SUPPLEMENTAL TEXT

Table of contents

Running time and sensitivity analyses of STR-FM

We analyzed the running time of the STR-FM pipeline on a quad-core AMD Opteron(tm) processor 8384 (Table S1). The running time increases linearly to the number of reads and the number of STRs found in them. For instance, for 10 million single-end 100-bp reads from the human genome, STR-FM detected 1.33 million reads that contain uninterrupted mononucleotide STRs with a length of ≥5 bp in approximately 72 minutes. We also measured the effect of the minimum length threshold of STRs to be detected (Table S2). As short STRs are more abundant in the genome, the running time decreases with a higher length threshold.

The numbers of reads containing STRs that passed through each filtering step are vary depending on STR class, stringency of parameters, redundancy and completeness of the reference genome assembly, and set of STR loci of interest in the reference (Table S3). For the human genome (hg19), 2-14% of STR containing reads were unmapped and 17-23% of mapped reads were not uniquely mapped using BWA with our pipeline. Approximately, 50 percent of reads passed through all filters and can be used for STR profiling (Table S3).

The percentages of STR loci that can be profiled using 30x sequencing depth are 82.32%, 85.93%, 82.12%, and 78.37% for mono-, di-, tri-, and tetra-nucleotide STRs respectively (Table S4). Higher sequencing depth can increase number of detected loci; however, the numbers of loci become saturated when the total percentage are about 80 percent. The percentages of loci that can be profiled also depend on transposable elements at STR loci. The ratios of STRs loci that can be profiled in non-repetitive are higher than those in STRs in *L1* or *Alu* repeat (Table S5).

Theoretically, longest STR repeat that can be profiled from a 100 bp read is 60 bp (requiring 20 bp flanking-bases on either sides). However, due to low sequencing quality scores of long STRs especially mononucleotide STRs (Fig. S2), the longest STR repeat lengths that could be profiled using STR-FM are approximately 22 bp, 50 bp, 55 bp, and 60 bp for mono-, di-, tri-, and tetra-nucleotide STRs respectively.

Supplemental methods

STR-FM algorithm for detecting interrupted STRs

If hamming distance (allow number of interrupted bases in microsatellites) is set to integer more than zero, the program will consider both uninterrupted and interrupted microsatellites. The process works as follows:

(1) Identify intervals that are highly correlated with the interval shifted by *'k'* (the repeat period). These intervals are called "candidates". The allowed level of correlation is 6/7. Depending on whether we want to look for more than one STR, we either find the longest such candidate (simple algorithm) or many candidates (more complicated algorithm). The following steps are then performed on each candidate.

(2) Find the most likely repeat motif in the candidate. This is done by counting all *k*-mers (of length *k*) and choosing the most frequent one. If that *k*-mer is in itself covered by a sub-repeat, we discard this candidate. For instance, we can ignore a 6-mer like ACGACG because we will find it when we are looking for 3-mers.

(3) Once we identify the most likely repeat motif, we then modify the interval, adjusting start and end coordinates to find the interval that has the fewest mismatches vs. a sequence of the motif repeated (hamming distance).

(4) At this point we have a valid STR interval (in the eyes of the program). It is subjected to some filtering stages (hamming distance or too close to an end), and if it satisfies those conditions, it's reported to the user.

Plasmid construction

The construction of all 6-kb 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). Briefly, tandem repeats were inserted in-frame between positions 110 and 111, or 111 and 112, of the HSV-*tk* sense strand by cloning of a small double-stranded fragment created by a polymerase reaction of a primed oligonucleotide including the STR of interest. The 11 plasmids containing artificially inserted tandem repeats that were used in this study are shown in Table S6. Because each of the tandem repeats was inserted into the same backbone vector, each of the 12 plasmids listed in Table S6 shared the exact same sequence with the exception of the inserted tandem repeat. Within the backbone vector, there exists several naturally occurring tandem repeats (Table S7) located throughout the plasmid. These intrinsic tandem repeats were also examined for STR errors among the short NGS sequencing reads.

Preparation of NGS libraries with a PCR-containing protocol

Plasmids were sheared to an average peak fragment length of 550 bp using a Covaris E220 with a peak incident power of 105, duty factor of 5%, and 200 cycles per burst for 100 seconds. Libraries were prepared using 1 ug of sheared plasmid DNA and the TruSeq DNA LT Sample Prep kit with Set A adapters (Illumina) per the manufacturer's protocol. After end-repair, A-tailing, and adapter ligation following the manufacturer's recommendations, 650-bp fragments (the 550-bp plasmid DNA fragment plus adapters) were size-selected by running a 2%-agarose gel (low-range ultra-agarose; BioRad) with SyBrGold nucleic acid stain (Invitrogen) in 1xTAE buffer and purified using the MinElute Gel Extraction kit (Qiagen). Fragments with adapters on both ends were enriched via 10 cycles of PCR per the manufacturer's protocol and purified using AmpPure beads (Beckman Coulter). Libraries were sequenced using 250x250 paired-end reads on a MiSeq. The resulting sequencing reads were deposited at Short Read Archive with an accession number SRP047377.

Preparation of PCR-free NGS libraries

Plasmids were sheared to an average peak fragment length of 550 bp as described above. Libraries were prepared using the TruSeq PCR-Free LT Sample Prep kit with Set A adapters (Illumina) per the manufacturer's protocol. After end-repair, fragments larger than 550 bp were removed by using a diluted mixture of SPRI beads (91 uL beads and 69 uL of H_2O; Beckman Coulter) per 100 uL of sample DNA. Fragments smaller than 550 bp were removed from the resultant supernatant by adding 20 uL of undiluted SPRI beads, providing a bead to DNA sample ratio (0.70) in which lower length fragments will remain in the supernatant and the 550-bp fragment will bind to the beads. After A-tailing, fragments from each plasmid were ligated to the same adapter that they were paired with using the PCR-containing libraries. PCR-free libraries were sequenced on the MiSeq as described above. The resulting sequencing reads were deposited at Short Read Archive with an accession number SRP047377.

Data analysis of the plasmid data

Error profiles of both PCR+ and PCR- plasmid data were estimated using the same pipeline and parameters we used for the X Chromosome data. Briefly, the raw reads from each library were screened for uninterrupted STRs with at least 5, 6, 9, and 12 bp for mono-, di-, tri-, tetra-nucleotide STRs, respectively. The Phred quality scores of all bases in STR regions and 20 bp flanking regions were required to be ≥20. The full lengths of both flanking regions for each STR were mapped to the plasmid reference genome, which is 6,136 bp long. The starting point of plasmid reference was relocated to position 2,751, therefore our artificially inserted STRs were located in the middle of the reference (position 3,494 according to the new coordinate system). The error profiles were calculated from variation of STR length of each clone that mapped to a certain locus. This variation should reflect library preparation and sequencing errors and mutations from clonal propagation. Because both PCR+ and PCR- plasmid data were analyzed following the same procedure, the differences in error profiles between them reflects differences in library preparation.

Construction of artificial heterogeneous genetic samples

The artificial heterogeneous genetic samples were created by mixing DNA from two different plasmids. These plasmids, pGEM-T-Easy-derivative R2 and pGEM-T-Easy-derivative Z1-1, were constructed with PCR-amplified D-loop control region of human mitochondrial DNA, which was extracted from human cell lines CHR and HL-60, respectively. The amplification of D-loop control region of mitochondrial DNAs was carried out with primers L15944 (5'-TCCAAGGACAAATCAGAGAAAAA-3') and H635 (5'-GTTTAGACGGGCTCACATC-3'). The amplicons were then ligated into pGEM-T-Easy vectors according to the procedure provided by the manufacturer (Promega), and the plasmids were extracted and purified using the QIAprep Spin Miniprep Kit (Qiagen).

The mixed clonal DNA samples were fragmented and prepared according to the customized Nextera XT protocol (McElhoe et al. 2014). To validate the fragment size and concentration of adapter-index-attached library molecules, UV spectrophotometry (the Agilent 2100 BioAnalyzer) and quantitative PCR (qPCR) (Library quantitation kits for Illumina sequencing platforms; Kapa BioSystems) were performed. The average length of library molecules was in the range of 400 to 700 base pairs. Libraries were diluted to 2 nM based on the result of qPCR quantification, and twelve libraries were pooled, diluted, and denatured to 17 pM as the final concentration for sequencing loading. Sequences were generated by 250-nucleotide (nt) paired-end reads in Illumina MiSeq sequencer using the 500-cycle v2 Reagent kit. More than 80% of total reads per run passed quality filter and reached high quality score (Phred≥30). The resulting sequencing reads were deposited in SRP047377.

STR-FM on Galaxy

Installation

The installation process can be done as follow:

1. Install and set configuration of local Galaxy

1.1 Download and install Galaxy (https://wiki.galaxyproject.org/Admin/GetGalaxy). Galaxy works on both Unix and Mac OS.

1.2 From your Galaxy directory, add your E-mail as admin E-mail to the Galaxy configuration file. Depending on the Galaxy version, this file can be either universe_wsgi.ini or config/galaxy.ini (https://wiki.galaxyproject.org/Admin/Interface)

1.3 Set directory for tool dependencies (step 2 in https://wiki.galaxyproject.org/Admin/Tools/AddToolFromToolShedTutorial).

