2436–2445 *Nucleic Acids Research, 2019, Vol. 47, No. 5* *Published online 30 January 2019*
doi: 10.1093/nar/gky1318

Short tandem repeat stutter model inferred from direct measurement of *in vitro* stutter noise

Ofir Raz[†], Tamir Biezuner[†], Adam Spiro, Shiran Amir, Lilach Milo, Alon Titelman, Amos Onn, Noa Chapal-Ilani, Liming Tao, Tzipy Marx, Uriel Feige and Ehud Shapiro[*]

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 761001, Israel

Received May 27, 2018; Revised December 19, 2018; Editorial Decision December 26, 2018; Accepted January 02, 2019

ABSTRACT

Short tandem repeats (STRs) are polymorphic genomic loci valuable for various applications such as research, diagnostics and forensics. However, their polymorphic nature also introduces noise during *in vitro* amplification, making them difficult to analyze. Although it is possible to overcome stutter noise by using amplification-free library preparation, such protocols are presently incompatible with single cell analysis and with targeted-enrichment protocols. To address this challenge, we have designed a method for direct measurement of *in vitro* noise. Using a synthetic STR sequencing library, we have calibrated a Markov model for the prediction of stutter patterns at any amplification cycle. By employing this model, we have managed to genotype accurately cases of severe amplification bias, and biallelic STR signals, and validated our model for several high-fidelity PCR enzymes. Finally, we compared this model in the context of a naïve STR genotyping strategy against the state-of-the-art on a benchmark of single cells, demonstrating superior accuracy.

INTRODUCTION

Short tandem repeats (STRs, also known as microsatellites) are repetitive elements of 1–6 bp long that constitute ∼3% of the human genome. They are best known for their highly mutative properties *in vivo*, which is due to polymerase slippage that results in repeat contraction/expansion. Although their mutation rates vary dramatically, even low estimates are 3–4 orders of magnitude larger than of random point mutations, highlighting STRs as a tool of growing interest for various applications (1). In disease, STRs are linked to tens of human diseases such as Huntington's disease (2); In several cancer types, mismatch repair deficiencies are analyzed utilizing STR polymorphic state, pointing to the disease progression (3). In genetics studies, STRs are utilized to study population genetics and phylogenetics (4,5). In regulatory genomics, the importance of STRs as regulatory elements was recently demonstrated (6). Recently, due to technological advancements in single cell (SC) genomics, SC STR analysis became of research interest for applications such as cell lineage phylogenetic analysis within an organism (7,8) and for pre-implantation genetic diagnosis (9).

A key challenge for STR analysis is that they undergo noisy amplification *in vitro*, similarly to *in vivo* replication slippage. This noise, often termed 'stutter', is commonly manifested by excessive peaks when STR length data is plotted in a histogram of lengths (see example in Figure 1B). Despite the value of the high polymorphicity of short unit STRs (e.g. in cancer diagnosis, forensics and phylogeny), they are still not commonly used for most assays due to excessive stutter noise. To address the stutter problem, simple noise models, such as highest peak analysis, are often employed when genotyping PCR-free NGS libraries or slowly mutating STR loci such as repeat units of three bases or more. These simple models do not apply to highly polymorphic STRs, such as mono and di repeats, specifically in samples, which undergo substantial amplification. Using such models in these cases is likely to result in false genotyping. The problem of genotyping highly polymorphic STRs is even more difficult when genotyping non-hemizygous loci (such as from autosomal chromosomes, X Chromosome in female and in copy number variation (CNV) cases) since it is compounded by amplification imbalance of the two alleles. Such unbalanced amplification is typical in SC studies, as the starting material for WGA is a single copy of each locus.

With the growing need of *in vitro* amplification as a tool for basic and applicative scientific research, straightforward *in vitro* STR amplification studies were performed, in order to calibrate amplification factors and conditions (5,10–12). A common STR stutter noise rule of thumb is that STR mutation rate both *in vivo* and *in vitro* is proportional to two main factors: (A) unit type length: short unit STRs (mono- and di-repeats) are more mutable than longer unit types. (B) STR length: Longer STRs (in repeat number) are more mutable than shorter STRs (1). Nevertheless, despite years

[*] To whom correspondence should be addressed. Tel: +972 8 934 4506; Fax: +972 8 934 2125; Email: ehud.shapiro@weizmann.ac.il
[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

