- data (about 1% lower accuracy). This is consistent with mononucleotide STRs having higher error rates (Figs. 1A, 1D, S5).

To investigate factors contributing to erroneous genotyping of mononucleotide STRs, we used logistic regressions with the correct prediction of an allele combination as the response, and with four predictors: the read depth, STR length, the difference in the lengths of STR alleles (for heterozygotes), and the ratio of read depths supporting each allele (for heterozygotes; see Methods for details). A separate logistic regression was run for PCR+ and PCR- data, and for heterozygous vs. homozygous alleles. All four predictors were significant in at least one of the four models (Table S11), although the sets of significant predictors and their relative contribution to the explained variability differed among models. However, we observed that high sequencing depth, small STR length, large difference between two alleles in heterozygotes, and similar ratio of read depths supporting each allele in heterozygotes increased the prediction power of each model (see Fig. 2 for examples).

We also evaluated the robustness of genotyping using technically replicated samples—two PCR- libraries (43x and 60x) of the whole genome for sample NA12882 from the Platinum Genome Project (Ajay et al. 2011). We profiled the STR lengths and genotyped those sites that had a minimum informative read depth of 5x in both samples. The ratio of loci that had different genotypes (the genotype discordance rate) between the two replicated libraries was only 0.29%.

Accuracy of the genotyping model in heterogeneous genetic samples. Our error correction model is able to accommodate applications others than standard diploid genotyping. For example, the allele contribution ratio can be modified to allow for a non-diploid allele frequency ratio commonly present in tumors and genomes of organelles (e.g., mitochondrial DNA). The user can find the percentage of heterogeneity that maximizes the likelihood of the observed

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

STR length array, or test the ratio that was suggested by a linked genetic marker (e.g., a SNP). To demonstrate this capability, we genotyped a sample of two plasmids with cloned mtDNA's D-loop sequence, pre-mixed at five different ratios -- 98:2, 99:1, 99.5:0.5, 99.75:0.25, and 99.9:0.1. The two plasmids differ at only one STR-containing region: the plasmid Z1-1 with a higher proportion of DNA in the mixture has a $(C)_8$ allele, while the lower proportion plasmid R2 has a $(C)_7$ allele (Rebolledo-Jaramillo et al. 2014). Each mixed DNA sample was sequenced on a MiSeq instrument using 250-bp paired-end reads at the depth of 300-2,000x per site. Each mixture was sequenced twice, to generate two replicates. We ran our STR-FM pipeline separately on each mixture and each replicate, and genotyped the STR locus in each case. Since the plasmids were sequenced at a very high depth, the prediction should reflect the competency of the model to detect the rare allele, and not the sequencing depth (see below). The minor allele was detected for both replicates at ratios of 98:2 and 99:1, and in one out of the two replicates for the other three mixing ratios (Table S12).

Genotyping disease-associated long trinucleotide STRs. We explored the abilities of the STR-FM pipeline and statistical model to genotype long trinucleotide repeats associated with diseases. Using the PCR-free data described above (NA12882 sequenced with 100-bp reads at 245x depth (Ajay et al. 2011)), we genotyped 10 disease-related trinucleotide STRs from Table S1 in (Castel et al. 2010). Five of them had read depth above 5x (repeat length 16-48) and thus could be genotyped with confidence >90%. These included a 48-bp (homozygote) CAG repeat in the *JPH3* gene (Wilburn et al. 2011), which is associated with Huntington's disease-like 2, and a 32-bp (also homozygote) CCG repeat in the *RELN* gene, which is associated with a higher risk of autism (Table S13). For the remaining five loci (including those associated with spinocerebellar ataxia 6, Jacobsen syndrome, dentatorubral-pallidouysian atrophy, and Huntington's disease; repeat length of 34-55 bp), the informative read depth was below 5x preventing us from genotyping such loci with confidence (Table S13).

STR germ-line mutation rates and patterns. The germ-line mutation rate is a fundamental parameter in molecular evolution; however, it has not been evaluated for human STRs on a genome-wide scale. We utilized our genotyping model to estimate STR germ-line mutation rates using a three-generation pedigree -- four grandparents, two parents, and 11 children (Fig. 3A) -- sequenced at high depth (43-64x) with a PCR- protocol as part of the Platinum Genome Project (Ajay et al. 2011). For each sequenced individual, we detected and genotyped uninterrupted

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

(perfect) STRs with at least 6, 8, 9, and 12 bp for mono-, di-, tri-, and tetra-nucleotide STRs, respectively.

Overall, the distribution of STR allele lengths as evaluated from sample NA12882 (one of the 11 children) was similar to that described in the recent studies (McIver et al. 2011; Payseur et al. 2011; Gymrek et al. 2012; Willems et al. 2014): the proportions of variable loci increased with repeat length and, at the same repeat length, were lower for STRs with longer motifs (Fig. S10). Here we confirm these trends for mononucleotide STRs, which were removed from previous studies due to high sequencing error rates or inadequacy of error profiles. We also confirm that alleles at most heterozygous STRs differ by just one motif unit (Fig. S11) (Payseur et al. 2011; Gymrek et al. 2012). Thus, accurate genotyping of such heterozygotes is important.

We studied germ-line indel mutations (insertion or deletion of whole repeat units) in autosomes that occurred in transmissions from grandparents to parents and that were confirmed by genotypes of children (Figs. S12-S13). To be scored as a germ-line mutation, the event was required to have the following characteristics: (a) a parent having an allele absent from his/her father or mother (a 'putative mutant allele'); (b) a putative mutant allele present in at least two children to confirm that it is a germ-line variant; and (c) a putative mutant allele absent in the second parent (Fig. S12A). If a variant STR allele is present in both parents, the child's genotype was required to unambiguously identify the parent of origin for the putative mutant allele (Fig. S12B, Table S14). We included all 11 children in our data set to capture the vast majority of germ-line mutations occurring between grandparents and parents. The cases when putative mutant alleles were present in only one child were not considered to reduce false positives resulting from incorrect genotyping in children.

We required STR-containing loci to be sequenced to at least 5x informative read depth in all children, parents, and grandparents (Fig 3A). The number of loci satisfying this requirement totaled 9,726,196 (4,863,098 loci in father and 4,863,098 in mother, equal to 19,452,392 transmitted STR alleles, as one allele is transmitted from each grandparent). At such loci, we detected 1,470 germ-line mutation events, which is equivalent to 7.6×10^{-5} mutations per locus per generation (Tables S14, S15). Among these mutations, 452 and 426 occurred in the male and female germ line (Table S14), respectively, corroborating male mutation bias (Kelkar et al. 2008; Sun et al. 2012). The parental origin of the remaining 592 de novo mutations could not be determined. The level of male mutation bias varied among STR classes and motifs (Figs. 3B,

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

S14). The germ-line mutation rate increased with repeat length and, among STRs of the same length, was highest for mononucleotide repeats (Fig. 3C). The STRs with C and AT motif are the most mutable among mono- and di-nucleotide STRs (Fig. S15) corroborating previous studies (Denver et al. 2004; Kelkar et al. 2008; Sun et al. 2012). For mono- and dinucleotide STRs the '1-unit-ins' were slightly more abundant than '1-unit-del' among STR de novo mutations (Figs. 3D and S16), even though the confidence intervals are highly overlapped. This is in contrast to STR NGS sequencing errors that were predominantly '1-unit-dels' (Fig. 1B, 1C, S6 and S7).

We next examined whether these mutations were distributed randomly along the genome. The observed number of mutant mononucleotide STR loci estimated in 50-Mb windows had high correlation with the number of mutations expected based on our estimated mutation rates by repeat number and corrected by the number of loci present in each window (Pearson correlation = 0.77; Spearman correlation = 0.79; Fig. S17), suggesting that these mutations were randomly distributed along the genome. Only two windows (chr6:0-50000000 and chr11:100000000-135006516) had a significantly higher number of mutations than expected (the p-value of the post-hoc binomial exact test of these two windows was 1.3×10^{-5} and 4.7×10^{-5}, respectively; Fig. S17). Our data are insufficient to test the randomness of the distribution at smaller windows, or for repeats with a larger motif size.

Required coverage for accurate genotyping. The power to accurately genotype a locus depends on many factors (Fig. 2), including the length of true alleles, the combination of true alleles, the ratio of read depths supporting each allele, and the informative read depth at that locus. Based on the error profile we assessed and the genotyping model we formulated, we can estimate the minimal read depth required to correctly determine an allele combination for a locus, given the observed STR length array from the reads mapping to this locus. Recall that informative read depth at a locus is the number of reads that span the entire STR and possess ≥20-bp flanks on both of its sides.

