Short Tandem Repeat stutter model inferred from direct measurement of in vitro stutter noise

Ofir Raz[1,2,3], Tamir Biezuner[1,2,3], Adam Spiro[1,2], Shiran Amir[1,2], Lilach Milo[1,2], Alon Titelman[1,2], Amos Onn[1,2], Uriel Feige[1] and Ehud Shapiro[1,2,*]

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

[2]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 761001, Israel

[3]These authors contributed equally to this work.

*** Corresponding author**

Ehud Shapiro

Address: Department of Computer Science and Applied Mathematics

234 Herzl Street, Rehovot 7610001 Israel

Phone: +972-8-934-4506, +972-8-934-2125

E-mail: Ehud.Shapiro@weizmann.ac.il

Abstract

Short tandem repeats (STRs) are polymorphic genomic loci valuable for various applications such as research, diagnostics and forensics. However, their polymorphic nature acts as a double-edged sword, as during *in vitro* amplification STRs undergo mutational processes that cause stutter noise, especially in the shorter, more mutable, repeat types. 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 known 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 proposed Markov model for the prediction of stutter patterns at any amplification cycle. By employing this model, we have managed to genotype accurately even cases of severe amplification noise, where as little as 3% of the reads accurately reflect the original STR size.

Introduction

Short tandem repeats (STRs, also known as microsatellites) are repetitive elements of 1-6 base pairs long that constitute about 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(Ellegren 2004). In disease, STRs are linked to tens of human diseases such as Huntington's disease(Mirkin 2007); In several cancer types, mismatch repair deficiencies are analyzed utilizing STR polymorphic state, pointing to the disease progression(Salipante et al. 2014). In genetics studies, STRs are utilized to study population genetics and phylogenetics(Willems et al. 2014; Fungtammasan et al. 2015). In regulatory genomics, the importance of STRs as regulatory elements was recently demonstrated(Gymrek et al. 2016). 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(Shapiro et al. 2013)(Biezuner *et al.*, accepted) and for pre-implantation genetic diagnosis(Eftedal et al. 2001).

A key challenge for STR analysis is that they undergo a 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 repeat numbers (see example in Figure 1B). Despite the value of the high polymorphicity of short unit STRs, they are commonly not used due to excessive stutter noise. Simple noise models such as highest peak do not apply to polymorphic STR since iterative stutter over amplification cycles 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(Byrd et al. 1965; Hite et al. 1996; Shinde et al. 2003; Fungtammasan et al. 2015). 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: (1) unit type length: short unit STRs (mono- and di-repeats) are more mutable than longer unit types. (2) STR length: Longer STRs (in repeat number) are more mutable than shorter STRs(Ellegren 2004). Nevertheless, despite years 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(Gymrek et al. 2012; Highnam et al. 2013; Fungtammasan et al. 2015), 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 (*e.g.* hemizygous alleles). 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.

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 synthetic library of STR plasmids, each containing a unique combination of STR type and length, spanning all naturally occurring di- repeats (namely: AC, AG, AT) in their full spectrum of their natural genomic occurrence(Subramanian et al. 2003) (Supplemental Table S1). The construct within each plasmid is sequencing-ready and includes a unique Illumina dual index combination. The plasmids served as a template to generate two PCR data sets, which correspond to two PCR amplification time points in the following amplification timeline (Figure 1A and Methods section):

T_0: Designed sequences without amplification noise as validated by clone Sanger-sequencing.

T_1: Amplification-free sequencing observed in the data obtained from Illumina sequencers. This is enabled by digestion of the plasmid using a type IIS restriction enzyme that cleaves outside of its recognition site, generating a final Illumina sequencing ready DNA, once purified.

T_2: A single extension PCR, with primers (see "Outer primers" in Figure 1A) that generate an Illumina sequencing library and incorporate a unique sequencing index combination to the sample (replacing the template index sequences).

T_3: two sequential PCRs in which the 1st PCR primers (see "Inner primers" in Figure 1A) generate a partial Illumina sequencing library (without adapters and indexes). The output of the 1st PCR is than diluted and used again in a 2nd PCR which is essentially the T_2 PCR setup (using the "Outer primers").

Overall, this experimental setup allows for a controlled amplification and sequencing of all highly mutable STRs at two independent time points (T_2, T_3) with the ability to measure the specific sequencing noise of each STR length and type (Figure 1B,C).