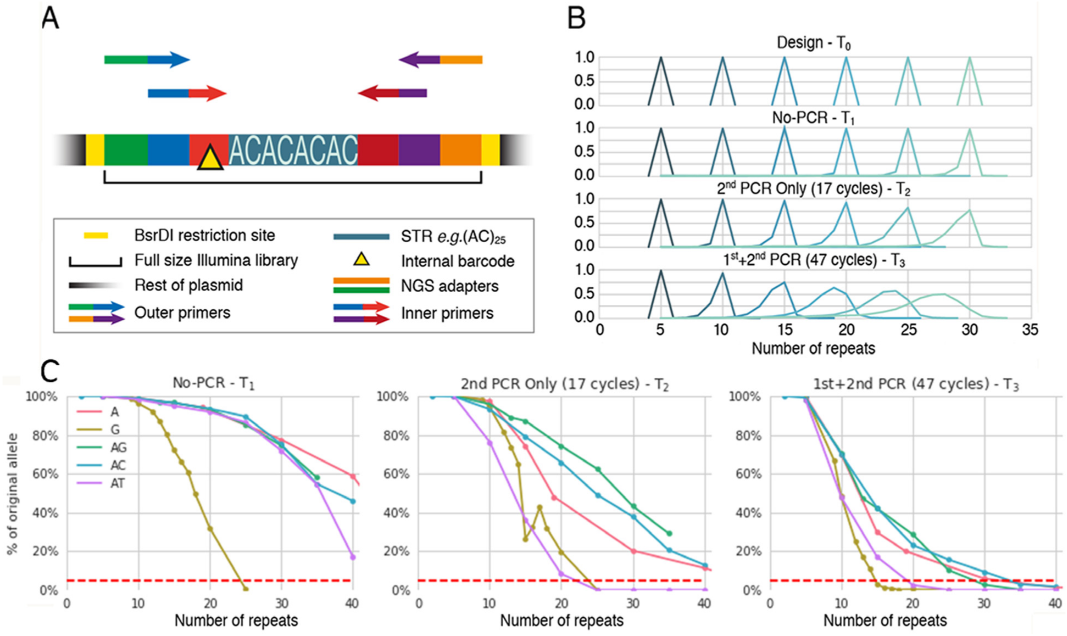

Figure 1. The synthetic STR experiment summary. (**A**) Schematic description of the synthetic library. In each plasmid, a different synthetic STR construct was designed, synthesized and clone-sequenced for various STR types and length. The STR was designed within a context of an Illumina Truseq-HT dual index library to enable for nested PCR amplification at two time points (T_2- amplification using outer primers only, T_3-amplification using inner primers followed amplification by outer primers). The library is flanked by BsrDI restriction sites to enable direct sequencing of the STR library without amplification (T_1). Internal barcode (yellow triangle) is a short sequence, unique to each STR length to detect for cross-contamination. See text and methods for elaboration and Supplemental Table S1 for the designed constructs. (**B**) AC STRs repeat-number histograms, as were interpreted from sequencing results (T_1, T_2 and T_3), compared to their expected length, T_0 (designed sequence). (**C**) Sequencing analysis results of each STR type, repeat-number and time point described as the percentage of the original (designed) signal from all the reads. Dashed line at the 5% marks the lower threshold of analysis: data points below the mark were deemed too noisy and were excluded from downstream analysis.

of STR research, a well-defined stutter behavior model is still lacking. The emergence of next generation sequencing (NGS) as a tool for large scale and detailed per-base analysis of STRs has re-emphasized the need for bioinformatics tools for STR analysis. While most current tools focus on mapping reads to the reference genome (5,13,14), their stutter error correction algorithms are mainly calibrated with statistical models based on indirect measurements such as STR distributions in progenies, in populations and/or in user-defined data sets. Here we present a method for controlled measurements of stutter behavior during amplification for various STR types and sizes. Utilizing these measurements, we calibrated a mathematical model that accurately captures and predicts the stutter pattern of *in vitro* STR amplification.

MATERIALS AND METHODS

Controlled amplification noise measurement of a synthetic STR library

STR plasmid design: Sequence verified cloned plasmids containing synthetic STRs of different types and sizes (Supplemental Table S1) were ordered from either IDT or GenScript (pIDT-kan and modified puc57-Kan vectors, respectively). Cloning vectors were validated to exclude BsrDI restriction sites. STRs were synthesized in the context of a complete Illumina NGS library (Truseq HT) to allow for nested amplification, and to enable a direct digestion using the Type IIS restriction enzyme BsrDI, thus creating a sequencing ready library. See elaboration in main text and in Figure 1. Immediate STR flanking sequences were validated to avoid partial STR repeat unit occurrence (e.g. $(AC)_X$ followed by 'A'). Internal 3-mer internal barcodes were inserted to allow for cross-contamination detection between samples. Several amplification time points were measured:

T_1 (No-PCR) control. T1 (No-PCR) control was performed by pooling all STR plasmid libraries at equal concentration and digestion with BsrDI enzyme (NEB) according to manufacturer protocol. Digestion was performed at 65°C for 16 h, followed by inactivation at 80°C for 20 min. Reaction was then processed for sequencing (see later description in 'Pooling and sequencing').

T_2 and T_3 PCR experiments. In the T_3 experiment, each STR plasmid (10^{-4} μg/μl) was loaded as template in an AccessArray (AA) PCR chip. Each primer inlet was loaded with the same primer solution ('Inner primers') composed of X1 Access Array Loading Reagent (Fluidigm) and primers: Control_Fw:

5′-CTACACGACGCTCTTCCGATCTTCCTAATCT TACGCGGCCATAAC-3′ and Control_Rev:

5′-CAGACGTGTGCTCTTCCGATCATGGACAG TCTTTAAGAGCCCATC-3′(IDT), at a concentration of 1 μM each. PCR reactions and purifications were performed as described in (8): In summary, a first PCR of 30 cycles PCR reaction is performed in the AA chip. Following sample harvesting, purification and dilution

1:100, a two-step second PCR of 17 cycles (5 cycles with annealing temperature of 55°C + 12 cycles with annealing temperature of 70°C) is performed to generate a dual indexed sequencing library (note that the 'Outer primers' sequences were as described for the second PCR primer sequences in (8). The first PCR (in the AA chip) is done using the manufacturer recommended enzyme: FastStart High Fidelity PCR System, dNTPack (Roche) while the second PCR is done using Q5 Hot Start High-Fidelity DNA Polymerase (NEB) with the addition of SYBR green I (LONZA) at a final concentration of X1, to enable real time tracking of amplification. Following second PCR, each sample was purified using SPRI beads.

T_2 PCR was performed by using 0.1–1 ng of each STR plasmid as a template. Samples were processed in accordance with the T_3 second PCR protocol.

Pooling and sequencing. All samples (T_1, T_2, T_3) were purified and concentrated using MinElute PCR purification kit (Qiagen), pooled together and size selected (200–500 bp) using a 2% agarose BluePippin gel cassette (Sage Science) utilizing an upgraded software that avoids blue light exposure after marker detection. Products were concentrated again (Minelute) and were sequenced by a 2 × 220 bp sequencing (Miseq, Illumina) using the manufacturer recommended sequencing primers (R1, Index) and custom R2 primer 5′-GTGACTGGAGTTCAGACGTGT GCTCTTCCGATC-3′ (HPLC grade, IDT).

Experimental validation of the model by using controlled synthetic templates

We opted to validate the model using five high fidelity PCR enzymes, using the controlled synthetic STRs as templates. The enzymes were: the two enzymes that were described above (Q5 High-Fidelity DNA Polymerase and FastStart High Fidelity PCR System, dNTPack), Phusion High-Fidelity DNA Polymerase (NEB), KOD Hot Start DNA Polymerase (Novagen) and KAPA HiFi HotStart PCR Kit (Kapa Biosystems).

Reactions were as performed in the $T_{2,}$ described above: 20 μl reactions in a 96-well format, with real time amplification tracking using SYBR green I, each time using a different enzyme and buffer composition, different templates, and different barcoding primers. The template for each PCR was 2 μl of 1ng/μl STR plasmids: $(AC)_{20}$, $(AC)_{25}$ or $(AC)_{30}$. Each reaction was duplicated to avoid PCR primer sequence effect (using different indexes). Negative control (water) was added to each PCR. In the serial dilution validation experiment, Q5 enzyme was used as described above, using the same STR plasmids as templates in three concentrations: 1 ng/μl (also used for the enzyme comparison experiment), 10^{-2} ng/μl and 10^{-4} ng/μl.

All Samples were purified, pooled and sequenced as described above.