1.4 Run local Galaxy from the command line by running ‘sh run.sh’ from your Galaxy directory.

1.5 Open your Galaxy from your browser at address http://localhost:8080 (https://wiki.galaxyproject.org/Admin/GetGalaxy)

1.6 Register using your admin E-mail in the ‘User’ tab on the top.

1.7 Refresh your browser

2. Install tools from and str_fm and dependency tools

Arkarachai Fungtamma..., 4/2/15 8:58 AM
Deleted: microsat_ngs

2.1 From your local galaxy, click ‘Admin’ tab on the top.

2.2 On the left panel, click ‘Search and browse tool sheds’ under ‘Tool sheds’. ‘Accessible Galaxy tool sheds’ will appear on main panel.

2.3 Click on ‘Galaxy main tool shed’ and select ‘Browse valid repositories’. (https://wiki.galaxyproject.org/Admin/Tools/AddToolFromToolShedTutorial)

2.4 Type ‘str_fm in search box and click enter.

Arkarachai Fungtamma..., 4/2/15 8:59 AM
Deleted: microsat’

2.5 The ‘suite_str_fm_0_1’ repository that has ‘arkarachai-fungtammasan’ as the owner will appear. The user may click on this repository name and click ‘Preview and install’.

Arkarachai Fungtamma..., 4/2/15 8:58 AM
Deleted: microsat_ngs_profiling_suite_trfm

The ‘Install to Galaxy’ button will appear on upper right corner. This button allows the user to install all our tools and workflows -- pipelines containing tools for specific purpose such as STR profiling from short read sequencing data, microsatellite detection of the reference genome, and estimating minimum informative read depth. None of our tools have any dependencies. However, some of the other tools that used in our workflows (e.g. SAM flag filter, unique element selection, etc.) are not included in the standard Galaxy installation. For the user’s convenience, we included all dependency tools for the workflows in this repository. Therefore, installing ‘suite_str_fm_0_1’ will be sufficient to operate all workflows we provided.

Arkarachai Fungtamma…, 4/2/15 8:58 AM
Deleted: microsat_ngs_profiling_suite_trfm

2.6 After clicking on ‘Install to Galaxy’ and ‘Install’ button in confirmation page, all our tools, workflows, and test datasets will be downloaded to your local Galaxy. After the download is completed, all our tools will be available on your local Galaxy. If the user wants to use the workflows that we suggested (i.e. STR profiling from short read sequencing data, microsatellite detection of the reference genome, and estimating minimum informative read depth), please proceed to step 3.

2.7 Refresh your browser

3. Install workflows

3.1 Click on the ‘Admin’ tab at the top again.

3.2 On the right panel, click ‘Manage installed tool shed repositories’ under ‘Server’. ‘Installed tool shed repositories’ will appear on main panel.

3.3 Click to open ‘str_fm’ repository.

Arkarachai Fungtamma…, 4/2/15 9:00 AM
Deleted: microsatellite_ngs

3.4 Scroll down to ‘Workflows’ section and select the workflow that you want to install. The SGV graphic of the workflow will appear.

3.5 Click on the ‘Repository Actions’ on the upper right corner and select ‘Import workflow to Galaxy’. If success, the ‘Workflow <workflow name> imported successfully’ will appear. Once the workflow is imported to your Galaxy, you can view and modify it from ‘Workflow’ tab on the top.

Tool description

Our tools in ‘str_fm’ can be used to: (1) profile STRs from short read data with STR-FM pipeline (tools: ‘Microsatellite detection’, ‘Read name modifier’, ‘Fetch flanking bases’, ‘Combine mapped flanked bases’, ‘Check microsatellite motif compatibility’, ‘Select

Arkarachai Fungtamma…, 4/2/15 8:58 AM
Deleted: microsat_ngs

uninterrupted microsatellites'); (2) genotype STRs with error correction (tool 'Correct genotype for microsatellite errors'); (3) estimate the minimum informative read depth from error rates (tools: 'Generate all possible combination of read profile', 'Evaluate the probability of the allele combination to generate read profile', 'Combine the probability to generate read profile'); (4) convert informative read depth to locus-specific and genome-wide sequencing depth (tool 'Convert informative read depth to sequencing depth'). The short description for each tool is provided below.

1. "Microsatellite detection" = Detect STRs from short reads (FASTQ), reference genome (FASTA), or alignments (SAM)

2. "Read name modifier" = Change space in read name to '_' to prevent read name truncation by mapping tools

3. "Fetch flanking bases" = Generate two FASTQ files containing flanking bases around STRs for mapping as faux paired-end reads

4. "Combine mapped flanked bases" = For each mapped faux paired-end reads, infer STR sequence in reference genome between the two mapped ends of the pair

5. "Check microsatellite motif compatibility" = Check if two STRs have the same motif

6. " Select uninterrupted microsatellites" = Select STRs that do not contain an interruption

7. "Correct genotype for microsatellite errors" = Build error correction model from pre-defined error rates and identify most likely genotype of the input data

8. "Generate all possible combinations of read profile" = Use STR error spectrum to generate all possible combinations of read profile at each read depth

9. "Evaluate the probability of the allele combination to generate read profile" = Calculate the probability of a given genotype to generate read profiles (instead of finding most likely genotype like tool number 7)

10. "Combine the probability to generate read profile" = Sum the probability of the given allele combinations to generate read profile at certain read depth

11. "Convert informative read depth to sequencing depth" = Calculate 'locus-specific' and 'genome-wide' sequencing depth from the given informative read depth

The detailed description for each tool is embedded within the tool.

SUPPLEMENTAL TABLES

Table S1. Running time (in seconds) and the number of STRs in 100-bp single-end sequencing reads detected by the STR-FM pipeline for datasets of different sizes. We detected uninterrupted mono-, di-, tri-, and tetra-nucleotide STRs, which have lengths of at least 5, 6, 9, and 12 bp, respectively, in raw DNA reads.

Amount of data		Mono-STRs	Di-STRs	Tri-STRs	Tetra-STRs
1.25 million reads	Detection time (s)	159	149	142	140
	Mapping time (s)	322	68	22	9
	Profiling time (s)	492	244	42	15
	Total time (s)	973	461	206	164
	Number of STRs	179,605	95,126	11,827	3,424
2.5 million reads	Detection time (s)	317	299	285	283
	Mapping time (s)	794	163	47	35
	Profiling time (s)	552	266	46	16
	Total time (s)	1663	728	378	334
	Number of STRs	335,106	188,000	22,976	6,059
5 million reads	Detection time (s)	636	599	575	569
	Mapping time (s)	1387	330	96	60
	Profiling time (s)	620	285	49	16
	Total time (s)	2643	1214	720	645
	Number of STRs	447,010	241,978	28,986	7,681
10 million reads	Detection time (s)	1272	1190	1139	1121
	Mapping time (s)	2174	487	134	74
	Profiling time (s)	925	416	66	20
	Total time (s)	4371	2093	1339	1215
	Number of STRs	1,332,286	650,884	75,534	20,719

Table S2. Running time in seconds for different minimum thresholds of STR lengths.

	Mononucleotide STRs > 5 bp	Mononucleotide STRs > 12 bp
Detection time (s)	636	568
Mapping time (s)	1387	5
Profiling time (s)	620	445
Total time (s)	2643	1018
Number of STRs	447,010	892

Table S3. Range of the percentage of remaining STR containing reads after each filtering step. Twenty exclusive sets each containing 25 million reads PCR- data (Ajay 2011) were tracked for the percentage of STR-containing reads that remain after different filtering steps. The range of initially detected STRs for mono-, di-, tri-, and tetra-nucleotide STRs are 2586985-2700353, 623310-667273, 511882-541487, 155719-162836 repeats.

Percentage of remaining reads	Mono STRs (≥6bp)	Dinucleotide STRs (≥8bp)	Trinucleotide STRs (≥9bp)	Tetranucleotide STRs (≥12bp)
Mapped to reference genome	97.85-98.03	86.05-86.92	94.31-94.97	98.36-98.58
Uniquely mapped	81.05-82.14	70.27-71.47	72.37-73.95	76.10-77.37
Correct read pair orientation and spacing*	69.44-71.3	59.1-61.02	60.31-62.43	58.98-60.79
Concur with the motifs and STR locations in reference genome**	50.23-51.57	40.17-41.71	50.78-52.52	45.42-46.7

* No overlappe of paired-end mapped reads and the distance between paired-end mapped reads is less than read length.

**STR locations in reference genome were also detected using our pipeline. STR loci that are closed to other STR loci than 10 bp were not included.