We first evaluated the minimum informative read depth required to accurately genotype a locus as a function of the allele length. We focused on heterozygous loci with the two alleles containing consecutive repeat numbers, since these are the hardest cases to genotype. We used simulations (see Methods) to compute the minimal number of informative reads required to correctly genotype with 90% accuracy (Fig. 4, Table S16). For instance, to correctly genotype

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

$(A)_{13}(A)_{14}$, 13 informative reads are required for the PCR+ protocol, while only seven informative reads are required for the PCR- protocol. For PCR- data, only five informative reads per locus is required for genotyping of STRs for all motif sizes and repeat numbers up to nine. The PCR+ data usually requires a higher number of informative reads for mono- and dinucleotide STRs. For tri- and tetranucleotide STRs up to 11 repeats, the number of informative reads required is only 5 (Fig. 4), as they have lower error rates (Figs. 1A, S3, S5). The sequencing depth requirements for longer tri- and tetranucleotide STRs could not be evaluated because we lacked error profile data for such long STRs. We also evaluated the effect of the motif itself, since some motifs are more error prone than others (Table S16). For the STR lengths studied, we observed a difference in the number of informative reads required to genotype dinucleotides of allele combination 7/8 -- seven informative reads are required to genotype $(AT)_7(AT)_8$ but only five for $(AC)_7(AC)_8$ and $(AG)_7(AG)_8$. The trends for minimum informative read depth for genotyping with 95% accuracy are similar (Table S17).

Next, we calculated the average genome-wide depth needed to achieve a given number of informative reads with 90% likelihood (see Methods). This depth is defined as the ratio of the number of sequenced bases to the genome length. The needed genome-wide depth is higher than the informative depth due to the randomness of the sequencing process and the need to have sufficiently long non-unique flanking regions. Because our calculations do not account for sequencing quality score, genome location sequencing bias, or mapping bias, they only serve as approximations for what is minimally needed in practice. Figure 5 shows the required average genome-wide depth as a function of repeat length of interest and read length utilized. The advantage of long reads is more pronounced for higher repeat lengths. For instance, to achieve 10x informative read depth for 10-bp STRs (independent of their repeat numbers), one needs to accomplish genome-wide sequencing depth of 26x and 17x for 100-bp and 300-bp single-end reads, respectively. However, to achieve the same depth for 50-bp STRs, one needs 104x and 21x genome-wide sequencing depth for 100-bp and 300-bp single-end reads, respectively.

The conversion from informative to genome-wide sequencing depth substantially increases the difference in depth requirements between PCR+ and PCR- protocols. For instance, in the $(A)_{13}(A)_{14}$ genotyping example above, the difference in the required informative sequencing depth between PCR+ and PCR- protocols was only 6x (13x vs. 7x). To achieve these informative read depths with 100-bp reads, the genome-wide sequencing depth required for

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

PCR+ and PCR- protocols is 35x and 21x, respectively, leading to the difference of 14x. This result emphasizes the advantage of the PCR- library preparation protocol in STR genotyping.

DISCUSSION

STRs are important causative agents of human diseases (Pearson et al. 2005; Castel et al. 2010) and are useful genetic markers (Wright and Bentzen 1994; Gupta and Varshney 2000; Miah et al. 2013; Abdul-Muneer 2014), but their genotyping with NGS technology is a challenge. To overcome this, we developed STR-FM, a versatile STR profiling method that is applicable to a wide range of STR lengths and motifs. We have thoroughly evaluated STR error profiles generated using Illumina PCR+ and PCR- library preparation protocols, using both computational and experimental approaches. As a result, we provide the scientific community with an STR genotyping tool including an error correction model, and a tool to evaluate the sequencing depth required for accurate genotyping of STR loci.

Factors that determine STR error rates and patterns. STR error profiles depend on repeat length (Figs. 1A, 1D, S5), motif length (Figs. 1A, 1D, S5), and motif composition (Fig. S8), corroborating previous findings (Shinde et al. 2003; Highnam et al. 2013). For the STR length ranges considered in this study, mononucleotide STRs have the highest error rate, followed by di-, tri-, and tetranucleotide STRs, respectively (Figs. 1A, 1D, S3, S5). The patterns of STR NGS errors we describe here are similar to those previously reported for STRs typed via gel electrophoresis (Murray et al. 1993; Walsh et al. 1996; Shinde et al. 2003): the most common STR error category resulting from slippage during PCR amplification is the '1-unit-del' (e.g., (Murray et al. 1993; Walsh et al. 1996)). The error rate of 10^{-6}-10^{-5}, observed by us here for five-unit mononucleotide STRs (Fig. 1), is consistent with an experimentally measured one-base deletion error rate within a five-base mononucleotide STR (~1 error per 20,000 nucleotides) synthesized by *Taq* polymerase (Eckert and Kunkel 1990). Also, our results based on the NGS data (Figs. S3, S5) agree with previous observations that tri- and tetranucleotide microsatellites have fewer stutter bands compared with di- and mononucleotide STRs (Ellegren 2004). The Illumina NGS data studied by us were generated using the Phusion polymerase (Ajay et al. 2011; Chen et al. 2012; Oyola et al. 2012); therefore the observed error patterns might be driven by the error profile of this enzyme. The error rates of the Phusion enzyme are ~50-fold lower than those of the *Taq* polymerase (Frey and Suppmann 1995). However, the similarity between our error patterns and those of the *Taq* polymerase (Shinde et al. 2003) suggest that the error patterns of these two polymerases are comparable.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

The PCR- library preparation protocol is expected to reduce bias in sequencing coverage (Kozarewa et al. 2009). Our findings reveal that eliminating the PCR step reduces STR errors as well. Based on our X Chromosome data, approximately a 1.5-2.8 fold and 2.0-3.7 fold error reduction was observed for mono- and dinucleotide STRs with greater than six and five repeats, respectively. This allows the PCR- protocol to achieve the equivalent power of the PCR+ with lower sequencing depth. For example, to distinguish $(A)_{14}$ from $(A)_{15}$ using 100-bp reads, we need 39x sequencing depth for PCR+ data, but we need only 29x depth for PCR- data.

Here we used the hg19 reference genome sequence. Mapping STRs to GRCh38 (Rosenbloom et al. 2015) might increase the number of loci that we will identify, but should not affect our overall conclusions. Indeed, we utilized stringent mapping parameters to map STR flanking regions without mismatches and indels. This minimized incorrect mapping to paralogous loci that are potentially more abundant in the new reference genome.

Minimum sequencing depth. We developed a method to estimate the informative and genome-wide sequencing depths required for STR genotyping. We demonstrate that the required sequencing depth depends on the STR error profile (as discussed above), sequencing read length, allele combination, and allele sequencing depth skew in a heterozygote. The depth-estimating tool integrated in Galaxy (Giardine et al. 2005; Goecks et al. 2010; Blankenberg et al. 2010, 2014) can be used for: (1) selecting sequencing depth while planning a new project; and (2) evaluating which STRs can be accurately genotyped from a completed sequencing project (e.g., using a publicly available data set). The desired level of resolution contributes to determining the required read depth. For instance, for some questions (e.g., for studying allele spectrum in search of loci evolving under selection) distinguishing alleles of consecutive lengths in a heterozygote is important. For other questions (e.g., scoring normal vs. pre-mutation disease-causing alleles separated by several repeat numbers) such high resolution might not be needed. Therefore, the recommended minimal depth for accurate genotyping is not a fixed number, and depends on the research question and STR loci of interest.

We found that to accurately genotype STRs with less than nine repeats using the PCR- library protocol, the required informative read depth is only 5x, which corresponds to genome-wide sequencing depths of 15x, 17x, 21x, and 26x for mono-, di-, tri-, and tetranucleotide STRs,

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

respectively (for 100-bp reads). These sequencing depths are comparable to those present in most publicly available data sets.

Our genotyping model and minimum sequencing depth study provide us with an opportunity to recommend the minimal depth needed for sequencing of disease-causing STRs. For instance, a change from 10 to 12-17 repeats at the $(CGC)_n$ repeat, present at the gene *PABPN1* encoding poly(A) binding protein, causes Oculopharyngeal Muscular Dystrophy (*OPMD*)(Brais et al. 1998; Pearson et al. 2005). Running our simulations, we came to a conclusion that at least five informative reads are required to correctly genotype the $(CGC)_{10}(CGC)_{12}$ heterozygote with a 90% success rate. This corresponds to a genome-wide sequencing depth of 20x for 100bp reads, using either the PCR+ or PCR- protocol.