The following exceptions were considered: (i) Activation, elongation and final elongation were adjusted to fit each enzyme's recommended protocol. (ii) Annealing temperature from the sixth amplification step and on was according to each enzyme's elongation temperature. (iii) PCR reaction was stopped when amplification reached a plateau. (iv) Due to failure of dNTPack to amplify using the standard two-step PCR protocol, we applied the same program as being performed in the first PCR of T_3 (in the AA chip). (v) Reactions mixes were according to manufacturer's protocols, with primer concentrations of 0.3–0.5 μM, with the exception of dNTPack, which composition was according to Fluidigm's recommended reaction mixture with primer concentration of 0.1 μM each and a final volume of 10.6 μl.

Experimental validation of the model by using single cell STR data

The high fidelity PCR enzymes were used in this study were: NEBNext Q5 Hot Start HiFi PCR Master Mix (NEB), NEBNext Ultra II Q5 Master Mix (NEB), FastStart High Fidelity PCR System, dNTPack (Roche), KOD Hot Start DNA Polymerase (Novagen), KAPA HiFi HotStart PCR Kit (Kapa Biosystems) and PrimeStar Max (Takara).

A recreation of the original amplicon targeted sequencing protocol as presented in (8) was performed in order to assess the error rate per polymerase enzyme using the STR stutter model. In summary, AA chip generates a mixture of 48× sample + PCR wells, with 48× primer mixes (1769 of amplicons in total, see Supplemental Table S2), ending up with 2304 nanoliter reactions, which are later harvested to each sample's inlets (48 reactions to a single well). Following sample harvesting, purification and dilution 1:100, a second PCR was performed at a final volume of 20 μl, each sample with its corresponding PCR enzyme from the first PCR reaction and using its protocol, unless otherwise mentioned. Purification and pooling procedures were as described in (8). The 'unified' AA first PCR protocol was composed in accordance with the thermal cycling protocol guidelines of all examined polymerases. Activation was performed at 98°C for 3 min followed by 5 cycles of 98°C for 20 s, 60°C for 15 s and 70°C for 15 s and 20 cycles of 98°C for 20 s, 70°C for 15 s and 70°C for 15 s. A final elongation step was added: 70°C for 5 min. Each polymerase reaction mixture was according to its manual. To avoid over-cycling, SYBR green (X1) was added to enable amplification tracking by real time PCR. Libraries were first shallow sequenced in 2 × 220-bp in a Miseq sequencer (Illumina), followed by normalization and pooling by number of total reads per sample, and deep-sequenced in NextSeq(Illumina) 2 × 151-bp.

Computational analysis

For the initial analysis of the synthetic STR experiments, enzyme comparison and biallelic genotyping, the pipeline presented by Biezuner *et al* (8) was used. In short, reads are processed using *cutadapt* (https://cutadapt.readthedocs.io/en/stable/) and PEAR (15), followed by unique mapping of the merged reads to their target using read alignment of only the read's edges corresponding to the primer pairs. STR repeat number is then determined by aligning the read to references containing a range of STR lengths and choosing the reference length with the highest alignment score.

For the initial analysis of the library that was used for comparing genotyping accuracy, FMSV (16) mapping was used to generate the input for *R&B* genotyping tool while BWA-MEM was used to generate the input for HipSTR (17).

An end-to-end implementation of the mapping and genotyping approaches described here, preceding and including the *R&B* genotyper is available at https://github.com/ofirr/clineage/tree/standalone.

RESULTS

Controlled amplification of synthetic STR molecules

In order to study the stutter pattern as a function of amplification, we have designed and ordered a library of plasmids (Figure 1A), each containing a unique combination of STR type and length, spanning all naturally occurring mono and di repeats (namely: A, C, AC, AG, AT) in the full spectrum of their natural genomic occurrence (18) (Supplemental Table S1). The construct within each plasmid is sequencing-ready and includes a unique Illumina dual index combination for direct sequencing (T_1) and a unique barcode for cross contamination control. Overall, the experimental setup allows for a controlled amplification and sequencing of all highly mutable STRs at three independent time points (T_1-no amplification, T_2-single amplification, T_3—two amplifications) using various nested PCR primers, with the ability to measure the specific sequencing noise and bias for each STR length and type (Figure 1B, C and Supplemental Figures S1–S5).

Fitting and model comparison

The data generated for the three time points (T_1, T_2 and T_3) was used for the calibration of a computational model that predicts the stutter pattern at any theoretical amplification cycle, given the repeat unit and length of the STR. Together with the assumption of perfect synthesis process (T_0 – the designed construct prior to any manipulation), supported by Sanger sequencing.

Our goal is to predict the stutter histogram H of repeat numbers for any amplification-time-point t and for any original length n in repeat units, $H(t,\ n)$, and we assess the performance by:

$$\frac{\sum_{l_0}^{l_n} \sum_{t_0}^{t_3} d\left(H_{model}\left(t,\ l\right),\ H_{seq}\right)^2}{n}$$

where $d(H_1,\ H_2)$ is the distance between the two histograms. We have examined multiple distance metrics for the sake of histogram comparison and found 1-correlation distance to be the most suitable (Supplemental Figure S11).

We attempted to fit a survey of multistep, length-dependent and length-independent exponential models (19) as well as several in-house polynomial models. All models were considered in the context of a single PCR-cycle and extended to the measured time points using discrete-time Markov chain (Figure 2). The models' fitness was assessed by minimizing the overall distance between their modeled histograms and the measured stutter patterns (Figure 1B and Supplemental Figures S1–S5) and their parameters were fitted using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) (20) optimization algorithm. We finally chose the model 'Linear1up3dw', a contraction-biased multistep linear model that best match the stepwise probabilities. This model obtained the best overall fit across the attempted mono and di STRs when calibrated individually for each repeat type.

We model the stutter as an iterative mutational process with multiple steps (PCR cycles in this case). For each of these steps, our genotype can contract by up to 3 repeat units or elongate by a single repeat. The probability of such a mutation is linearly dependent on the STR's current length.

Validation and genotyping comparison

To confirm the model, we propose *R&B*, a naïve genotyping algorithm implementing an exhaustive strategy to call the original STR length from a population of reads with different STR lengths by scoring it against all possible predicted populations of any amplification time and STR length:

$$\arg\min_{t,l} d\left(H_{model}\left(t, l\right),\ H_{seq}\right)$$

Following a meticulous STR genotyping comparison by Willems *et al.* (17), we compare this heuristic only to the current state-of-the-art, HipSTR genotyping tool, on a benchmark experiment first presented by Biezuner *et al.* (8). This experiment involves cells from a controlled *ex vivo* cell lineage tree experiment, picked and extracted for their DNA, while documenting their sampling lineage. STR mapping issues were tackled using an STR-targeted enrichment panel (rather than shotgun sequencing) and mapping the known primers panel to the reads in order to identify them. Using a similar strategy (FMSV (16)), we can isolate the problem of genotyping stutter patterns and avoid possible mapping bias.