Table S4. The number and percentage (in parenthesis) of STR loci that can be profiled for each sequencing depth for PCR- data

	Mono STRs (≥6bp)	Dinucleotide STRs (≥8bp)	Trinucleotide STRs (≥9bp)	Tetranucleotide STRs (≥12bp)
All loci	5231663 (100.00%)	1047170 (100.00%)	1328530 (100.00%)	448018 (100.00%)
5x depth	3610847 (69.02%)	741191 (70.79%)	837420 (60.03%)	251032 (50.03%)
10x depth	4058889 (77.58%)	843748 (80.57%)	989611 (74.49%)	309170 (69.01%)
15x depth	4189076 (80.07%)	873856 (83.45%)	1043665 (78.56%)	332601 (74.24%)
20x depth	4240897 (81.06%)	885497 (84.56%)	1065726 (80.22%)	342088 (76.36%)
25x depth	4273314 (81.68%)	892662 (85.25%)	1078064 (81.15%)	346420 (77.32%)
30x depth	4306653 (82.32%)	899839 (85.93%)	1090930 (82.12%)	351108 (78.37%)

Table S5. The number and the proportion of loci that can be profiled (uniquely mapped and passed all the filters) using 30x depth sequencing data in non-repetitive regions, L1, and *Alu*

STR	Non-repetitive regions			L1			*Alu*		
class	all	detected	ratio	all	detected	ratio	all	detected	ratio
Mono	2.4e+06	2.2e+06	0.92	1.3e+06	1.0e+06	0.78	7.2e+05	3.7e+05	0.51
Di	4.9e+05	4.5e+05	0.92	1.7e+05	1.4e+05	0.84	5.1e+04	3.1e+04	0.60
Tri	5.9e+05	5.4e+05	0.91	2.2e+05	1.6e+05	0.74	2.0e+05	1.2e+05	0.58
Tetra	1.1e+05	1.0e+05	0.90	5.7e+04	4.3e+04	0.75	1.3e+05	8.7e+04	0.68

Table S6. The number of STRs on the X Chromosome investigated in the study of NGS error profiles at STRs

	Number of STRs[a]	
	150x PCR+ (Chen 2012)	245x PCR- (Ajay 2011)
Mono	16,256,003	37,891,287
Di	7,734,491	18,183,519
Tri	844,333	2,084,657
Tetra	227,950	551,937

[a] of length ≥ 5 bp, 6 bp, 9 bp, 12 bp for mono-, di-, tri-, and tetra-nucleotide STRs, respectively obtained after all filtering

Table S7. Artificial[a] Tandem Repeat sequences located within plasmid DNA

Tandem Repeat motif	Length, units	Sequence Context	Number of informative reads in PCR+ data	Number of informative reads in PCR- data
None[b]	n.a.	GCG CGT TCT CGC		
Mononucleotide				
[A/T][c]	8	GCG CG**A [A]$_6$ A**CT CGA	39,791	27,500
[T/A]	5	GCG CG**T [T]$_3$ T**CT CGA	29,022	36,643
	8	GCG CG**T [T]$_6$ T**CT CGA	40,450	36,531
[G/C]	6	GCG C**GG [G]$_3$ G**CT CGA	28,349	24,429
	9	GCG C**GG [G]$_6$ G**CT CGA	22,805	25,627
Dinucleotide				
[GT/CA]	4	GCG C**GT [GT]$_3$** TCT CGA	37,897	38,839
	7	GCG C**GT [GT]$_6$** TCT CGA	38,495	36,519
	10	GCG C**GT [GT]$_9$** TCT CGA	18,643	24,611
	16	GCG C**GT [GT]$_{15}$** TCT CGA	23,796	10,813
	19	GCG C**GT [GT]$_{18}$** TCT CGA	9,127	4,248
	25	GCG C**GT [GT]$_{24}$** TCT CGA	1,853	815

[a] Artificial tandem repeats are those that were inserted in-frame between positions 110 – 111 or 111 – 112 of the HSV-*tk* gene; see Methods.

[b] Control plasmid containing no artificial tandem repeat.

[c]The first sequence is the tandem repeat motif located on the HSV-*tk* sense strand and the second sequence is the tandem repeat motif located on the HSV-*tk* antisense strand.

Table S8. Intrinsic[a] Tandem Repeat sequences located within plasmid DNA.

Tandem Repeat	Length[b], bp	Number of Occurrences	Number of informative reads in PCR+ data	Number of informative reads in PCR- data
Mononucleotide				
[A/T]	5	19	9,722,111	9,077,482
	6	5	3,012,065	2,820,945
	7	4	1,886,222	1,699,131
[G/C]	5	15	4,962,104	4,617,951
	6	5	1,379,195	1,242,968
	7	1	114,635	73,602
Dinucleotide				
[AT/TA]	6	3	1,247,469	990,649
	7	1	623,578	606,127
[AC/GT]	6	4	998,155	853,332
[CG/GC]	6	6	1,505,874	1,404,656
Trinucleotide				
[ACC/TGG]	11	1	18,880	4,769

[a] Intrinsic tandem repeats are tandem repeats that are naturally occurring within the plasmid DNA sequence.

[b] The minimum length of TRs used for this table is 5bp, 6bp, 9bp, and 12bp for mono-, di-, tri-, and tetranucleotide TRs, respectively.

Table S9: Number of reads for each motif and STR length in PCR+ analysis.

(See a separate file)

Table S10: Number of reads for each motif and STR length in PCR- analysis.

(See a separate file)

Table S11. Logistic regression models contrasting correct and incorrect predictions for mononucleotide STRs.

	Predictor	Coefficient [a]	VIF [b]	P-value	Relative contribution [c]
PCR+ heterozygous data	Informative read depth	0.09	2.79	3.37e-06	7.38
	STR length [d]	- 0.72	1.07	< 2e-16	55.21
	Length difference [e]			NS	
	Coverage balance [f]	6.96	2.74	3.78e-06	6.88
	Pseudo R-squared [g]				26.27
PCR- heterozygous data	Informative read depth			NS	
	STR length [d]	- 0.81	2.29	2.0e-09	13.74
	Length difference [e]	3.51	1.75	4.9e-11	45.47
	Coverage balance [f]	76.81	1.74	2.3e-13	85.12
	Pseudo R-squared [g]				65.72
PCR+ homozygous data	Informative read depth	+0.14	1.37	9.81e-7	4.87
	STR length	- 0.65	1.37	<2e-16	23.27
	Pseudo R-squared [g]				37.515
PCR- homozygous data	Informative read depth	+ 0.25	NA	9.47e-07	100
	STR length	-		NS	
	Pseudo R-squared [g]				33.55

[a] Signs of coefficients represent the directional effect of predictors. A positive coefficient means that the higher the value of a predictor, the more likely the model predicts genotype correctly.

[b] VIF (Variance Inflation Factor) measures the autocorrelation among predictors. In this analysis all autocorrelations are low.

[c] Relative Contribution is the level of contribution of each predictor to the predictive power of the regression model (see Methods). Predictors with higher values are more influential to the model.

[d] In case of heterozygotes, the longer STR allele was used.

[e] Difference in the lengths of STR alleles in heterozygote (bp).

[f] Ratio of reads from two alleles at heterozygous loci. The frequency of the allele that had fewer reads was used; therefore, this predictor has value between 0-0.5. A higher value means higher evenness of read depth from both alleles.

[g] Pseudo R-squared is the ratio of deviance that was explained by the regression model compared to the null model.

NS= Not significant

NA= Not applicable

Table S12. The detection of a minor allele for mixed DNA samples.

Mixing ratio	Replicate 1		Replicate 2	
(C_8: C_7)	Informative read depth	Minor allele detection	Informative read depth	Minor allele detection
98.00: 2.00	654	detected	848	detected
99.00: 1.00	614	detected	1652	detected
99.50: 0.50	292	not detected	1134	detected
99.75: 0.25	1220	not detected	1300	detected
99.90: 0.10	438	detected	1924	not detected

Table S13. Genotype of disease-related long trinucleotide STRs of NA 12882 (Ajay et al. 2011). Genotypes were inferred from 245x depth PCR free data using our genotyping model. List of trinucleotide repeat diseases came from Table S1 of Castel et al. 2010. Only pure (uninterrupted) STRs with at least one previously identified (in other studies) allele shorter than 60 bp (this is the maximum detection range for 100-bp reads) were included. Genotype columns in gray are loci that have less than 5x informative read depth, thus the confidence of sampling both allele is less than 90%.

Motif	Genomic coordinates	Gene	Part of gene	Disease association (Castel 2010)	Normal length (bp) (Castel 2010)	Disease length (bp) (Castel 2010)	genotype(bp)	Informative read depth (the number of reads)
CAG	chr16:87637893 - 87637935	*JPH3*	CDS of antisense to *JPH3* (Wilburn 2011)	Huntington’s disease-like 2	18-81	153-171	Homozygote 48	10
AGC	chr13:36050433 - 36050490	*MAB21L1*	5’UTR	No confirmed disease association	18-93	>150	Heterozygote 29,41	42
CCG	chr7:103629803-103629827	*RELN*	5’UTR	Risk of Autism	24-30	36-39	Homozygote 32	17
CCG	chr10:95462279 - 95462303	*FRA10AC1*	5’UTR	No confirmed disease association	24-42	>600	Homozygote 23	18
GAC	chr19:18896844-18896859	*COMP*	CDS	Multiple Skeletal dysplasias	15	12, 18, 21	Homozygote16	44
CAG	chr1:154842199 - 154842241	*KCNN3*	CDS	No confirmed disease association	21-84	NA	One allele is 55	1
CAG	chr4:3076603 - 3076660	*HTT*	CDS	Huntington’s disease	30-102	>105	One allele is 50	1
CAG	chr12:7045891 - 7045936	*ATN1*	CDS	Dentatorubral-pallidouysian atrophy	21-75	147-264	Homozygote 47 or Heterozygote 44,47	4
CCG	chr11:119076999 - 119077032	*CBL*	5’UTR	Jacobsen syndrome	33	300-3000	One allele is 34	4
CAG	chr19:13318672 - 13318711	*CACNA1A*	CDS	Spinocerebellar ataxia 6	12-54	60-87	One allele is 34	2

Table S14. Germ-line mutation rates detected by 1) a putative mutant allele absents in the second parent (Fig. S12A), 2) the other allele in first parent is not the same as child's non-mutant allele (Fig. S12B), or both criteria.