Tool for STR length profiling. The STR-FM pipeline allows one to: (1) detect a broad range of STR lengths and motifs (including mononucleotide STRs and short STRs; Table S1); (2) clearly distinguish between perfect (uninterrupted) and interrupted STRs (Table S2); and (3) be flexible with respect to the algorithm to be used for mapping the STR flanking sequences. Unlike other approaches, STR-FM can profile mononucleotide STRs, can profile heterogeneous genetic samples that do not follow a 50:50 allele frequency distribution (down to an allele frequency of 0.01), and can detect STRs in species without an annotated reference genome. Because STR-FM aligns only the flanking sequences and not the whole reads, it is not susceptible to bias of STR length and can be used to compare the mutational spectrum across different lengths. Our pipeline is integrated into Galaxy (Giardine et al. 2005; Goecks et al. 2010; Blankenberg et al. 2010, 2014) and thus can be customized via a graphical user interface.

The STR-FM pipeline also could be used to extract STRs from mapped long reads generated by SMRT technology (Pacific Biosciences, (Roberts et al. 2013; Chaisson et al. 2014) or MinION (Oxford Nanopore, (Schneider and Dekker 2012). The long reads from these technologies would allow us to study longer microsatellites, which exceed the current read length of Illumina technology. Due to the differences in read lengths and sequencing errors between Illumina and these technologies, however, such data would require special mapping procedures and the development of separate error models. Alternatively, STR-FM can take mapped reads in standard SAM format, detect STRs in the mapped reads, and profile the STR length. While the unbiased property of flank-based mapping will be lost this way, this demonstrates the versatility of the STR-FM in profiling STRs, even when an *ad hoc* mapping algorithm is utilized.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Germ-line mutation rates of TRs in the human genome. We report here the first genome-wide study of STR mutation rates. Using the NGS data, we genotyped >4.8 million STR loci, as compared to 2,477 loci examined in a previous study (Sun et al. 2012). The pattern of *de novo* STR mutations we observed differs from that of STR sequencing errors (Figs. 1B, 1C, 3D, S6, S7, S14), a result that demonstrates the importance of applying our error correction genotyping model when studying STR mutations in genomes. Specifically, germ-line mutations have comparable rates for expansions and deletions among the repeat lengths we studied, whereas STR sequencing errors have a very strong deletion bias.

The male mutation bias of dinucleotide STRs observed here is lower than that described in another recently published study (Sun et al. 2012). The latter study was not genome-wide and included a higher proportion of AC-repeats than we find genome-wide (compare Table S3 in (Sun et al. 2012)) with our Table S18). Because AC-repeats have the highest male mutation bias among dinucleotide repeats we studied (Fig S14), this could have inflated male mutation bias in the study by Sun et al. (2012). Also, a confounder in our study is that we could identify the parental origin for only 60% of *de novo* mutations. We could not assign a parental mutation bias when both grandparents carry an allele identical to the mutant allele in the next generation (e.g. mutation from $(A)_6$ to $(A)_7$ when both grandparents have $(A)_6$ allele), which could underestimate the real male mutation bias in our study.

The germ-line mutation rates we estimated for STRs are higher than recently obtained estimates (Sun et al. 2012). Several factors could contribute to this difference. We hypothesize that a major factor is that we used perfect (uninterrupted) STRs exclusively in this study, while the study of Sun and colleagues (2012) included interrupted STRs that, as we have shown, have significantly lower mutation rates genome-wide (Ananda et al. 2014). Additionally, our study presents a genome-wide analysis including all STR-containing loci in the genome for which we could obtain data of high depth. In contrast Sun and colleagues analyzed loci used in large-scale gene-mapping studies, and thus potentially affected by ascertainment bias. Finally, our analysis is based on the analysis of a single family, while Sun and colleagues (2012) included hundreds of families, and this can also lead to differences in mutation rates.

We can apply these STR germ-line mutation rates and patterns in several areas. In population genetics, these rates can be used to estimate divergence times of recent population splits.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Since STRs have higher mutation rates than SNPs (Ellegren 2004; Campbell and Eichler 2013), the former can accumulate many mutations in a short period of time and thus can provide higher resolution. In forensics, knowing mutation rates can aid in estimating the level of STR polymorphism, in computing the number of STR markers necessary for individual identification or kinship testing, and in computing the probability of relatives having different alleles due to a mutation. Also, in medical research, we can evaluate the likelihood of pre-mutation STR loci to become disease-causing STRs in the next generation. Although most studied disease-causing STRs are long trinucleotide repeats (Pearson et al. 2005), some short STRs were also shown to have strong implications in several diseases (Cuppens et al. 1990; Chu et al. 1991; Pearson et al. 2005).

METHODS

Data sets used to assess STR error rates. The PCR+ data were downloaded from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) with accessions SRR345592-SRR345594. SRR345592 is a 100-bp paired-end dataset sequenced from blood of a human male at the 50x depth. SRR345593 and SRR345594 are 100-bp paired-end datasets sequenced from saliva of a human male at the 100x depth.

The PCR- data were downloaded from the European Sequence Read Archive (http://www.ebi.ac.uk/ena/) with accession numbers ERR194151, and ERR174342-ERR174360. This male individual is a Utah resident (NA12882 from the CEPH collection) whose blood was sequenced at the 245x depth.

Using STR-FM to assess error profiles in the context of STRs for Illumina data

Detecting STRs in short reads Our STR detection module for uninterrupted STRs executes the following three steps: (1) Scanning reads using sliding windows. For a given '*k*' (e.g. *k*=2 for dinucleotide STRs), we compared consecutive *k*-mer window size sequences, with a step size of *k*. If a sequence at a given position matches the previous k-mer, it was marked with a *plus* and with a *minus* if otherwise. (2) Since we do not allow mutations in the reported STR, consecutive *plus* signal sequence means that a *k*-mer STR is present in this sample. (3) Report *k*-mer STRs if the length is larger than the threshold provided by the user.

The method to detect interrupted STRs can be found in the Supplemental text. Using the STR-FM pipeline, we detected mono-, di-, tri-, and tetra-nucleotide STRs, with length ≥ 5, ≥6, ≥9, and

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

≥12 bp, respectively, in raw sequencing reads. These four classes of STRs were analyzed separately throughout the analysis. We also required that all bases in the STRs and the 20 flanking bases on both sides have a Phred quality score of ≥20.

Mapping flanking bases of STRs to the reference genome. 20 bases flanking the STRs upstream and downstream in the STR-containing reads were extracted as FASTQ files and mapped to the human reference genome (GRCh37) with BWA version 0.7.5a-r405 (Li and Durbin 2009) as paired-end reads using aln and sampe mode. To minimize incorrect mapping, mismatches and indels were not allowed in mapping. The resulting alignments were filtered to retain only the uniquely mapping flank-pairs with proper orientation of the upstream and downstream flanks. We discarded STRs that mapped to the Y Chromosome (pseudoautosomal regions) and human self-alignment regions from the UCSC Genome Browser to obtain only unique hemizygous loci on the X Chromosome (reads mapping to autosomes were discarded). Duplicated reads were not removed throughout the analysis to get accurate magnitude of errors generated by the PCR+ and PCR- library preparation protocols.

Profiling STR length at each locus. At each STR locus in the reference genome (identified using the same script to detect STRs in reads starting from the FASTA-formatted human reference genome (GRCh37)), all mapped STR-containing reads were binned and the STRs contained in these reads were collated. Only uninterrupted reference STR loci were considered. STRs with the same motif size (mono-, di-, tri-, or tetranucleotide repeats) on the X Chromosome that are closer than 10 bp were excluded to avoid incorrect STRs profiling from overlapping STRs. In total, 789,401, 378,730, 59,383, and 19,322 loci mono-, di-, tri-, and tetra-nucleotide STRs, respectively, were studied. The STRs were checked for motif compatibility with STR motifs in the reference genome. The major allele length for each locus was determined for each STR at each reference locus. Loci supported by <5 reads or with frequency of major allele <50% were removed. Then we grouped loci by their major allele lengths to evaluate their error profiles. Any STR length differences from the major allele length were considered to be erroneous calls.

Construction of bootstrap confidence intervals. For every STR motif size, motif composition, and allele length, using the observed counts of error variants of different lengths, we generated multinomially distributed random counts of error variants, and computed the error rate using these counts. This was repeated 10,000 times, and the 95% confidence intervals for the error rate were constructed from the 2.5^{th} to the 97.5^{th} percentiles of the error rates. The bootstrap

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

confidence intervals were not drawn when the sampling numbers of error counts at particular percentiles were zero. For example, several data points miss the lower bound of confidence intervals due to the small number of error counts.

Plasmid construction, sequencing, and analysis. The construction of plasmids containing artificially inserted tandem repeat sequences have been previously described (Eckert et al. 2002; Kelkar et al. 2010; Ananda et al. 2013; Baptiste et al. 2013; Ananda et al. 2014). The standard procedures were used for library preparation and the sequencing on Miseq. The analysis was performed using comparable parameters with the analysis in human X Chromosome. See Supplemental Text for complete description.