The known lineage topology of individually analyzed SCs provides a solid reference for the comparison of any genotyping tool. To do so, we have devised the following metric to assess the accuracy of genotyping algorithms.

Let $A : T_{leaves} \rightarrow A$ be the set of alleles assigned to the leaves of tree T by a genotyping algorithm. $P(A,\ T)$ is the maximum parsimony or the minimal number of mutations required to explain set of alleles A on the leaves of tree T.

$D\left(A\right) = \sqrt{\sum_{a\epsilon A} \left(\#_a A - 1\right)^2}$ is the allele diversity.

We define F as the *reference tree fitting*:

$$F\ \left(T,\ A\right) = |A|\ - 1 - P\left(T,\ \ A\right) - D\left(A\right)$$

The *reference tree fitness* aims to balance the diversity of alleles found within this cell group, which provides information describing the topology of T, with the adherence of the genotypes to T. We compensate for the fact that diverse genotypes inherently have a lower parsimony, even when correct.

Using this metric, *Loci* that add valid information regarding the tree will be awarded positive scores while *loci* whose genotyping results contradict the topology will be negatively scored. A *locus* for which there is no relevant information (either no genotyping or a single allele across all cells) will receive a zero score.

Both genotyping methods, *R&B* and HipSTR, provide a measure of confidence together with each locus they attempt to genotype. While these confidence metrics are

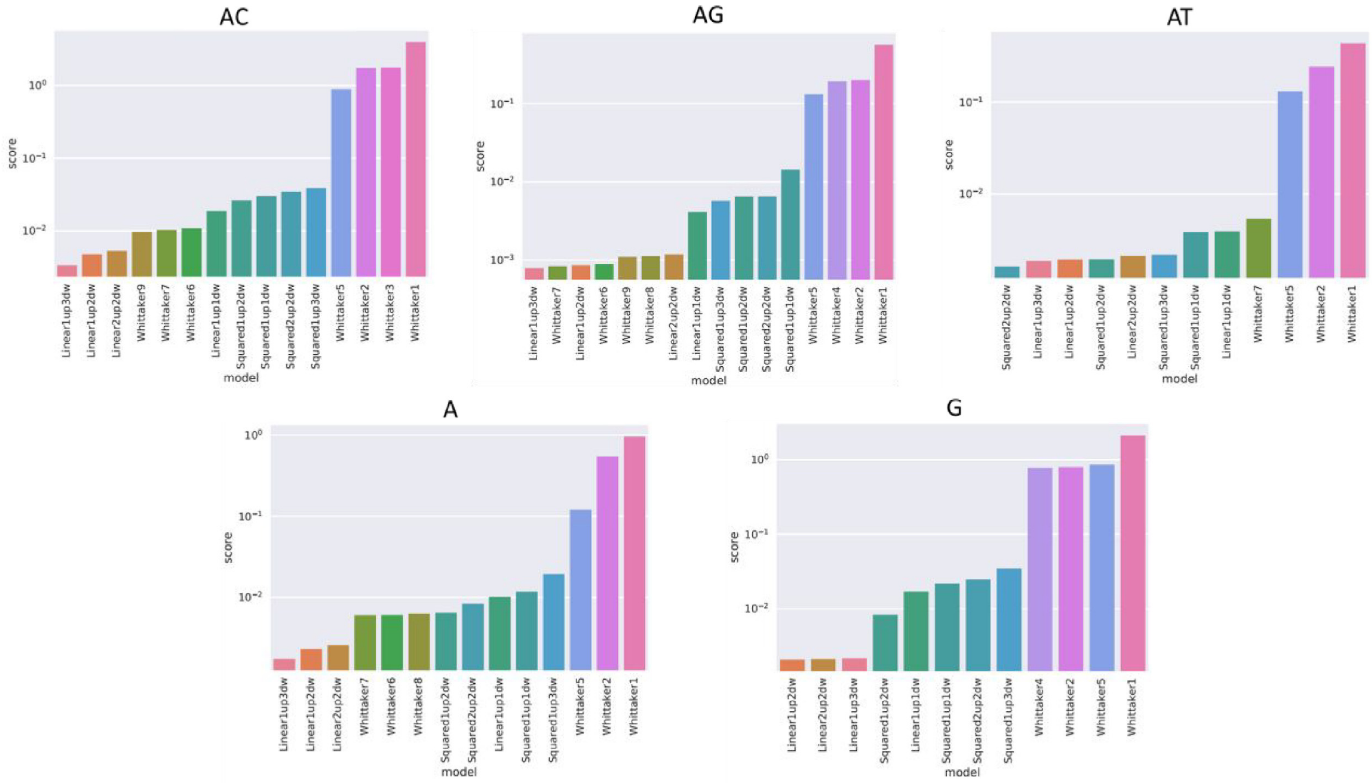

Figure 2. Model fitness to synthetic dataset. Each model parameters were optimized to best fit the dataset measured from the synthetic plasmids at different amplification time-points. The scores reflect the squared sum of distances (here distance = 1 – correlation) across all measurements for each STR repeat unit type (i.e. AC, AG…) divided by the number of samples. Models include Whittaker1–9 as introduced by Whittaker *et al.* (19) and polynomial models named after their number of variables and degrees. Models that failed in the optimization for a particular repeat type were not included in the respective sub-figure. For a depiction of the compared models, see Supplemental Figures S6–S10.

very different and have different distributions across the attempted cells/loci population, we can try to compare them by referring to percentiles of the full scores set, the top 10%, top 50% or any other threshold. To compare similar confidence genotyping attempts of both tools despite the large difference in the number of successfully genotyped loci, we compared only the cells/loci combinations for which both tools provide a genotyping attempt ranked with sufficient confidence (Figure 3A, B). Here we can see that across most confidence levels, when both tools attempt to provide a genotype, *R&B* attempts are more in line with the true tree topology.

To maintain simplicity, we only account for mono-allelic loci from the X chromosome of the cancerous cell line used in this experiment (human male DU145). Other chromosomes were found to have major copy-number abnormalities.

Experimental validation by controlled synthetic templates and real genomic data

To provide experimental-based confirmation for the model validity we opted to measure its simulated amplification cycles analysis in a series of controlled experiments. First, by using synthetic STRs as controlled templates for serial dilution analysis, and later, using synthetic STRs and real genomic data to demonstrate the robustness of the model analysis to the utilized PCR enzymes.