	Criteria 1 (Fig. S12A)	Criteria 2 (Fig. S12B)	Union
Male germ-line mutation	231	446	452
Female germ-line mutation	234	423	426
Both	9	11	11
Unknown	469	135	581

Table S15. Germ-line mutation rates binned to the nearest full repeats

STR	Number of mutations	Number of transmissions	Mutation rate per locus per generation
Mononucleotide			
6 bp	67	8083617	8.29e-06
7 bp	124	3007897	4.12e-05
8 bp	127	1029624	1.23e-04
9 bp	183	550843	3.32e-04
10 bp	206	312632	6.59e-04
11 bp	190	136173	1.40e-03
12 bp	141	49609	2.84e-03
13 bp	40	10169	3.93e-03
14 bp	5	963	5.19e-03
15 bp	1	51	1.96e-02
All repeat length[1]	1191	13181908	9.04e-05
Dinucleotide			
8 bp	5	1581659	3.16e-06
10 bp	10	826281	1.21e-05
12 bp	11	184146	5.97e-05
14 bp	14	72480	1.93e-04
16 bp	26	34344	7.57e-04
18 bp	12	17137	7.00e-04
20 bp	21	9019	2.33e-03
22 bp	11	5410	2.03e-03
24 bp	8	3649	2.19e-03
26 bp	4	2094	1.91e-03
28 bp	6	1160	5.17e-03
30 bp	4	487	8.21e-03
All repeat length[1]	171	2738156	6.25e-05
Trinucleotide			
9 bp	2	2185400	9.15e-07
12 bp	4	420717	9.51e-06
15 bp	9	87351	1.03e-04
18 bp	7	29251	2.39e-04
21 bp	5	11319	4.42e-04
24 bp	2	4324	4.63e-04
27 bp	4	1687	2.37e-03
30 bp	1	643	1.56e-03
All repeat length[1]	38	2741124	1.39e-05
Tetranucleotide			
12 bp	6	429316	1.40e-05
16 bp	13	258154	5.04e-05
20 bp	16	69628	2.30e-04
24 bp	10	23899	4.18e-04
28 bp	4	7220	5.54e-04
32 bp	4	2050	1.95e-03
36 bp	1	561	1.78e-03
40 bp	0	185	0.00e+00

44 bp	1	118	8.47e-03
48 bp	1	45	2.22e-02
All repeat length[1]	70	791204	8.85e-05

[1] All repeat length row also includes cases that the actual length cannot be determined. For example, if father has A_7A_6, mother has A_8A_8, and kids have A_8A_8, we will not know if it is a male germ-line mutation of A_7-> A_8 or A_6 ->A_8

Table S16. Minimum informative sequencing depth to detect heterozygotes with probability of at least 0.9. Only the allele combinations in which both alleles have reliable error rates (as least 500 reads were used to calculate error rates) are shown. For trinucleotide STRs, allele combinations that were estimated from smaller number of read (<500 reads) are also shown (as marked with asterisk). Also, all trinucleotide STR motifs were combined.

	PCR+						**PCR-**					
Allele combination by repeat number	**A**	**C**	**AT**	**AC**	**AG**	**All tri-STR**	**A**	**C**	**AT**	**AC**	**AG**	**All tri-STR**
3/4			**5**	**5**	**5**	**5**			**5**	**5**	**5**	**5**
4/5			**5**	**5**	**5**	**5**			**5**	**5**	**5**	**5**
5/6	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
6/7	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**	**5**
7/8	**5**	**5**	**7**	**5**	**5**	**5***	**5**	**5**	**5**	**5**	**5**	**5**
8/9	**5**	**5**		**7**	**7**	**5***	**5**	**5**	**5**	**5**	**5**	**5**
9/10	**5**			**7**		**5***	**5**		**5**	**5**	**5**	**5***
10/11	**5**			**7**		**5***	**5**			**5**		**5***
11/12	**8**						**5**			**5**		**5***
12/13	**10**						**5**			**5**		
13/14	**13**						**7**			**7**		
14/15	**14**						**10**			**7**		
15/16							**10**			**7**		

Table S17. Minimum informative sequencing depth to detect heterozygotes with probability of at least 0.95. Only the allele combinations in which both alleles have reliable error rates (as least 500 reads were used to calculate error rates) are shown. For trinucleotide STRs, allele combinations that were estimated from smaller number of read (<500 reads) are also shown (as marked with asterisk). Also, all trinucleotide STR motifs were combined.

	PCR+						PCR-					
Allele combination by repeat number	A	C	AT	AC	AG	All tri-STR	A	C	AT	AC	AG	All tri-STR
3/4			6	6	6	6			6	6	6	6
4/5			6	6	6	6			6	6	6	6
5/6	6	6	6	6	6	6	6	6	6	6	6	6
6/7	6	6	8	6	6	6	6	6	6	6	6	6
7/8	6	6	8	8	8	6*	6	6	6	6	6	6
8/9	6	6		8	10	6*	6	6	6	6	6	6
9/10	6			8		6*	6		8	6	6	6*
10/11	9			8		6*	6			6		6*
11/12	11						9			8		6*
12/13	14						9			8		
13/14	20						11			8		
14/15	23						11			8		
15/16							14			11		

Table S18. Number of dinucleotide STRs in genome by motif and repeat length

	AC	AG	AT	CG
Total loci (>6 bp)	2572569	3215981	2107265	60759
loci longer than 10 bp	115974	85881	64960	729
loci longer than 20 bp	34983	5272	9534	3
loci longer than 30 bp	23438	1661	4867	0
loci longer than 40 bp	10786	665	2866	0

SUPPLEMENTAL FIGURES

Figure S1. Overview of the STR-FM pipeline. The left panel shows the overall process of the STR-FM pipeline, while the right panel shows the status of the reads that correspond the left schematic

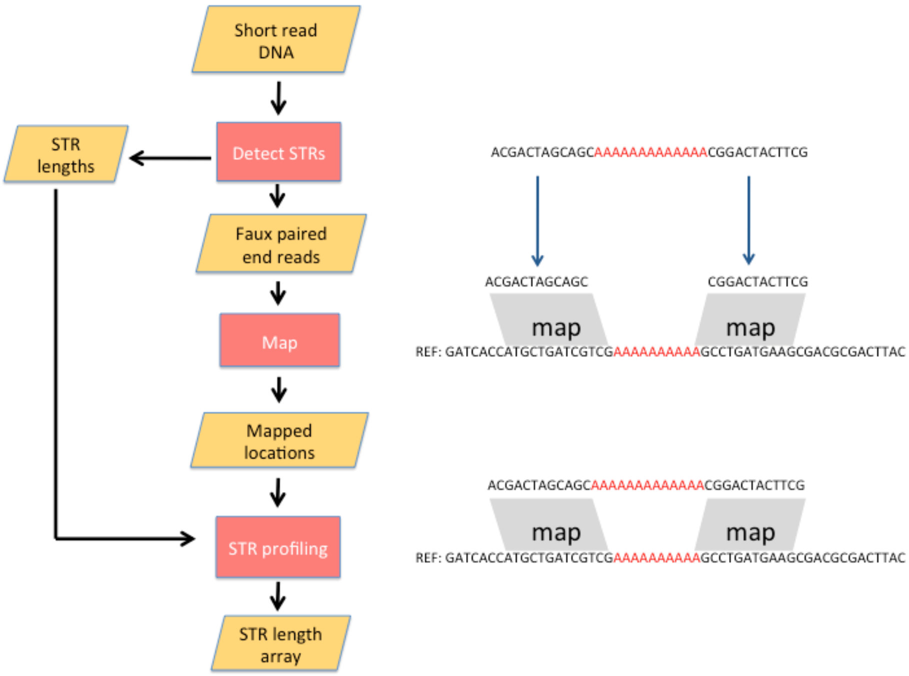
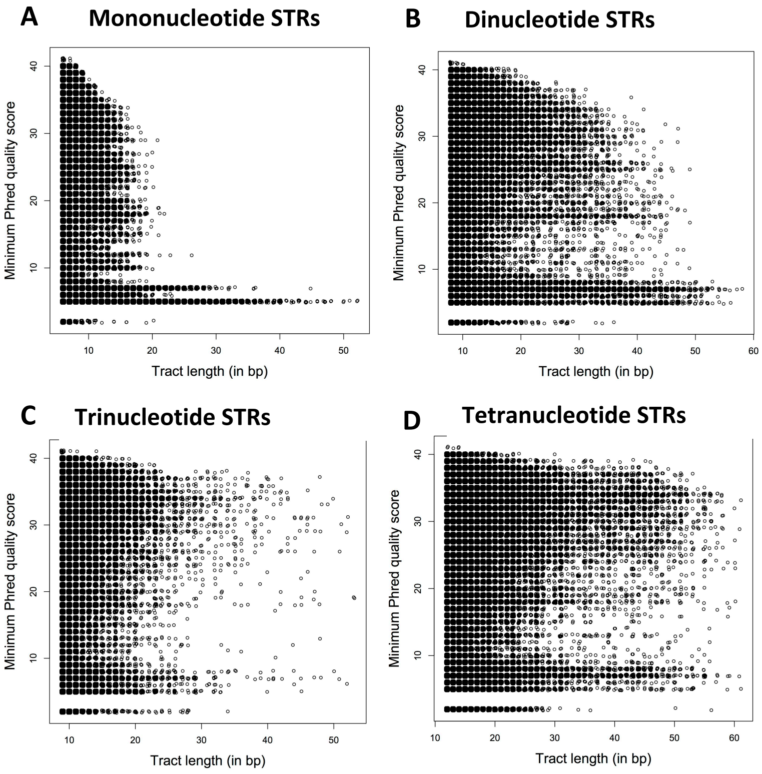