Building error correction model for genotyping. The raw reads were processed to generate STR length array using the STR-FM pipeline. The STR length array for each locus contained repeat length(s), number of repeats, and motif. The three most common STR lengths from the input STR length array of a single locus were considered for formulating homozygotes and heterozygotes.

The probability for any given allele length to be observed was retrieved from either PCR+ and PCR- error profiles generated from X Chromosome data. When the conversion could not be observed from an error profile (e.g. $(AG)_3$ was not observed to change to $(AG)_8$), the base substitution rate (Kong et al. 2012) was used. The probability value for all the repeat lengths in the input STR length array was multiplied to calculate the probability to generate the whole STR length array from either a homozygote or a heterozygote.

For homozygotes, each of the three most common STR lengths was considered. The allele that gave the highest probability to generate the observed STR array at a locus was considered to represent the probable homozygous allele. For heterozygotes, the three most common STR lengths were grouped into three sets of alleles. For each set, the probability of the two alleles to generate the observed repeat lengths was averaged. The most probable set of alleles was used as a probable heterozygous form. The log-odds ratio of the homozygous allele to the heterozygous allele was then reported, with a score greater than zero indicating a homozygote and a score below zero indicating a heterozygote.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Evaluation of the error correction model's correct prediction rate for diploid samples. We used the male X Chromosome sequencing data from the iPOP study (Chen et al. 2012)(for PCR+) and Platinum Genome Project (Ajay et al. 2011) (for PCR-) to in silico generate pseudo-diploid data to test the genotyping model (see Results). We generated 10,000 homozygous loci by combining two loci with the same motif and major allele length. Heterozygous loci were generated similarly. To ensure that the test dataset had an error profile similar to the actual data, we focused on loci in which the major alleles were at least one repeat longer than our minimum cut-off (≥6, ≥8, ≥12, and ≥16bp, for mono-, di-, tri-, and tetra-nucleotide STRs respectively) to allow STR contractions, which are the major category of STR errors. We considered only the range of STR lengths that had at least 500 reads in both PCR+ and PCR- data in the STR genotyping model to ensure reliability of error rates. Motifs that had only one STR length that passed the criteria were removed because they would always have a 100% correct prediction and therefore would not be useful for evaluating our error correction model. The prediction was considered to be correct when the allele combination was determined correctly.

Logistic regressions were performed using standard R package (R Core Team 2013). Pseudo R-squared of a model was calculated as $(D_o–D)/D_o$, where D_o is the null deviance and D is the residual deviance of the model. Relative contribution of each predictor to a model was calculated using $[(D_o–D)-(D_o–D_{(-p)})]/(D_o–D)$, where $D_{(-p)}$ is the deviance of a model obtained by removing the predictor of interest (Fungtammasan et al. 2012).

The genotype concordance/discordance rate was evaluated on libraries ERR194151 and ERR324433 from the European Sequence Read Archive (http://www.ebi.ac.uk/ena/). The genotype discordance was defined as the incongruence of at least one allele between genotypes of two replicated libraries (Pompanon et al. 2005). Minimum read depth was set at 5x.

Evaluation of the genotyping model's correct prediction rate for heterogeneous genetic samples. To assess the accuracy and precision of the error correction model, we prepared heterogeneous genetic samples by mixing DNA from two different plasmids (See Supplemental Text for details). The ratios of artificial heterogeneous genetic samples (clones R2 to Z1-1) were designed as 0.25%, 0.5%, 1.0%, 1.5%, and 2.0%. The sequencing data of mixed clones were profiled using the STR-FM pipeline described above. The sequences from the two plasmids

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

were also profiled to check STR lengths in the inserted DNA region that were different between the two plasmids. The error correction and genotyping were performed using the error rates estimated from the two unmixed Z1-1 and R2 plasmids. Also, the ratios of contributed alleles were adjusted from 50:50 to the ratios present in each mixed sample.

Using error correction model to evaluate germ-line mutation rates and patterns. PCR-whole-genome sequencing data of the three-generation pedigree from the Platinum Genome Project (Ajay et al. 2011)was downloaded from http://www.ebi.ac.uk/ena/data/view/ERP001960. This family consists of four grandparents, two parents, and 11 children. The data include libraries ERR194146-ERR194148, ERR194151-ERR194152, ERR194154-ERR194155, ERR194157-ERR194162, ERR218433 and ERR324432-ERR324435. Sequencing depth for each individual is 43.76-64.98x. We considered only autosomal STR-containing loci, which are sequenced to at least 5x informative read depth in all children, parents, and grandparents. STRs on sex chromosomes were excluded. We genotyped all individual of the pedigree using error rates estimated from X Chromosome of PCR- data which was evaluated from an individual of this pedigree (NA12882) to reduce potential differences due to batch effect. The germ-line mutations were classified by the gender of mutated germ line (male vs. female), original length, and mutated length. The germ-line mutation rates were calculated as the fraction of mutations per the total number of transmissions (each locus has two transmissions). The expected numbers of mutations in 50 Mb windows were calculated using the germ-line mutation rates obtained in this study separately for different STR classes, lengths, and motifs multiplied with the numbers of correspond STR loci that were genotyped in all members of pedigree. The deviation from expected number of mutations in 50 Mb windows was tested using binomial exact test.

Computing minimum informative read depth required for accurate genotyping. We simulated series of read profiles of all possible combinations of alleles with consecutive repeat numbers for 2-50x, 60x, 70x, 80x, and 100x depth for motifs A, T, AT, AC, AG, combined motif of trinucleotide STRs, and combined motif of tetranucleotide STRs. First, for each motif and STR length combination, we selected the range of observed STRs from our error profiles and generated all possible combinations of STR arrays. In this step, we did not include the observed STR lengths that have the probability of occurrence <0.01 (rare error form), to reduce computational time. We limited our study to range of STR lengths for which we observed at least 500 reads in our error profiles in both PCR+ and PCR- data. Second, we used our error

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

correction models to analyze these STR arrays and select only those that can be predicted to be heterozygotes. Third, we calculated the probability of our model to generate the observed STR arrays from predicted heterozygous allele pairs and multiplied by the number of all possible rearrangements in these STR arrays. A sum of these products was equal to the probability to detect heterozygous alleles for each allele pair at each depth.

Estimation of required sequencing depth. First, we derived an equation to calculate the *locus-specific* sequencing depth needed to achieve a certain informative read depth by assuming uniform sequenced reads along the genome. Suppose that the sequenced reads of length L uniformly represent parts of a genome that has an infinite size and no redundant sequences. If we place these reads on every position of this genome, the sequencing depth will be equal to read length L. If an STR of interest is r bp long and we require at least F bp of flanking regions both upstream and downstream, there will be $L+2F+r-1$ reads that are associated with at least one base of STR or flanking regions. Among these reads, $2F+r$ start positions and $2F+r$ end positions of reads locate within STR or flanking regions. These positions are mutually exclusive as long as reads are longer than $2F+r$. Thus, $2(2F+r-1)$ reads do not cover at least one base of STR and F bases either upstream or downstream (-1 is the correction for start and end position at the outer edge of flanking regions which can still map). Since the reads that do not cover the whole STR of length r or do not cover at least the required minimum flanked bases are not useful for profiling, the informative read depth of L is $L+(2F+r-1) - 2(2F+r - 1) = L-(2F+r-1)$. To achieve the informative read depth of X, an STR at a specific locus needs to be sequenced at depth:

$$\frac{X * L}{L - (2F+r-1)} \quad \text{(equation 1)},$$

where L is read length, F is the number of flanked bases required at each flank, r is the expected repeat length of an STR of interest. We will use $y_{required}$ to represent this value. This equation is true for all flank-based mapping approach programs that require certain lengths for both flanking regions.

In reality, the sequenced reads are not evenly distributed along the genome. Therefore, we used Poisson distribution to estimate the required genome-wide sequencing depth (the average depth for all loci in the genome) that guarantees that a certain percentage of STRs in the genome (e.g. 90%) have at least a certain level of locus-specific sequencing depth and informative read depth. Using y to represent specific level of sequencing depth and λ to

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

represent genome-wide sequencing depth, we can write the equation that the probability of observing a locus with depth y or P(Y=y) is:

$$(\lambda^{y} * e^{-\lambda}) / y! \qquad \text{(equation 2)}.$$

We are interested in finding the required genome-wide sequencing depth λ that gives sufficient informative read depth for certain percentage (e.g. 90%) of the genome. In other words, we want to find minimum value of λ that makes P(Y ≥ $y_{required}$) greater than or equal to a specific percentage of genome. The P(Y < $y_{required}$) can be calculated from:

$$P(Y=0) + P(Y=1) + \ldots + P(Y=y_{required}-1) \quad \text{(equation 3)}.$$

Each term of the Poisson distribution can be calculated from equation 2, and $y_{required}$ can be calculated from equation 1. The tool to estimate locus-specific $y_{required}$ and genome-wide sequencing depth 'λ' is provided in the Galaxy toolshed.