Experimental validation of the model by using controlled synthetic templates

(1) Controlled amplifications of synthetic STRs in a serial dilution experiment. Using the synthetic STRs that were used for the model calibration above, we generated highly accurate NGS data originated from amplification of known and controlled templates. First, we have generated an NGS dataset generated from a single PCR amplification using the Q5 enzyme (NEB), as previously described for the T_2 experiment, of three different templates: $(AC)_{20}$, $(AC)_{25}$ and $(AC)_{30}$, each using three serially diluted templates (by 10-fold each). Our model's simulated cycles linearly correlate with the actual number of amplification cycles performed, as expected from serially diluted samples (Supplemental Figure S12A, B).

(2) Model robustness to PCR enzyme by an enzyme comparison assay. First, we performed a small-scale PCR enzyme comparison by applying five commercially available PCR enzymes on the same synthetic templates as used above at an equal template concentration (using a subset of the generated data of Q5 from the above-mentioned experiment (1) and four other enzymes). We show that the model accurately captures the stutter variability between different polymerases within a single degree of freedom, its simulated cycles (Supplemental Figure S12A, C).

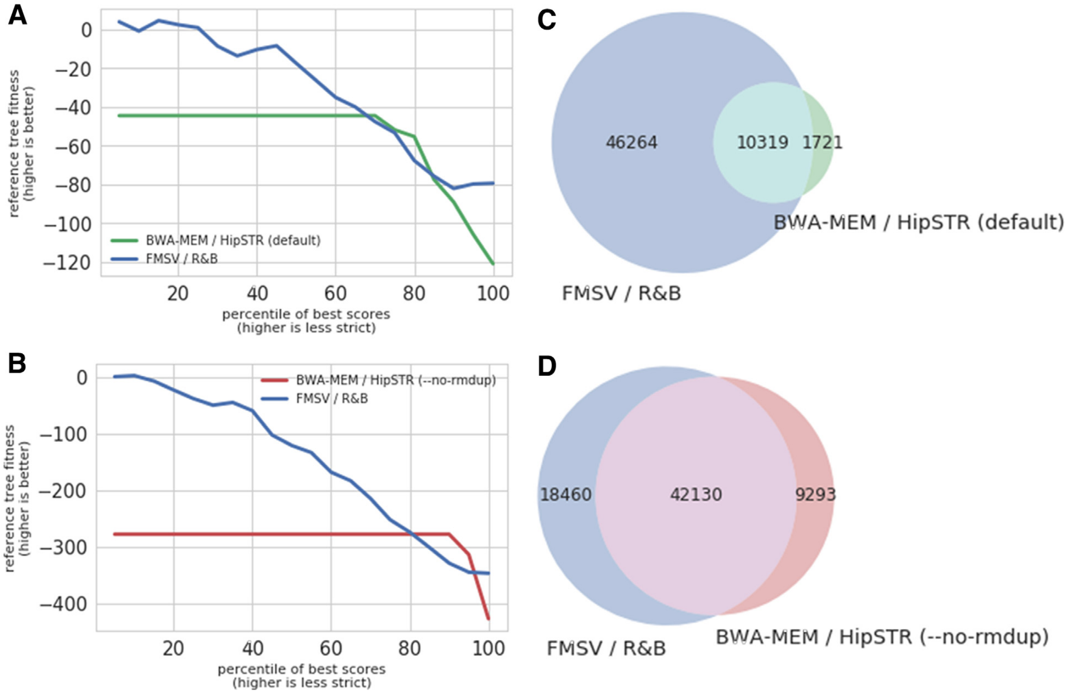

Figure 3. Genotyping results. Comparison of the proof of concept genotyping method, *R&B*, with HipSTR genotyping tool (17) under both the default parameters and with the '–no-rmdup' flag, appropriate for PCR amplified results (According to HipSTR's documentation: '*no-rmdup: Don't remove PCR duplicates. By default, they'll be removed; Why? Your sequencing data is for PCR-amplified regions*'). We compare the results' quality in tiles A and B by measuring their ability to accurately genotype the sequencing results of the *ex-vivo* cell lineage tree (reference tree fitness) as a function of their subjective confidence metrics (confidence greater than percentile threshold, lower values mean higher confidence but less loci). We compare loci that were genotyped by both genotyping methods within similar confidence percentiles (**A**, **B**) and the total quantity of the produced genotypes (**C**, **D**). We see that *R&B* excels in both quality and quantity. Across all cases, we used a minimal coverage of 5×, no confidence filters prior to percentile calculation and no stutter filtering for HipSTR. To maintain simplicity, we only account for haploid loci from the X chromosome of the cancerous cell line used in this experiment (human male DU145).

Experimental validation of the model by using single cell STR data

Following the successful proof of concept of polymerase comparison using synthetic templates, we opted to enlarge the validation to thousands of data points per each polymerase to create a statistical significant polymerase error rate comparative assay based upon the measured error rate per each thousands of genomic STR loci.

To generate a valid comparison we opted to utilize the same polymerase for the entire targeted sequencing protocol as outlined in (8) using 1769 amplicons (Supplemental Table S2). We first selected six high-fidelity enzymes and opted to apply them in parallel to a collection of single cell WGA DNA templates, picked from H1 cell line, which demonstrates a normal karyotype, thus reducing copy number artifacts. To fit all PCR enzymes in a single preliminary AA chip, we composed a 'unified' first PCR thermal cycler protocol that meets the requirements of all enzyme manuals (see methods section), with as little digression as possible from manufacturers' recommended protocols. Second PCR was performed with each enzyme's original protocol.

We first performed a preliminary experiment (Supplemental Table S3, Supplemental Figure S13) with two single cell DNA and two control templates, positive and negative, all in duplicates. We eliminated the KOD enzyme from further experiments due to its low success rate (mapped reads/total reads ratio) but maintained dNTpack despite even lower success rate, in light of its successful performance in a previous experiment (8). We regard this inconsistency to the difference in thermal cycler programs between dNTpack's manufacturer manual and the 30 cycles Fluidigm protocol, as described in (8).

In a follow-up experiment, we have enlarged the cohort of samples to 22 single cell DNA samples and two controls (positive and negative) which were used as templates for the different PCR enzymes, in duplicates (48 samples per enzyme). dNTPack enzyme was used twice. First, with expected negative outcome, following the same failed protocol mentioned above and labeled here 'dNTPack'. Second, with dNTPack as the first PCR enzyme of the original AA protocol (8), and with UltraII serving as the second PCR enzyme.