Short read DNA
Detect STRs
STR lengths
Faux paired end reads
Map
Mapped locations
STR profiling
STR length array
ACGACTAGCAGCAAAAAAAAAAAAAACGGACTACTTCG
ACGACTAGCAGC
CGGACTACTTCG
map
map
REF: GATCACCATGCTGATCGTCGAAAAAAAAAAAGCCTGATGAAGCGACGCGACTTAC
ACGACTAGCAGCAAAAAAAAAAAAAACGGACTACTTCG
map
map
REF: GATCACCATGCTGATCGTCGAAAAAAAAAAAGCCTGATGAAGCGACGCGACTTAC

Figure S2. Scatterplot of minimum Phred sequencing quality score of all bases in STRs and 20 bp upstream and downstream flanking regions by repeat length. Two hundreds thousand repeats were plotted for each STR class. Random noise in coordinated were applied by jitter function in R to reduce overlapped plotting. Each circle in the plot represents one repeat. (A) Mononucleotide STRs. (B) Dinucleotide STRs. (C) Trinucleotide STRs. (D) Tetranucleotide STRs.



A Mononucleotide STRs

B Dinucleotide STRs

C Trinucleotide STRs

D Tetranucleotide STRs

Figure S3. **Erroneous call rates between Illumina sequencing with PCR (PCR+) and without PCR (PCR-) for tri- and tetranucleotide STRs and human male X Chromosome data.** Only repeat lengths with ≥100 reads support are plotted. The dotted lines represent the 95% confidence interval of the multinomial sampling. **(A)** Trinucleotide STRs. **(B)** Tetranucleotide STRs.

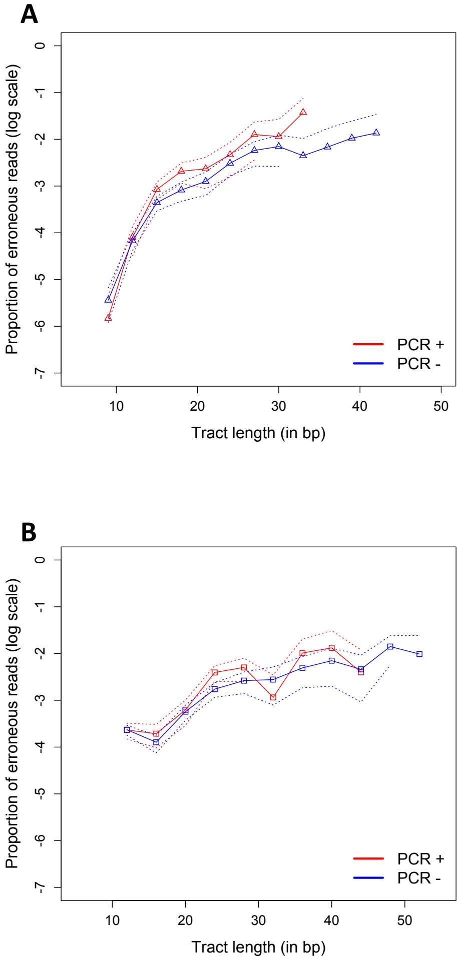


Figure S4. **Erroneous call rates for PCR+ saliva and PCR+ blood data.** Only repeat lengths with ≥100 read support are plotted. The dotted lines represent the 95% confidence interval of the multinomial sampling.

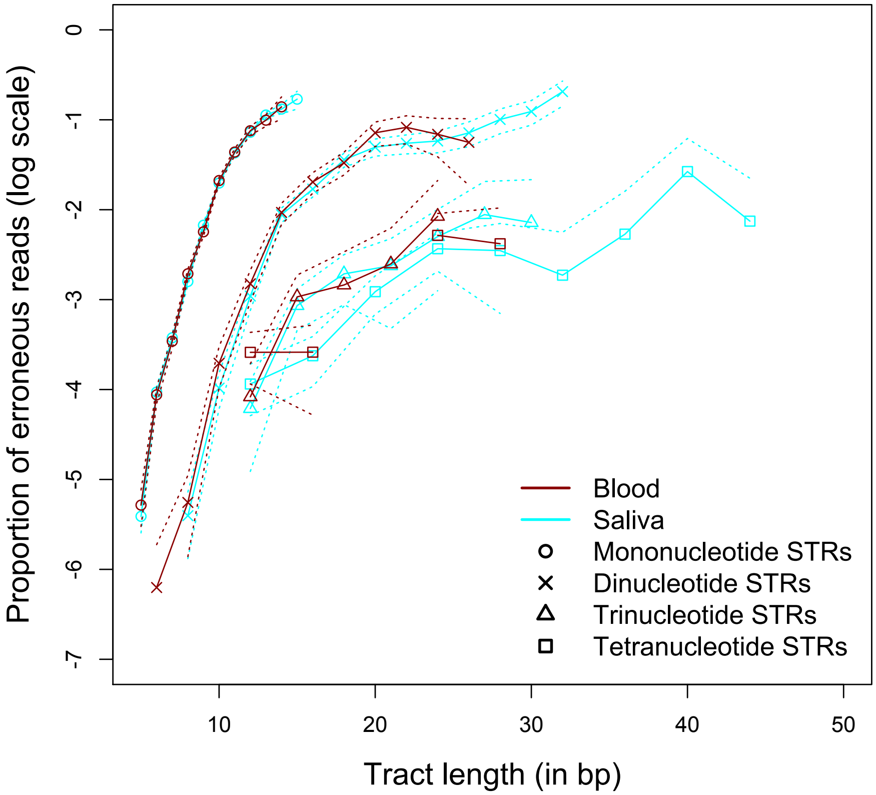


Figure S5. Erroneous call rates for PCR+ data and full PCR- data. Only repeat length with ≥100 read support were plotted. The dotted lines represent the 95% confidence interval of the multinomial sampling.

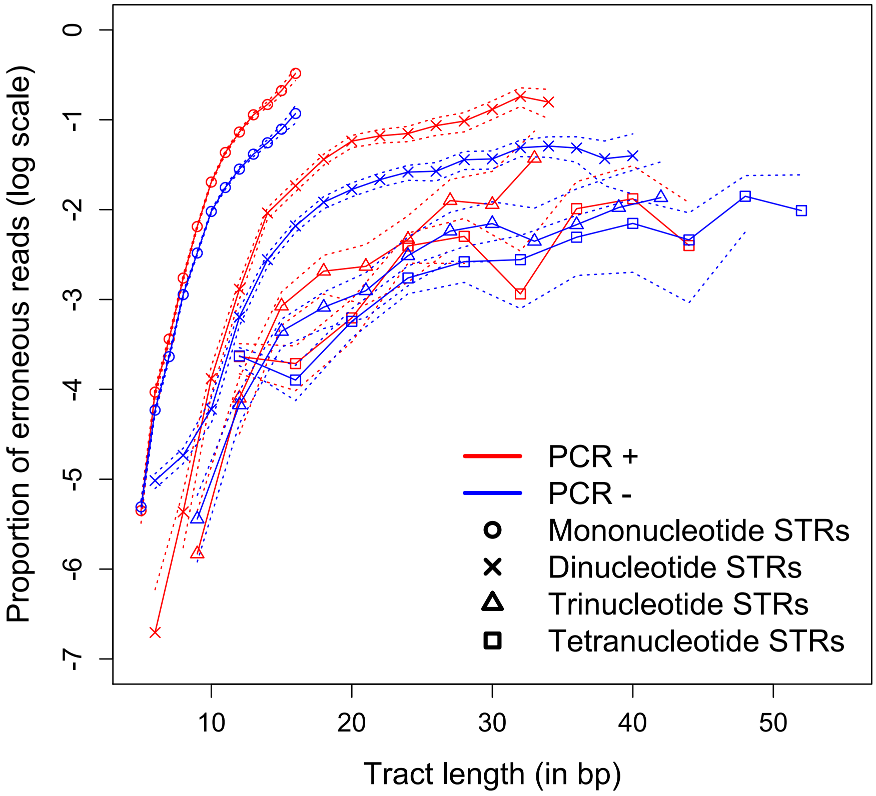


Figure S6. Erroneous call rates of PCR+ data by STRs error category and repeat numbers. Only STR repeat numbers with ≥100 reads (all loci combined) are plotted. The dotted lines represent the 95% confidence intervals of the multinomial sampling. (A) Mononucleotide STRs. (B) Dinucleotide STRs. (C) Trinucleotide STRs. (D) Tetranucleotide STRs.