GALAXY tool description. The tools from the STR-FM pipeline are freely available on Galaxy (https://usegalaxy.org/) (Giardine et al. 2005; Goecks et al. 2010; Blankenberg et al. 2010, 2014). Users can install local Galaxy and download all tools used in this research from repository 'microsatellite_ngs' (See Supplemental text for details). The STR error rates estimated in this study can be downloaded from https://usegalaxy.org/u/guru%40psu.edu/h/error-rates-files.

Data Access

The sequencing data of the plasmid containing artificially inserted tandem repeat sequences in both PCR+ and PCR- systems, all the ratios of heterogeneous genetic samples, and the original plasmids before mixing have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP047377.

Acknowledgements

We thank Sarah J. Carnahan Craig for her comments on the manuscript; Howard Fescemyer, Wilfried Guiblet, Rahulsimham Vegesna, Monika Michalovová, Marta Tomaszkiewicz, Maria Krasilnikova, and Samarth Rangavittal for suggestions on sequencing data analysis; Francesca Chiaromonte and Prabhani Kuruppumullage Don for statistical comments; Daniel Blankenberg

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

and Martin Cech for their help with integration of our tools in Galaxy; and the Huck Institute of the Life Sciences, Genomics Core Facility-University Park for their help with the plasmid sequencing analyses. This work was supported in part by NIH grant R01-GM087472 to K.A.E. and K.D.M., NSF grant DBI-0965596 to KDM, Penn State Clinical and Translational Sciences Institute, National Science Foundation instrumentation grant OCI–0821527, USDA-AFRI graduate fellowship to A.F., and the Pennsylvania Department of Health using Tobacco CURE Funds (SAP# 4100042746). The Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

FIGURE LEGENDS

Figure 1. Erroneous call rates for Illumina data. The dotted lines represent the 95% confidence intervals of the multinomial sampling. Only repeat lengths with ≥100 read support (all loci combined) were plotted. (**A**) Human male X Chromosome for PCR-containing (PCR+) and PCR-free (PCR-; down-sampled data) library preparation protocols. (**B**) PCR- Mononucleotide STR error rates by error categories. See text for the explanation of the categories. (**C**) PCR- Dinucleotide STR error rates by error categories. (**D**) Error rates for ultra-deep sequencing of plasmids with PCR+ and PCR- protocols.

Figure 2. Factors that affect the accuracy of STR heterozygote genotyping, using the error correction model. (A) STR length. **(B)** The length difference in a heterozygote. **(C)** The ratio of read depths supporting each allele. **(D)** Read depth. All read profiles are generated from $(A)_m$ $(A)_n$ heterozygotes. The X-axis shows different STR length arrays, with the number of reads indicated for each genotype. The Y-axis (confidence of correct prediction) shows the ratio of the probability for a locus to be a heterozygote vs. a homozygote depending on the read length profile. The magnitude of the bar imply more confident in genotyping. The negative value indicates incorrect genotyping. PCR- error profiles from human X Chromosome data were used to calculate probabilities that the read profiles correspond to homozygous or heterozygous loci, which reflects the accuracy of genotyping, since the true genotype is a heterozygote.

Figure 3. Germ-line mutation rate analyses using STR-FM. (A) Cartoon of the pedigree used, with arrows showing branches where mutations were detected. **(B)** The numbers of *de novo* mutations arising in the male and female germ lines. **(C)** The germ-line mutation rates for STRs with different length and motif size. The dotted lines represent the 95% confidence interval of the multinomial sampling. Only repeat lengths with ≥2000 loci support were plotted**.** **(D)** The frequencies of different mutation categories for mononucleotide STRs**.** The dotted lines represent 95% confidence intervals from multinomial sampling. Only repeat numbers with ≥2000 loci support were plotted.

Figure 4. The minimal informative read depth required for correct genotyping of STR heterozygous alleles of consecutive repeat number with success rate of 90%. $(A)_n$ was used for mononucleotide STRs, and $(AC)_n$ – for dinucleotide STRs. All tri- and tetranucleotide

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

TRs were used. (A) PCR-containing library preparation protocol; (B) PCR-free library preparation protocol.

Figure 5. Genome-wide sequencing depth required for 10x informative read depth of 10-50 bp repeats with success rate of 90% using different read lengths.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 1. Evaluation of the STR genotyping model by pseudo-diploid genotyping.

Repeat class	Percentage of correctly determined genotypes from 10,000 simulations			
	PCR+		PCR-	
	Homozygous loci	Heterozygous loci	Homozygous loci	Heterozygous loci
Mononucleotides	98.41	98.73	99.73	99.63
Dinucleotides	99.83	99.73	100.00	99.68
Trinucleotides	99.81	100.00	99.77	99.90
Tetranucleotides	100.00	100.00	100.00	99.98

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Abdul-Muneer PM. 2014. Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. *Genet Res Int* **2014**: 691759.

Abdulovic AL, Hile SE, Kunkel TA, Eckert KA. 2011. The in vitro fidelity of yeast DNA polymerase δ and polymerase holoenzymes during dinucleotide microsatellite DNA synthesis. *DNA Repair* **10**: 497–505.

Ajay SS, Parker SCJ, Abaan HO, Fajardo KVF, Margulies EH. 2011. Accurate and comprehensive sequencing of personal genomes. *Genome Res* **21**: 1498–1505.

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. 2011. Dindel: accurate indel calls from short-read data. *Genome Res* **21**: 961–973.

Ananda G, Hile SE, Breski A, Wang Y, Kelkar Y, Makova KD, Eckert KA. 2014. Microsatellite Interruptions Stabilize Primate Genomes and Exist as Population-Specific Single Nucleotide Polymorphisms within Individual Human Genomes. *PLoS Genet* **10**: e1004498.

Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, Chiaromonte F, Makova KD. 2013. Distinct Mutational Behaviors Differentiate Short Tandem Repeats from Microsatellites in the Human Genome. *Genome Biol Evol* **5**: 606–620.

Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I. 2010. Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. *Bioinforma Oxf Engl* **26**: i420–425.

Baptiste BA, Ananda G, Strubczewski N, Lutzkanin A, Khoo SJ, Srikanth A, Kim N, Makova KD, Krasilnikova MM, Eckert KA. 2013. Mature Microsatellites: Mechanisms Underlying Dinucleotide Microsatellite Mutational Biases in Human Cells. *G3 GenesGenomesGenetics* **3**: 451–463.

Baptiste BA, Eckert KA. 2012. DNA polymerase kappa microsatellite synthesis: Two distinct mechanisms of slippage-mediated errors. *Environ Mol Mutagen* **53**: 787–796.

Blankenberg D, Kuster GV, Bouvier E, Baker D, Afgan E, Stoler N, $author.lastName $author firstName, Taylor J, Nekrutenko A. 2014. Dissemination of scientific software with Galaxy ToolShed. *Genome Biol* **15**: 403.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol Ed Frederick M Ausubel Al* **Chapter 19**: Unit 19.10.1–21.

Brais B, Bouchard JP, Xie YG, Rochefort DL, Chrétien N, Tomé FM, Lafrenière RG, Rommens JM, Uyama E, Nohira O, et al. 1998. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat Genet* **18**: 164–167.

Campbell CD, Eichler EE. 2013. Properties and rates of germline mutations in humans. *Trends Genet* **29**: 575–584.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Castel AL, Cleary JD, Pearson CE. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol* **11**: 165–170.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2014. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*.

Chambers GK, Curtis C, Millar CD, Huynen L, Lambert DM. 2014. DNA fingerprinting in zoology: past, present, future. *Investig Genet* **5**: 3.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, et al. 2012. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. *Cell* **148**: 1293–1307.

Chu CS, Trapnell BC, Murtagh JJ, Moss J, Dalemans W, Jallat S, Mercenier A, Pavirani A, Lecocq JP, Cutting GR. 1991. Variable deletion of exon 9 coding sequences in cystic fibrosis transmembrane conductance regulator gene mRNA transcripts in normal bronchial epithelium. *EMBO J* **10**: 1355–1363.

Cuppens H, Marynen P, Berghe HV den, Cassiman JJ, Boeck CD, Eggermont E, Baets FD. 1990. A child, homozygous for a stop codon in exon 11, shows milder cystic fibrosis symptoms than her heterozygous nephew. *J Med Genet* **27**: 717–719.

Denver DR, Morris K, Kewalramani A, Harris KE, Chow A, Estes S, Lynch M, Thomas WK. 2004. Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of Caenorhabditis elegans. *J Mol Evol* **58**: 584–595.

Eckert KA, Kunkel TA. 1990. High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Res* **18**: 3739–3744.