Applying our STR noise model on the sequencing data received from the large scale follow-up experiment, we plot the simulated cycle scores of all single cells in the experiment (duplicates included, Figure 4A, see also results summary in Supplemental Table S4). PrimeStar demonstrated a significantly lower simulated cycle number compared to the other enzymes. UltraII and Q5, both neck to neck second best in the number of simulated cycles category, an expected result as both enzymes are based essentially on the same Q5 enzyme with a different mix composition, emphasizing the robustness of the model.

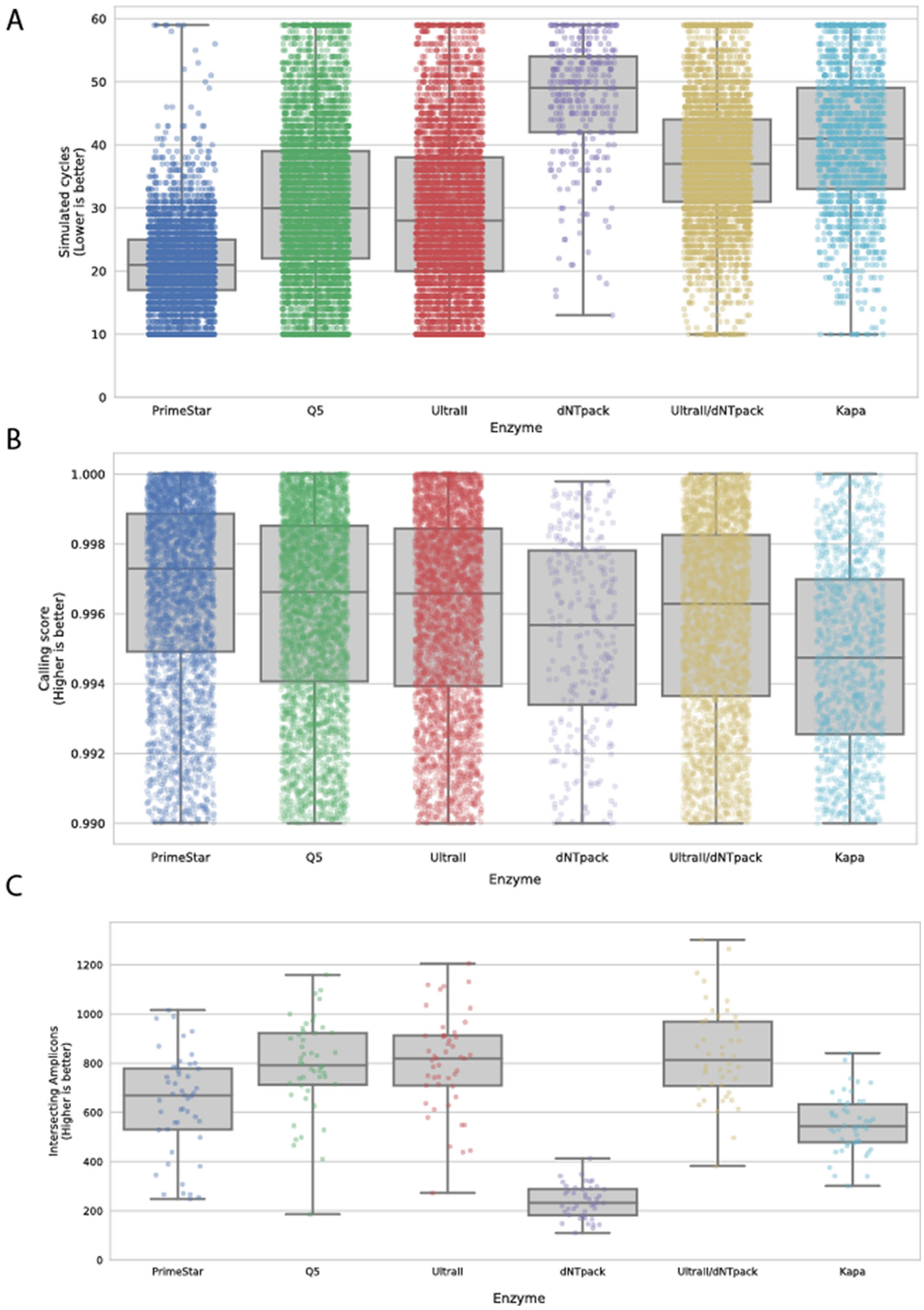

Figure 4. Comparison of genotyping results for various PCR enzymes using targeted PCR on a template of single cells WGA DNA. (**A**) Comparing the number of simulated PCR cycles that best fit the measured histogram reflects the STR-specific stutter noise that is produced by a fixed number of actual PCR cycles. (**B**) Comparing the fitness (correlation) between the simulated histograms and the measured ones. (**C**) Loci counts that were retrieved from each SC.

Overall, we show that the model accurately captures the variability between different polymerases within a single degree of freedom, its simulated cycles, making it robust to any switch in utilized biochemical methods.

Biallelic calling—genomic data

Heterozygous STR genotyping is often hindered by the relative similarity of the original underlying alleles, overlaid stutter noise, amplification bias and the often-missing phased SNVs that can allow easy haplotyping (17).

We opted to try and fit biallelic loci that amplified unevenly during the WGA process on SCs by extending the exhaustive search to nearly all possible allele combinations and at any proportion from the set: 0.1/0.9, 0.2/0.8, . . . , 0.5/0.5, . . . , 0.9/0.1 (Supplemental Figure S14). In order to assess our ability to accurately discover the true alleles that compose a stuttered biallelic histogram, we have selected autosomal loci from a SC population of H1 stem cells (8) that consistently alternated between two alleles when genotyped as mono-allelic (Figure 5A–C, first column). Since the two alleles can appear simultaneously and at any proportion (Figure 5D), we can assume these cases presented the biallelic locus' alleles at a proportion of 0/1 or 1/0 and that occurrences of this loci that failed to be genotyped as mono-allelic would present both alleles.

By applying our model and comparing histograms of overlapping bi-allelic signals (Figure 5A–C, third column, blue) with simulated overlapping stutter patterns (Figure 5A–C second column) we see that the expected allele lengths can be recovered by selecting the closest simulated stutter pattern (Figure 5A–C, third column, green).