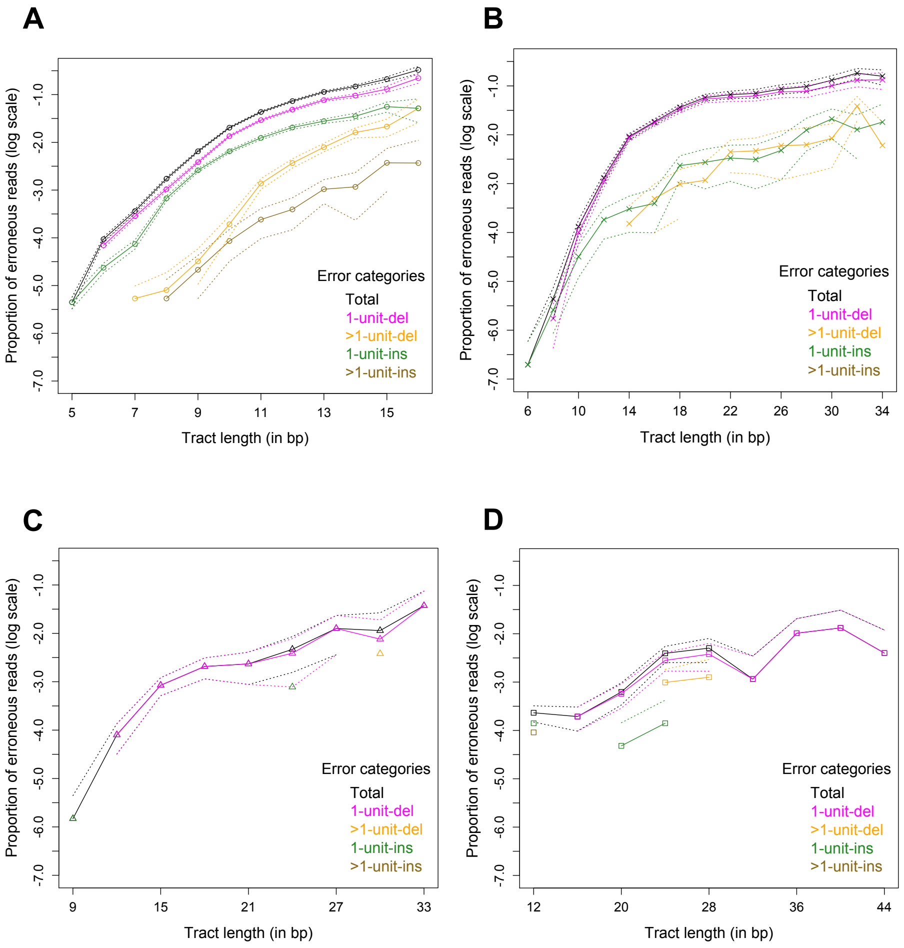


Figure S7. The erroneous call rates for PCR- data by STR error category and repeat number. Only STR repeat numbers with ≥100 reads (all loci combined) are plotted. Dotted lines represent the 95% confidence intervals from multinomial sampling. (A) Trinucleotide STRs. (B) Tetranucleotide STRs.

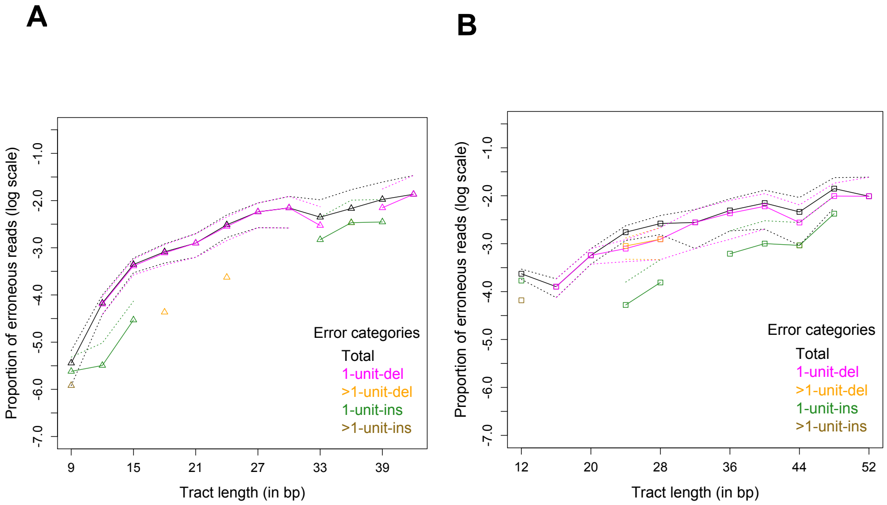


Figure S8. Erroneous call rates by STR motif and repeat number. Only STR repeat numbers with ≥100 reads (all loci combined) are plotted. Dotted lines represent the 95% confidence intervals from multinomial sampling. (A) PCR+ Mononucleotide STRs. (B) PCR+ Dinucleotide STRs. (C) PCR- Mononucleotide STRs. (D) PCR- Dinucleotide STRs.

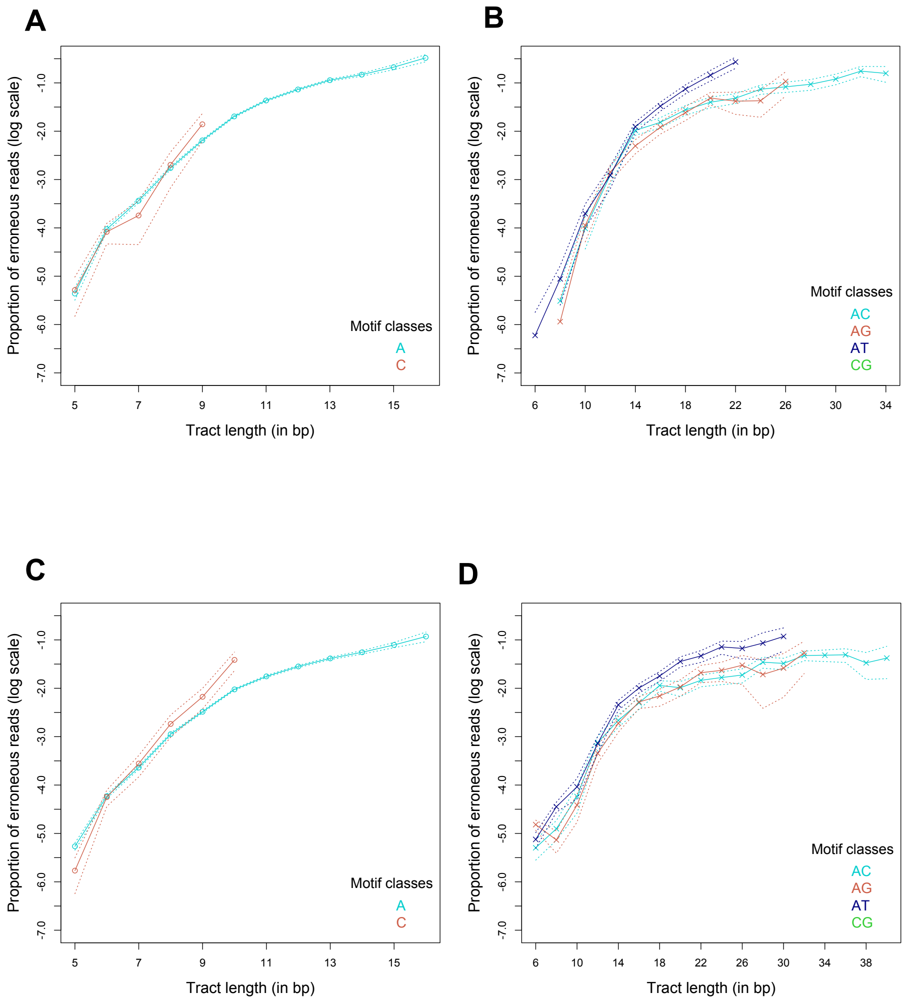


Figure S9. The erroneous call rates for plasmid analysis by STR error category and repeat number. Only STR repeat numbers with ≥100 reads (all loci combined) are plotted. Dotted lines represent the 95% confidence intervals from multinomial sampling. (A) PCR+ mononucleotide STRs. (B) PCR+ dinucleotide STRs. (C) PCR- mononucleotide STRs. (D) PCR- dinucleotide STRs.