Eckert KA, Mowery A, Hile SE. 2002. Misalignment-Mediated DNA Polymerase β Mutations: Comparison of Microsatellite and Frame-Shift Error Rates Using a Forward Mutation Assay†. *Biochemistry (Mosc)* **41**: 10490–10498.

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

Frey B, Suppmann B. 1995. Demonstration of the ExpandTM PCR System's Greater Fidelity and Higher Yields with a lacI-based PCR Fidelity Assay. *Biochemica* **2**: 34–35.

Fungtammasan A, Walsh E, Chiaromonte F, Eckert KA, Makova KD. 2012. A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome Res* **22**: 993–1005.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**: 445–477.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, et al. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Res* **15**: 1451 –1455.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Goecks J, Nekrutenko A, Taylor J, $author.lastName $author firstName. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**: R86.

Gryfe R, Di Nicola N, Lal G, Gallinger S, Redston M. 1999. Inherited colorectal polyposis and cancer risk of the APC I1307K polymorphism. *Am J Hum Genet* **64**: 378–384.

Gupta PK, Varshney RK. 2000. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* **113**: 163–185.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**: 1154–1162.

Hagelberg E, Gray IC, Jeffreys AJ. 1991. Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* **352**: 427–429.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* **41**: e32.

Hile SE, Yan G, Eckert KA. 2000. Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. *Cancer Res* **60**: 1698–1703.

Jarne P, Lagoda PJL. 1996. Microsatellites, from molecules to populations and back. *Trends Ecol Evol* **11**: 424–429.

Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. 2010. What Is a Microsatellite: A Computational and Experimental Definition Based upon Repeat Mutational Behavior at A/T and GT/AC Repeats. *Genome Biol Evol* **2**: 620–35.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* **18**: 30–38.

Kim KS, Sappington TW. 2013. Microsatellite data analysis for population genetics. *Methods Mol Biol Clifton NJ* **1006**: 271–295.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475.

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**: 291–295.

Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, Hampel H, Markowitz A, Klimstra D, et al. 1997. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* **17**: 79–83.

Lander E. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Levinson G, Gutman GA. 1987. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. *Nucleic Acids Res* **15**: 5323–5338.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* **26**: 589–595.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl* **25**: 1754–1760.

Lim KG, Kwoh CK, Hsu LY, Wirawan A. 2013. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Brief Bioinform* **14**: 67–81.

Li Y-C, Korol AB, Fahima T, Nevo E. 2004. Microsatellites Within Genes: Structure, Function, and Evolution. *Mol Biol Evol* **21**: 991–1007.

Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. 2012. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS ONE* **7**: e30087.

McIver LJ, Fondon JW, Skinner MA, Garner HR. 2011. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* **97**: 193–199.

Miah G, Rafii MY, Ismail MR, Puteh AB, Rahim HA, Islam KN, Latif MA. 2013. A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. *Int J Mol Sci* **14**: 22499–22528.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS, et al. 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* **23**: 749–761.

Murray V, Monchawin C, England PR. 1993. The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. *Nucleic Acids Res* **21**: 2395–2398.

Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, MacInnis B, Kwiatkowski DP, Swerdlow HP, et al. 2012. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. *BMC Genomics* **13**: 1.

Payseur BA, Jing P, Haasl RJ. 2011. A Genomic Portrait of Human Microsatellite Variation. *Mol Biol Evol* **28**: 303–312.

Pearson CE, Edamura KN, Cleary JD. 2005. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* **6**: 729–742.

Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

Pellegrini M, Renda ME, Vecchio A. 2010. TRStalker: an efficient heuristic for finding fuzzy tandem repeats. *Bioinformatics* **26**: i358–i366.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* **6**: 847–846.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rebolledo-Jaramillo B, Su MS-W, Stoler N, McElhoe JA, Dickins B, Blankenberg D, Korneliussen TS, Chiaromonte F, Nielsen R, Holland MM, et al. 2014. Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc Natl Acad Sci* 201409328.

Roberts RJ, Carneiro MO, Schatz MC. 2013. The advantages of SMRT sequencing. *Genome Biol* **14**: 405.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**: D670–681.

Schneider GF, Dekker C. 2012. DNA sequencing with nanopores. *Nat Biotechnol* **30**: 326–328.

Shinde D, Lai Y, Sun F, Arnheim N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi□likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res* **31**: 974–980.

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.

Sunnucks P. 2000. Efficient genetic markers for population biology. *Trends Ecol Evol* **15**: 199–203.

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

Walsh PS, Fildes NJ, Reynolds R. 1996. Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Res* **24**: 2807–2812.

Wang DY, Chang C-W, Lagacé RE, Oldroyd NJ, Hennessy LK. 2011. Development and Validation of the AmpFℓSTR® Identifiler® Direct PCR Amplification Kit: A Multiplex Assay for the Direct Amplification of Single-Source Samples*,†. *J Forensic Sci* **56**: 835–845.

Wan Q-H, Wu H, Fujihara T, Fang S-G. 2004. Which genetic marker for which conservation genetics issue? *Electrophoresis* **25**: 2165–2176.

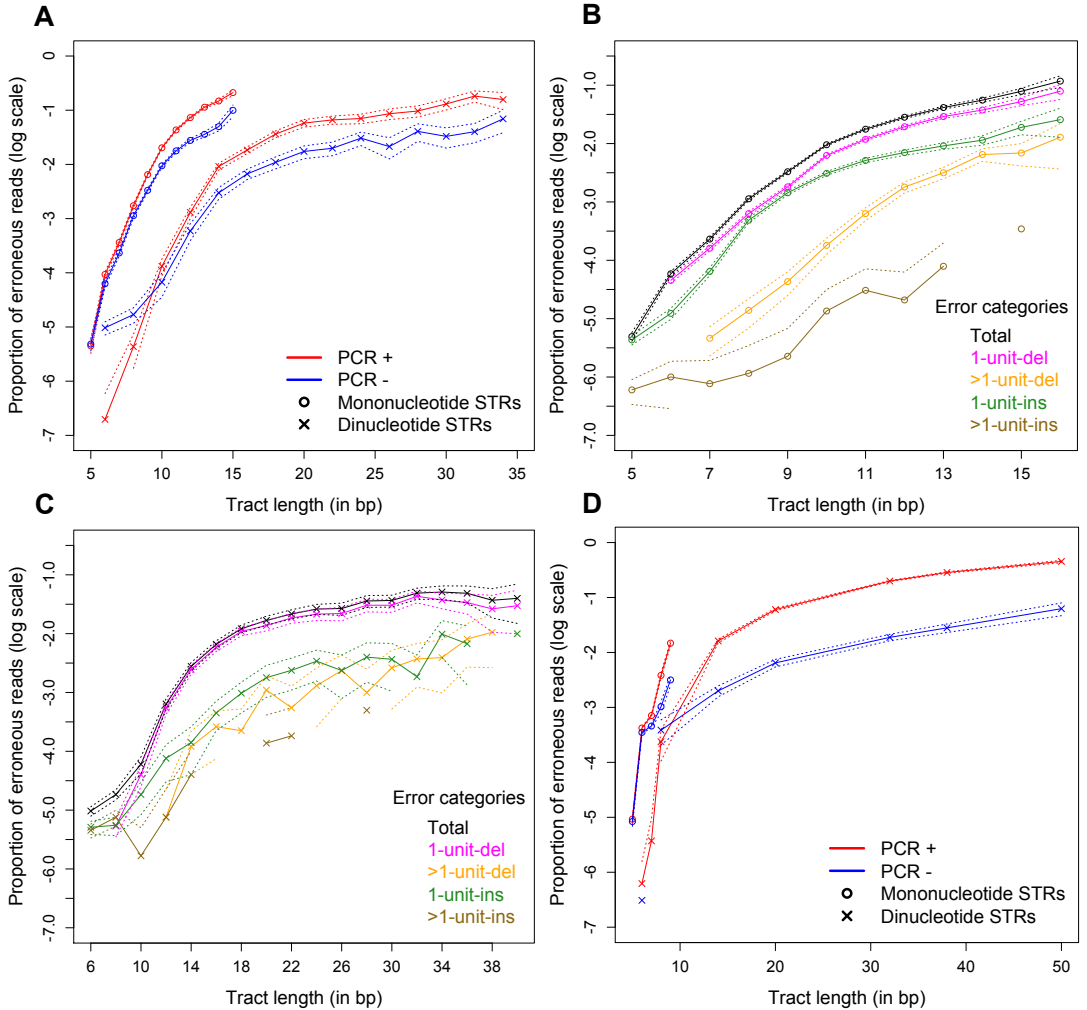
Wilburn B, Rudnicki DD, Zhao J, Weitz TM, Cheng Y, Gu X, Greiner E, Park CS, Wang N, Sopher BL, et al. 2011. An antisense CAG repeat transcript at JPH3 locus mediates