DISCUSSION

STR usage in scientific research is increasing. High throughput sequencing opens a new frontier for STR science, both for basic (4,6) and for applicative research (21,22). With that understanding, in recent years, bioinformatics tools were developed to map and genotype STRs in a high-throughput genome-wide scale with improved accuracy and speed over standard mapping algorithms (5,13,14). However, current tools still struggle with the *in vitro* amplification stutter noise that is typical to STRs, and in particular to highly mutable STRs. Recent biochemical advances have enabled PCR-free protocols that substantially decreased the effect of stutter noise in STR analysis (5). However, these protocols have some limitations: (1) they require bulk amounts of template, making it incompatible with SC analysis, which requires whole genome amplification (2). In most cases, only a fraction of the STRs in the genome is required for analysis and therefore targeted amplification is required (23). Overall, this work lays the foundation for a better understanding of STR behavior in the NGS era. Although STR enrichment and sequencing kits are now available, a comprehensive assessment of the STR sequencing capabilities of extant sequencing machine was not systematically carried out, except for known constraints of some technologies such as mononucleotides sequencing in pyrosequencing based technologies (24) and inferred estimation of such noise from old Illumina platforms (25). Here we provided a controlled measurement of noisy sequencing at different amplification conditions and even in amplification free STR molecules.

We described here a new stutter model for the highly mutable STRs over *in vitro* amplification. The novelty of this model is that it is calibrated with NGS data generated by a controlled amplification of a range of di-repeat STRs of different types and sizes (according to their genomic occurrence in human). One key element in our model is that it takes into account that during amplification, the molecule lengths stochastic mutations can be accurately predicted, according to its inputs, the STR type, and the input length distribution of the previous amplification step. We chose to model the STR noise as a discrete-time Markov chain (DTMC). Our model enables easy calibration of different types of STRs. However, our data clearly shows a distinct and unusual pattern of noisy amplification of AT, which currently cannot be determined by either Markovian or binomial models, and may require modified model in the future. This variation in mutational mechanism was suggested previously (1).

We provided three types of experimental-based evidence for the effectiveness of our model:

(1) Controlled amplification of STR plasmids. First, by utilizing it to measure an accurate amplification difference between known STR templates of various types and concentration, and second, by validating it against various types of polymerases.
(2) Comparative analysis of STR amplification of thousands of genomic single cell STRs.
(3) Both experiments have demonstrated the model robustness, such that although calibrated by a specific set of polymerases and conditions can be trustfully used as a quantitative tool for analyzing mutational processes by any NGS downstream process. Future work will enable a large-scale utilization of this model for assaying and/or optimizing other mutational processes, such as WGA.
(4) Utilization of NGS genomics datasets from SCs by accurately analyzing STRs from biallelic histograms, from drifted histogram, unclear determination of single peaks, and unbalanced allelic representation.

We also compared our model to a state-of-the-art genotyping tool (14). As the highly mutable STRs studied here undergo frequent *in vivo* mutations even within the context of clonal expansion, a natural reference to compare our genotyping results against cannot be obtained. While synthetic STRs libraries such as the one used here to study the stutter patterns behavior, can provide a solid reference for genotyping, they are expensive to scale and do not encompass the complexities and degrees of freedom of genomic STR loci. We therefor approached a previously published dataset of *ex vivo* controlled cell lineage tree (8) that provides a solid phylogenetic reference to natural somatic mutations occurring in hundreds of STR loci, allowing indirect but accurate and large scale genotyping comparison. Our model outperforms both by the number of STR genotypes and both by the calling confidence, when compared with respect to the *ex vivo* tree.

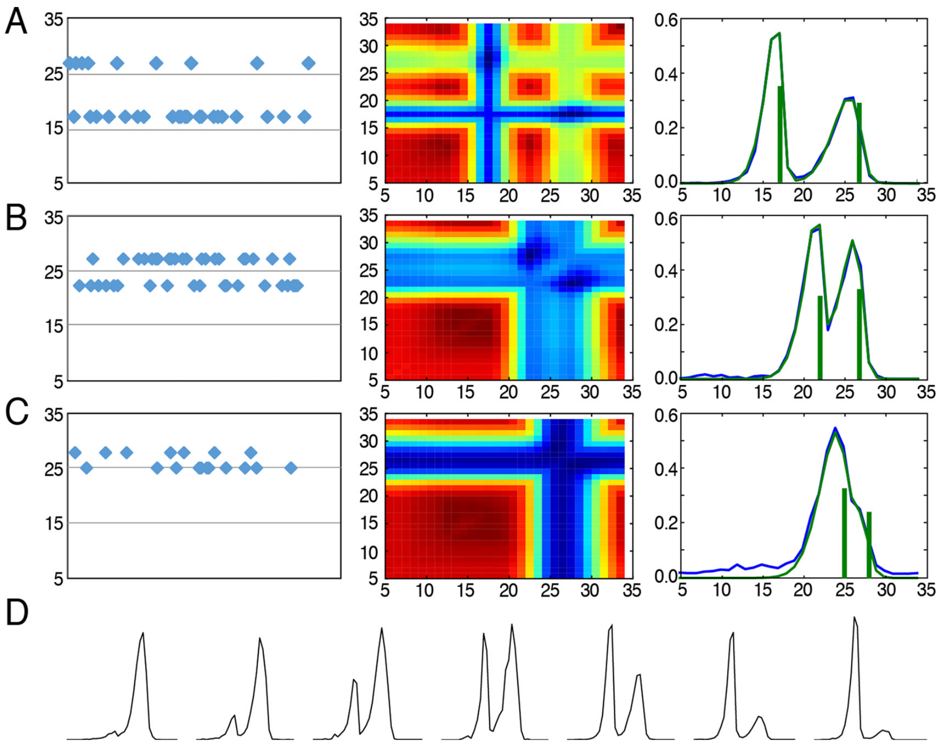

Figure 5. Biallelic genotyping using overlaid model histograms. Figure rows A, B and C show the successful genotyping of biallelic loci (AC repeats) within a SC population of H1 stem cells (8). (**A**) Recognizing overlapping alleles spanning 17 and 27 repeats, (**B**) 22 and 27 repeats, and (**C**) 25 and 28 repeats. First column—monoallelic genotypes recognized in the clonal population. Second and third columns—in biallelic SC signal: second column: Heatmap of the correlation scores between the predicted and the measured histograms across the space of possible alleles; Third column: overlaid model prediction (green histogram) on top of the measured histogram (blue histogram). The resulting genotypes are marked as vertical green lines that also depict the alleles' proportion in their height. (**D**) Examples of asymmetric allele proportions.