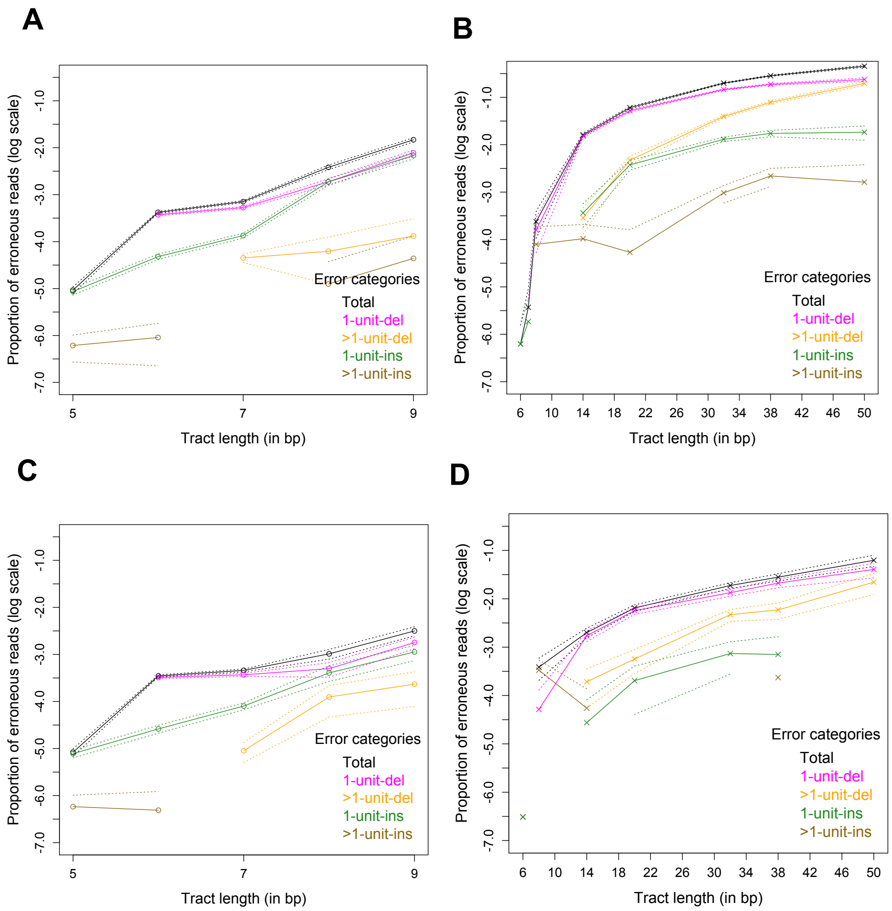


Figure S10. The percentage of heterozygous loci in NA12882 (Ajay 2011) by STR motif size and repeat number. Only repeat number with at least 1000 loci support were plotted.

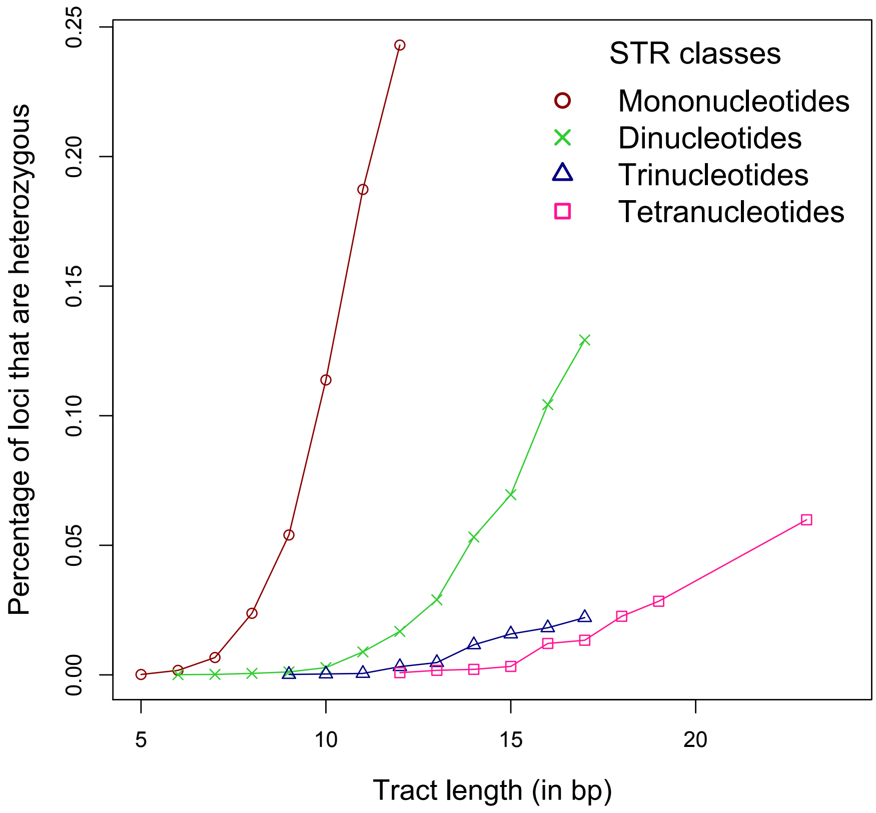


Figure S11. The absolute frequency of heterozygotes depending on the length difference between their alleles. X-axis is the length difference in bp. Y-axis is the number of loci in NA12882 (Ajay 2011).

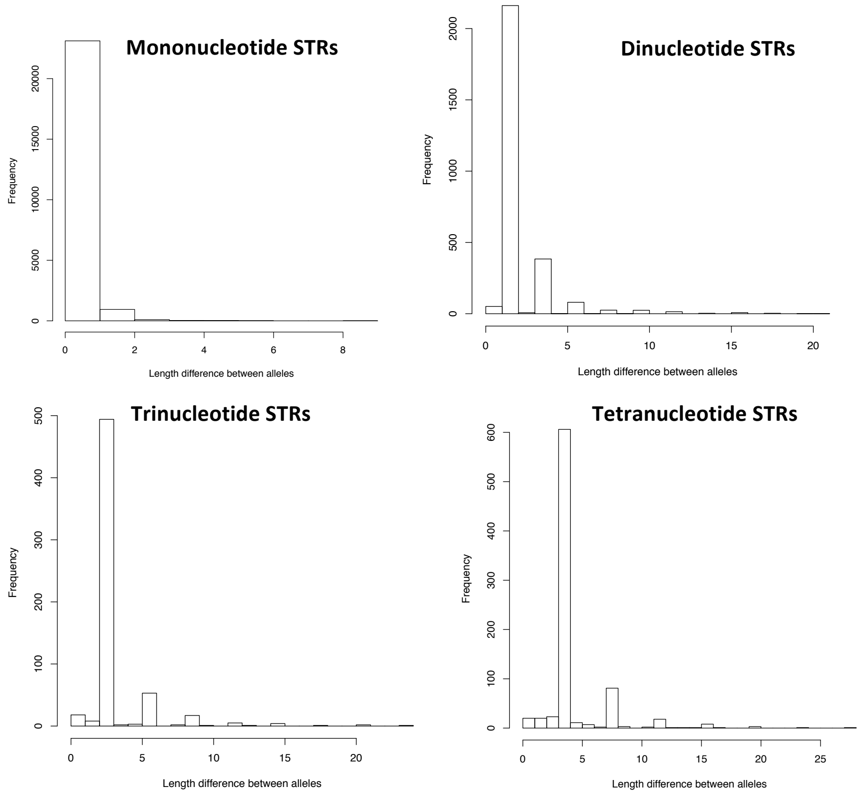


Figure S12. Examples of germ-line mutations in grandparental germ line that can be verified by transmission to grandchildren. Pairs of numbers are repeat numbers of STR alleles. Mutant alleles are in red. (A) Mutant allele does not exist in the other parent, thus it is transmitted from the mutant parent. (B) Non-mutant allele in the child exists only in the other parent, thus the mutant allele is transmitted from the mutant parent.

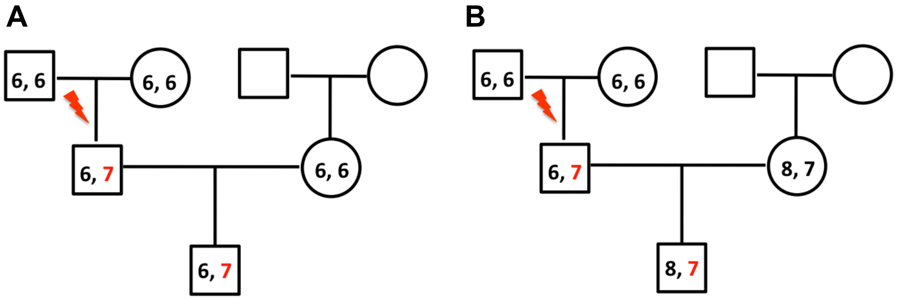


Figure S13. Examples of germ-line mutations in grandparental germ line that failed to be verified by transmission to grandchildren. Pairs of numbers are repeat numbers of STR alleles. Mutant alleles are in red. (A) Mutant allele is not present in a child. (B) STR with the same length as mutant allele present in both parents. In this scenario, only child that have genotype 7,7 can verify transmission of mutant allele.