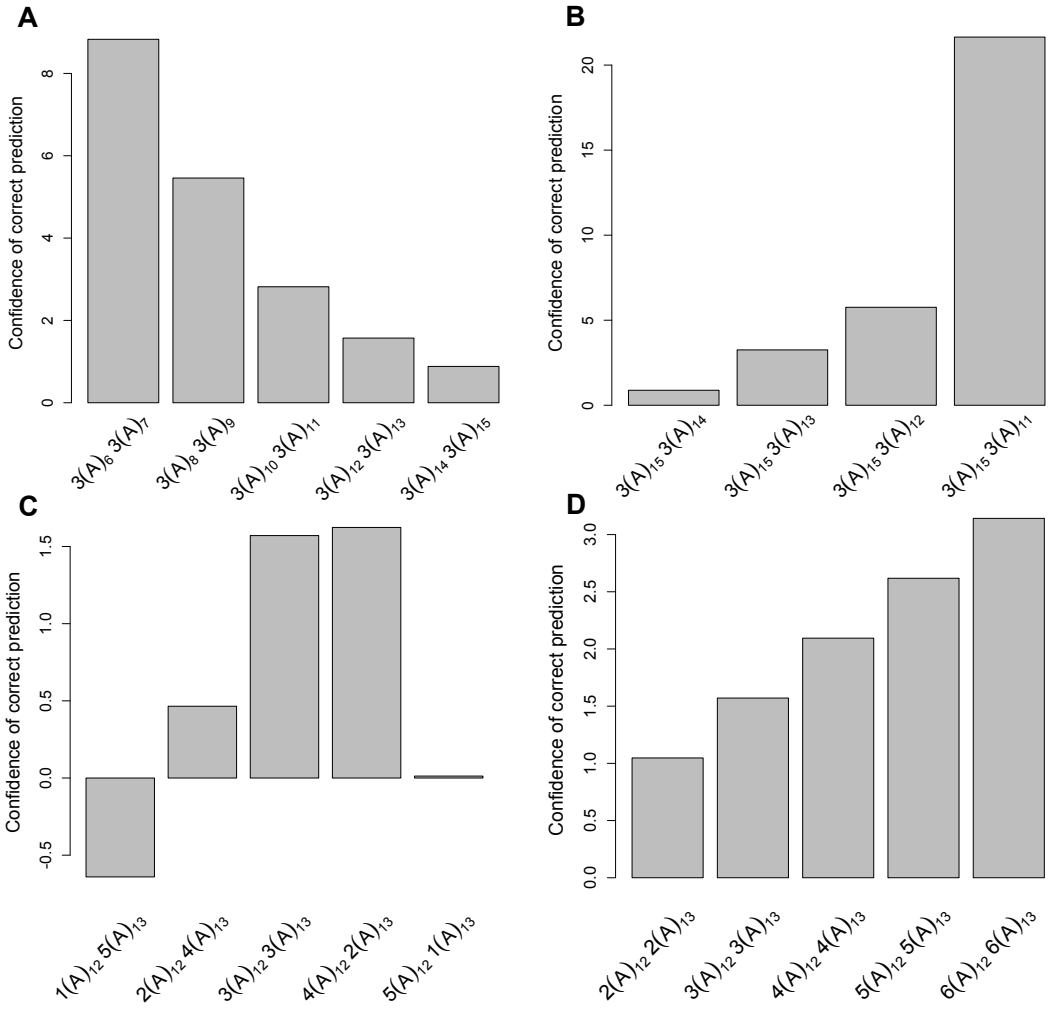
Downloaded from genome.cshlp.org on November 8, 2026 . Published by Cold Spring Harbor Laboratory Press

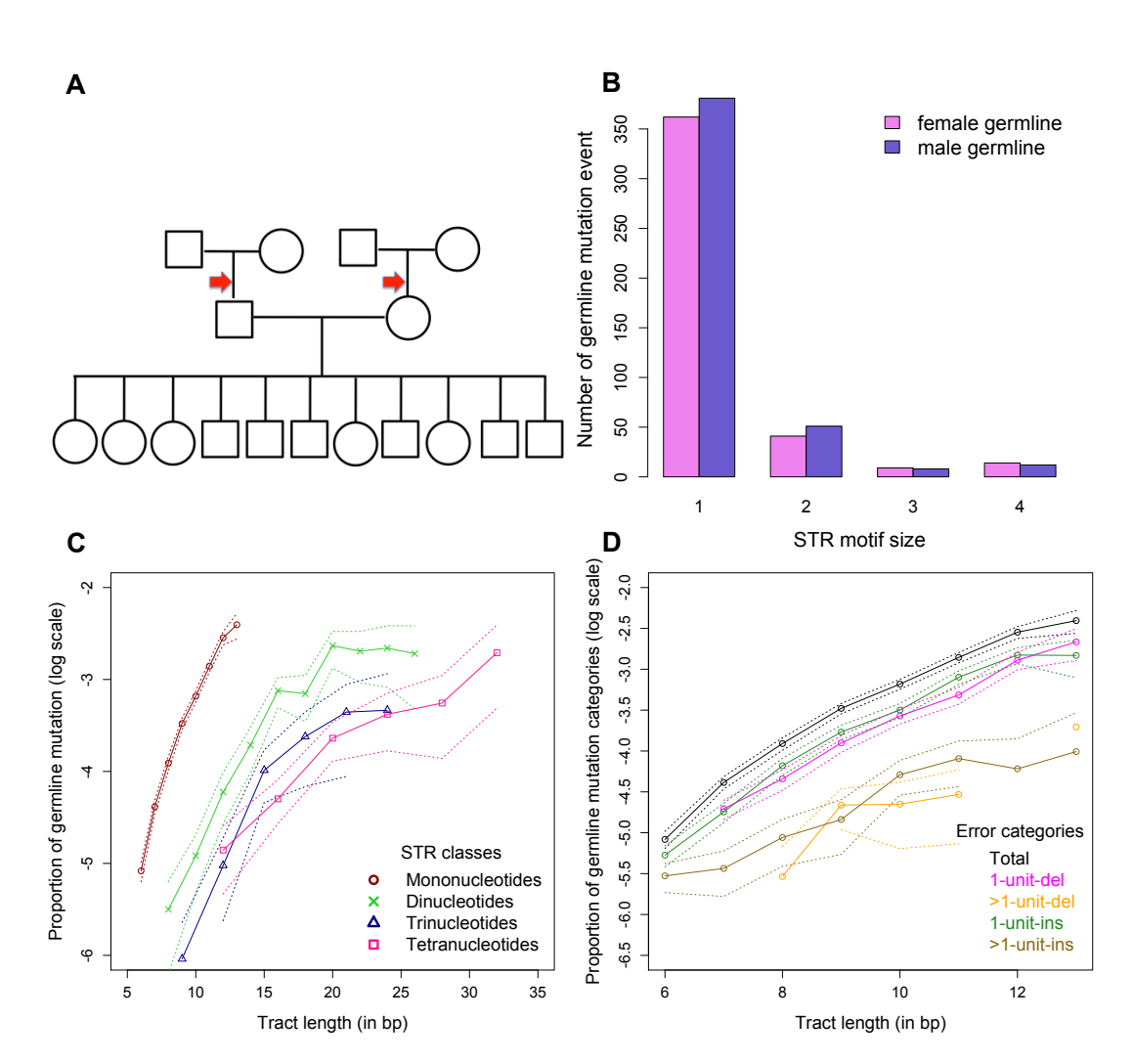
expanded polyglutamine protein toxicity in Huntington's disease-like 2 mice. *Neuron* **70**: 427–440.

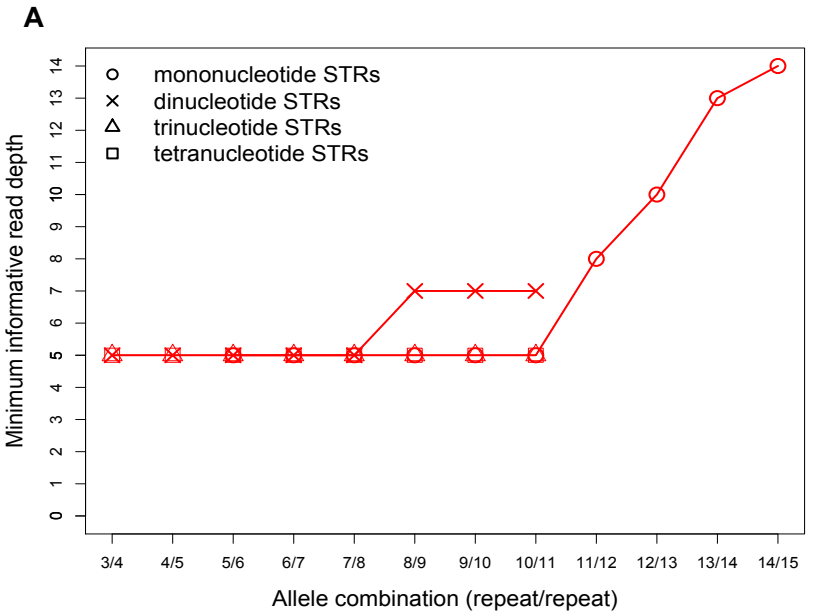
Willems T, Gymrek M, Highnam G, The 1000 Genomes Project Consortium, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res*.