We acknowledge that the bioinformatic improvement we provide here is mainly the stutter model itself, where current tools, mainly HipSTR, are implemented as a more inclusive STR genotyping tools in terms of phasing, haplotyping and interfaces with standard bioinformatics pipelines. Nevertheless, we recommend this model as an integrative step for STR noise analysis, specifically for SC analysis, where the sequenced samples undergo extensive amplification or in high sensitivity STR analysis, e.g. diagnosis of Microsatellite Instability (MSI) in cancer samples (26). The tolerance of our model to noisy STR signal allows for a more flexible experimental design and opens the gate for highly mutable STR sequencing research.

DATA AVAILABILITY

Sequencing data generated in this study have been submitted to ArrayExpress (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-7602. The sequencing data used for the comparison conducted in Figure 3 can be found under accession E-MTAB-6411.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank O. Bechar for the prompt design of figures.

FUNDING

European Union [ERC-2008-AdG (Project No. 233047) and ERC-2014-AdG (Project No. 670535)]; Israel Science Foundation grants: Individual Research Grant [456/13]; Joint Broad-ISF Research [422/14 and 2012/15]; German Research Foundation [DFG SH grant 867/1-1]; Kenneth and Sally Leafman Appelbaum Discovery Fund. E.S. is the Incumbent of The Harry Weinrebe Professorial Chair of Computer Science and Biology. Funding for open access charge: European Union [ERC-2014-AdG [Project No. 670535)].

Conflict of interest statement. None declared.

REFERENCES

1. Ellegren,H. (2004) Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
2. Mirkin,S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.
3. Salipante,S.J., Scroggins,S.M., Hampel,H.L., Turner,E.H. and Pritchard,C.C. (2014) Microsatellite instability detection by next generation sequencing. *Clin. Chem.*, **60**, 1192–1199.
4. Willems,T., Gymrek,M., Highnam,G., Mittelman,D. and Erlich,Y. (2014) The landscape of human STR variation. *Genome Res.*, **24**, 1894–1904.
5. Fungtammasan,A., Ananda,G., Hile,S.E., Su,M.S., Sun,C., Harris,R., Medvedev,P., Eckert,K. and Makova,K.D. (2015) Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res.*, **25**, 736–749.
6. Gymrek,M., Willems,T., Guilmatre,A., Zeng,H., Markus,B., Georgiev,S., Daly,M.J., Price,A.L., Pritchard,J.K., Sharp,A.J. *et al.*

(2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.*, **48**, 22–29.

7. Shapiro,E., Biezuner,T. and Linnarsson,S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14**, 618–630.
8. Biezuner,T., Spiro,A., Raz,O., Amir,S., Milo,L., Adar,R., Chapal-Ilani,N., Berman,V., Fried,Y., Ainbinder,E. *et al.* (2016) A generic, cost-effective and scalable cell lineage analysis platform. *Genome Res.*, **26**, 1588–1599.
9. Eftedal,I., Schwartz,M., Bendtsen,H., Andersen,A.N. and Ziebe,S. (2001) Single intragenic microsatellite preimplantation genetic diagnosis for cystic fibrosis provides positive allele identification of all CFTR genotypes for informative couples. *Mol. Hum. Reprod.*, **7**, 307–312.
10. Byrd,C., Ohtsuka,E., Moon,M.W. and Khorana,H.G. (1965) Synthetic deoxyribo-oligonucleotides as templates for the dna polymerase of escherichia coli: new dna-like polymers containing repeating nucleotide sequences*. *Proc. Natl. Acad. Sci. U.S.A.*, **53**, 79–86.
11. Shinde,D., Lai,Y., Sun,F. and 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.
12. Hite,J.M., Eckert,K.A. and Cheng,K.C. (1996) Factors affecting fidelity of DNA synthesis during PCR amplification of d(C-A)n.d(G-T)n microsatellite repeats. *Nucleic Acids Res.*, **24**, 2429–2434.
13. Gymrek,M., Golan,D., Rosset,S. and Erlich,Y. (2012) lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.*, **22**, 1154–1162.
14. Highnam,G., Franck,C., Martin,A., Stephens,C., Puthige,A. and Mittelman,D. (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.*, **41**, e32.
15. Zhang,J., Kobert,K., Flouri,T. and Stamatakis,A. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.
16. Tao,L., Raz,O., Marx,Z., Biezuner,T., Amir,S., Milo,L., Adar,R., Onn,A., Chapal-Ilani,N., Berman,V. *et al.* (2017) A duplex MIPs-based biological-computational cell lineage discovery platform. *BioRxiv*, doi:10.1101/191296.
17. Willems,T., Zielinski,D., Yuan,J., Gordon,A., Gymrek,M. and Erlich,Y. (2017) Genome-wide profiling of heritable and. *Nat. Methods*, **14**, 590.
18. Subramanian,S., Mishra,R.K. and Singh,L. (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol.*, **4**, R13.
19. Whittaker,J.C., Harbord,R.M., Boxall,N., Mackay,I., Dawson,G. and Sibly,R.M. (2003) Likelihood-based estimation of microsatellite mutation rates. *Genetics*, **164**, 781–787.
20. Byrd,H.R., Lu,P., Nocedal,J. and Zhu,C. (1994) A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Comput.*, **16**, 19.
21. Churchill,J.D., Schmedes,S.E., King,J.L. and Budowle,B. (2016) Evaluation of the Illumina((R)) beta version ForenSeq DNA signature prep kit for use in genetic profiling. *Forensic Sci. Int. Genet.*, **20**, 20–29.
22. Kim,E.H., Lee,H.Y., Yang,I.S., Jung,S.E., Yang,W.I. and Shin,K.J. (2016) Massively parallel sequencing of 17 commonly used forensic autosomal STRs and amelogenin with small amplicons. *Forensic Sci. Int. Genet.*, **22**, 1–7.
23. Mertes,F., ElSharawy,A., Sauer,S., van Helvoort,J., van der Zaag,P., Franke,A., Nilsson,M., Lehrach,H. and Brookes,A. (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief. Funct. Genomics*, **10**, 374–386.
24. Huse,S.M., Huber,J.A., Morrison,H.G., Sogin,M.L. and Welch,D.M. (2007) *Genome Biol*. Vol. **8**, p. R143.
25. Albers,C.A., Lunter,G., MacArthur,D.G., McVean,G., Ouwehand,W.H. and Durbin,R. (2011) Dindel: accurate indel calls from short-read data. *Genome Res*, **21**, 961–973.
26. Maruvka,Y.E., Mouw,K.W., Karlic,R., Parasuraman,P., Kamburov,A., Polak,P., Haradhvala,N.J., Hess,J.M., Rheinbay,E., Brody,Y. *et al.* (2017) Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nat. Biotechnol.*, **35**, 951.