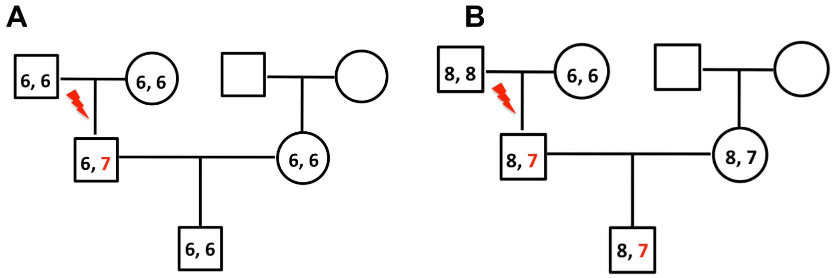


Figure S14. Male and female germ-line mutations by STR motif. (A) mononucleotide STRs. (B) Dinucleotide STRs.

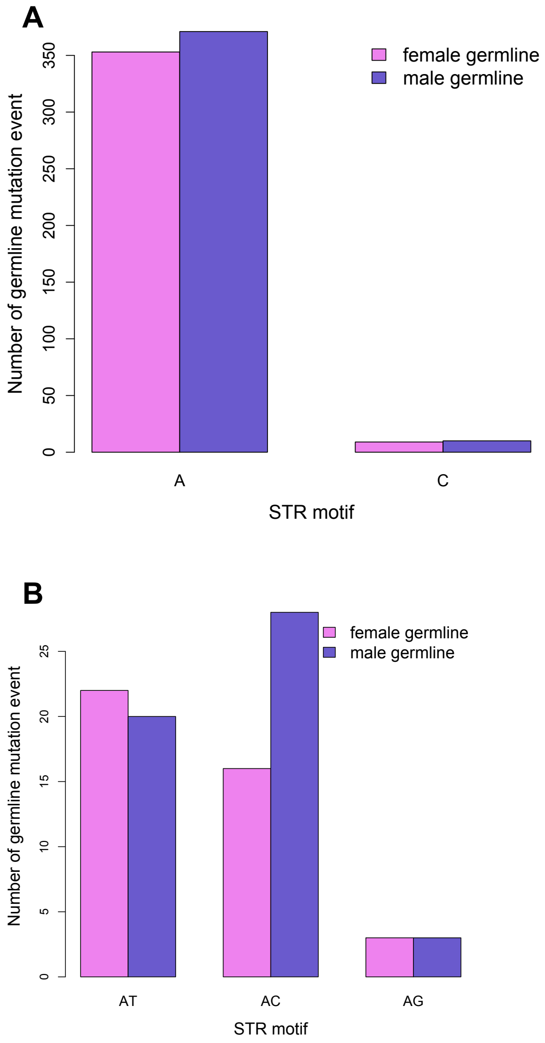


Figure S15. Germ-line mutation rates by STR motif and repeat number. Only STR repeat numbers supported by ≥100 loci are plotted. Dotted lines represent 95% confidence intervals from multinomial sampling. (A) mononucleotide STRs. (B) Dinucleotide STRs.

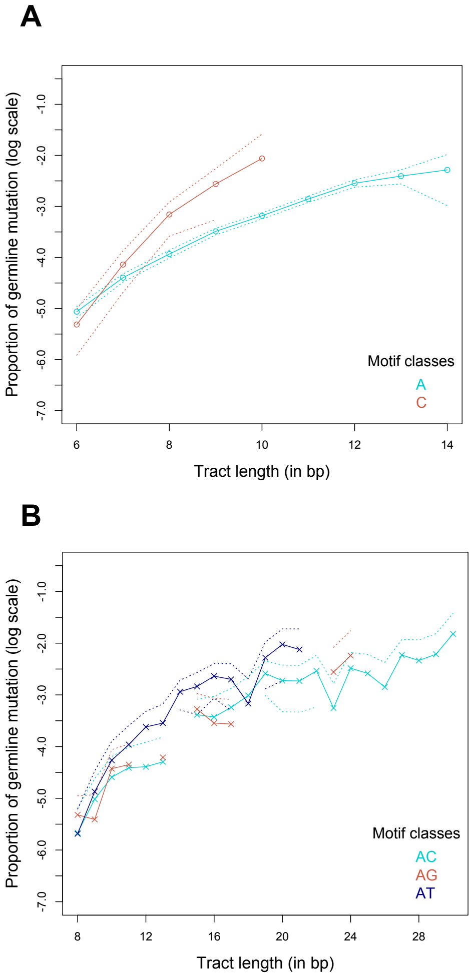


Figure S16. Germ-line mutation rates for different mutation categories. OnlySTR repeat numbers supported by ≥2,000 loci are plotted. Dotted lines represent 95% confidence intervals from multinomial sampling. (A) Dinucleotide STRs. (B) Trinucleotide STRs. (C) Tetranucleotide STRs.

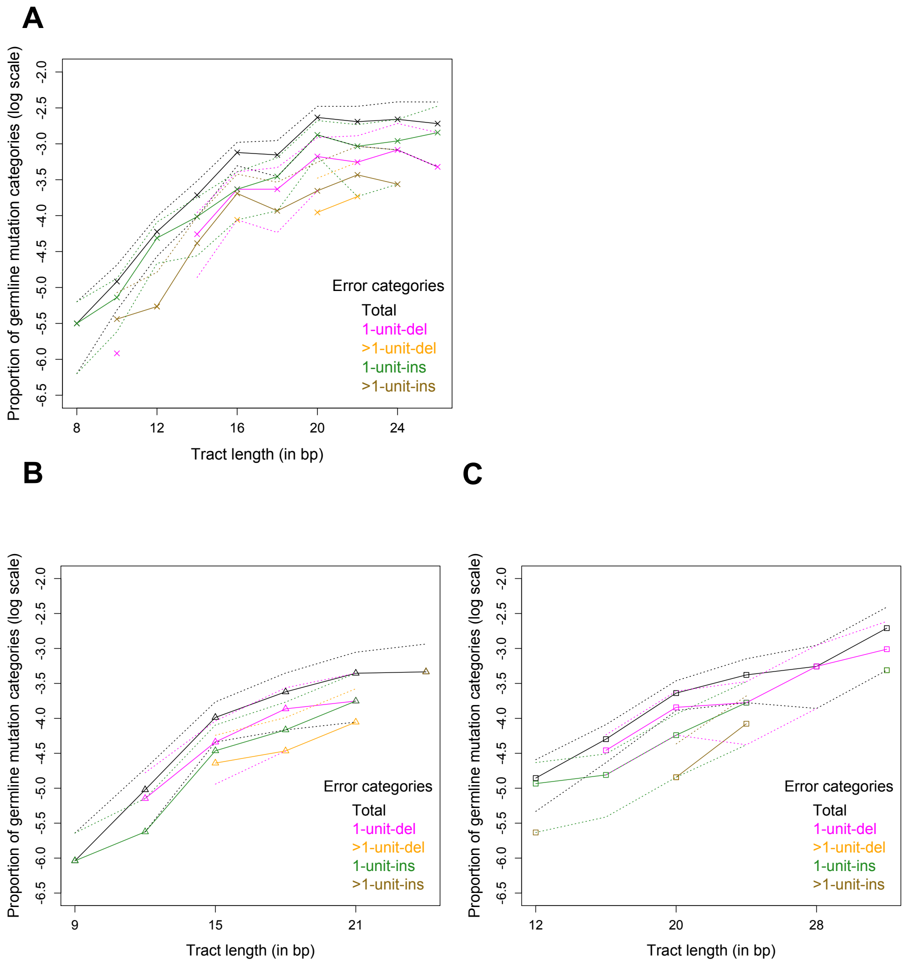
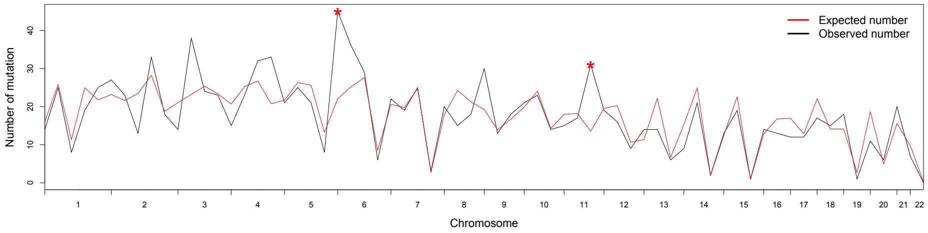


Figure S17. Expected and observed germ-line mutation distribution in the genome at 50 Mb window. Tick mark on X-axis indicates start and stop windows of each chromosome. The starting window (starting coordinate) is on the left. Windows that are shorter than 50 Mb were also included in the plot. Red stars indicate windows that have significantly higher than expected numbers of mutation.


Expected number
Observed number
Number of mutation
Chromosome
*
*