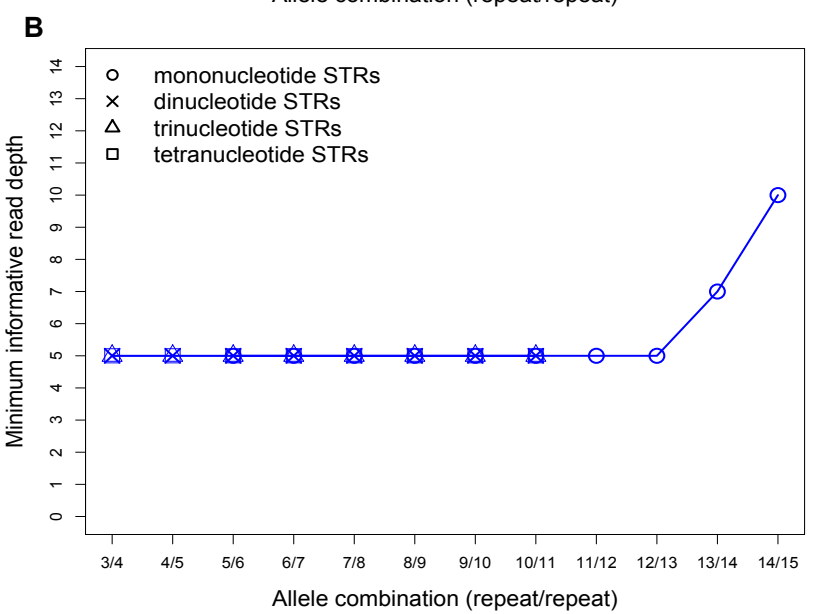
Wright JM, Bentzen P. 1994. Microsatellites: genetic markers for the future. *Rev Fish Biol Fish* **4**: 384–388.

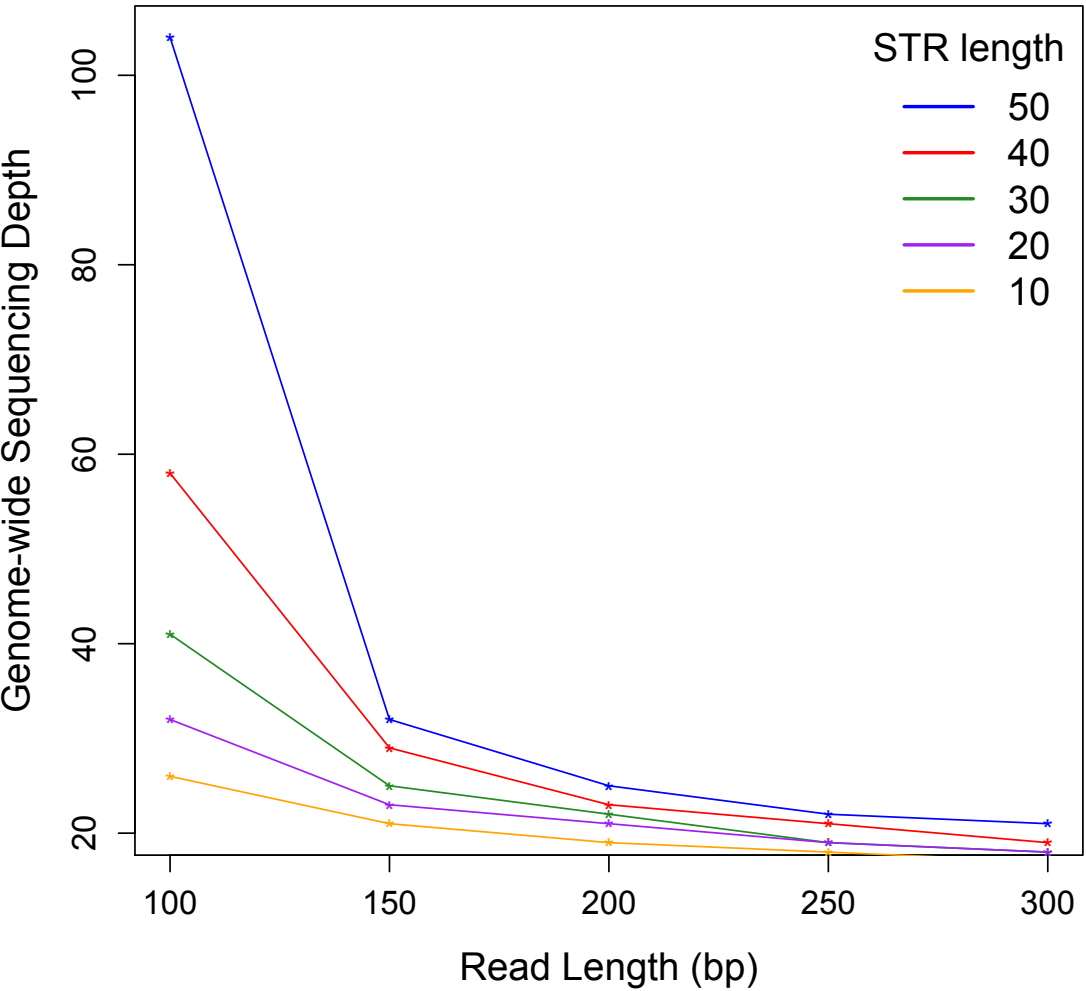

A
Proportion of erroneous reads (log scale)
Tract length (in bp)
PCR +
PCR -
Mononucleotide STRs
Dinucleotide STRs
B
Proportion of erroneous reads (log scale)
Tract length (in bp)
Error categories
Total
1-unit-del
>1-unit-del
1-unit-ins
>1-unit-ins
C
Proportion of erroneous reads (log scale)
Tract length (in bp)
Error categories
Total
1-unit-del
>1-unit-del
1-unit-ins
>1-unit-ins
D
Proportion of erroneous reads (log scale)
Tract length (in bp)
PCR +
PCR -
Mononucleotide STRs
Dinucleotide STRs


A
Confidence of correct prediction
0
2
4
6
8
3(A)6 3(A)7
3(A)8 3(A)9
3(A)10 3(A)11
3(A)12 3(A)13
3(A)14 3(A)15
B
Confidence of correct prediction
0
5
10
15
20
3(A)15 3(A)14
3(A)15 3(A)13
3(A)15 3(A)12
3(A)15 3(A)11
C
Confidence of correct prediction
-0.5
0.0
0.5
1.0
1.5
1(A)12 5(A)13
2(A)12 4(A)13
3(A)12 3(A)13
4(A)12 2(A)13
5(A)12 1(A)13
D
Confidence of correct prediction
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2(A)12 2(A)13
3(A)12 3(A)13
4(A)12 4(A)13
5(A)12 5(A)13
6(A)12 6(A)13


A
B
female germline
male germline
Number of germline mutation event
0
50
100
150
200
250
300
350
1
2
3
4
STR motif size
C
Proportion of germline mutation (log scale)
-2
-3
-4
-5
-6
5
10
15
20
25
30
35
Tract length (in bp)
STR classes
Mononucleotides
Dinucleotides
Trinucleotides
Tetranucleotides
D
Proportion of germline mutation categories (log scale)
-2.0
-2.5
-3.0
-3.5
-4.0
-4.5
-5.0
-5.5
-6.0
-6.5
6
8
10
12
Tract length (in bp)
Error categories
Total
1-unit-del
>1-unit-del
1-unit-ins
>1-unit-ins


A
mononucleotide STRs
dinucleotide STRs
trinucleotide STRs
tetranucleotide STRs
Minimum informative read depth
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
3/4 4/5 5/6 6/7 7/8 8/9 9/10 10/11 11/12 12/13 13/14 14/15
Allele combination (repeat/repeat)


B
mononucleotide STRs
dinucleotide STRs
trinucleotide STRs
tetranucleotide STRs
Minimum informative read depth
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
3/4 4/5 5/6 6/7 7/8 8/9 9/10 10/11 11/12 12/13 13/14 14/15
Allele combination (repeat/repeat)


STR length
50
40
30
20
10
Genome-wide Sequencing Depth
20
40
60
80
100
100
150
200
250
300
Read Length (bp)