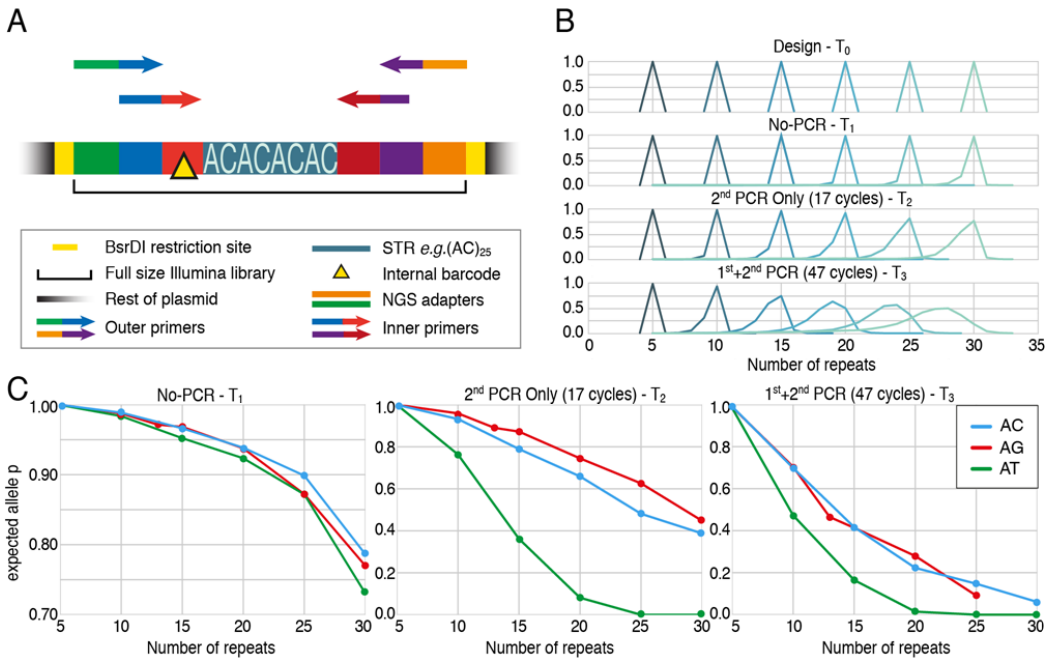

Figure 1. The synthetic STR experiment summary. A. Schematic description of the synthetic library. In each plasmid, a different synthetic STR construct was designed, synthesized and clone-sequenced for various STR types and length (dashed pink line). The STR was designed within a context of an Illumina 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 STR type histograms (normalized

by the sum of squares), 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 (bold circles) of each STR type, size and time point described as the percentage of expected p signal from overall reads.

Computational methods

The data generated for the 3 time points (T_1, T_2, and T_3) was then 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.

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, $H(t)(n)$. We label our data as t_0-t_4 in accordance with the amplification steps T_0-T_4 described above.

We chose a Markov model for the mutational process the STR undergoes during amplification (WGA, PCR and NGS). We model these processes as an iterative mutation process with multiple steps. 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 happening is dependent on the STR's current length.

We model the probability of each mutation step as a linear function $P(n)$ and fit its parameters using Broyden–Fletcher–Goldfarb–Shanno (BFGS)(H. Byrd et al. 1994) optimization algorithm with the following optimization problem:

$$\underset{P_{-3}(n),\dots,P_{+1}(n)}{\arg\max} \sum_{L_0}^{L_n} \sum_{t_0}^{t_3} d\left(H_{model}\left(t, P_{-3}(n), \dots, P_{+1}(n)\right), H_{seq}\right)$$

With $d(H_1, H_2)$ being the distance between the two histograms.

We attempted varying numbers of mutational steps and solved each as a similar optimization problem. We then assessed which of these steps were negligible and restricted our model to only $-3, -2, -1, 0, +1$ steps. These steps proved sufficient to accurately predict the majority of experimental measurements we have encountered (Figure 1). We found that our model can effectively predict both AC and AG with the same mutational steps and different probabilities, however AT's self-annealing nature as well as the stutter patterns we have measured, suggest a very different set of mutational steps such as –n/2 or even a different model. Since the occurrence of AC predominates that of AG in the human genome(Subramanian et al. 2003), in the next paragraphs we will focus on the analysis of AC.

Genotyping

To confirm the model, we propose *R&B*, a naïve genotyping algorithm, 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:

$$\arg\max_{t} d\left(H_{model}\left(t, P_{-3}(n), \ldots, P_{+1}(n)\right), H_{seq}\right)$$

We have examined multiple distance metrics for the sake of histogram comparison and found correlation (Numpy, Jones et al. 2001) as the most suitable (Supplemental Fig S1).

We than demonstrate the robustness of the model by applying it to an NGS dataset generated from a single PCR amplification, as previously described for the T_2 experiment, of 3 different templates: $(AC)_{20}$, $(AC)_{25}$ and $(AC)_{30}$, each using 3 serially diluted templates (by 10 fold each). Our model's simulated cycles linearly correlate with the actual number of amplification cycles performed on the sample (Supplemental Fig S2 A, B). We also validated the model by applying it to an NGS dataset generated from a single PCR amplification of the same three templates as above, at the same concentration but with different commercially available polymerases. We show that the model accurately captures the variability between different polymerases within a single degree of freedom, its simulated cycles (Supplemental Fig S2 A, C).

We than 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 Fig S3). 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 (Biezuner *et al.*, accepted) that consistently presented alleles A and B when genotyped as mono-allelic (Figure 2A,B,C first column). Since alleles A and B can appear at any proportion (Figure 2D), 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 A and B.

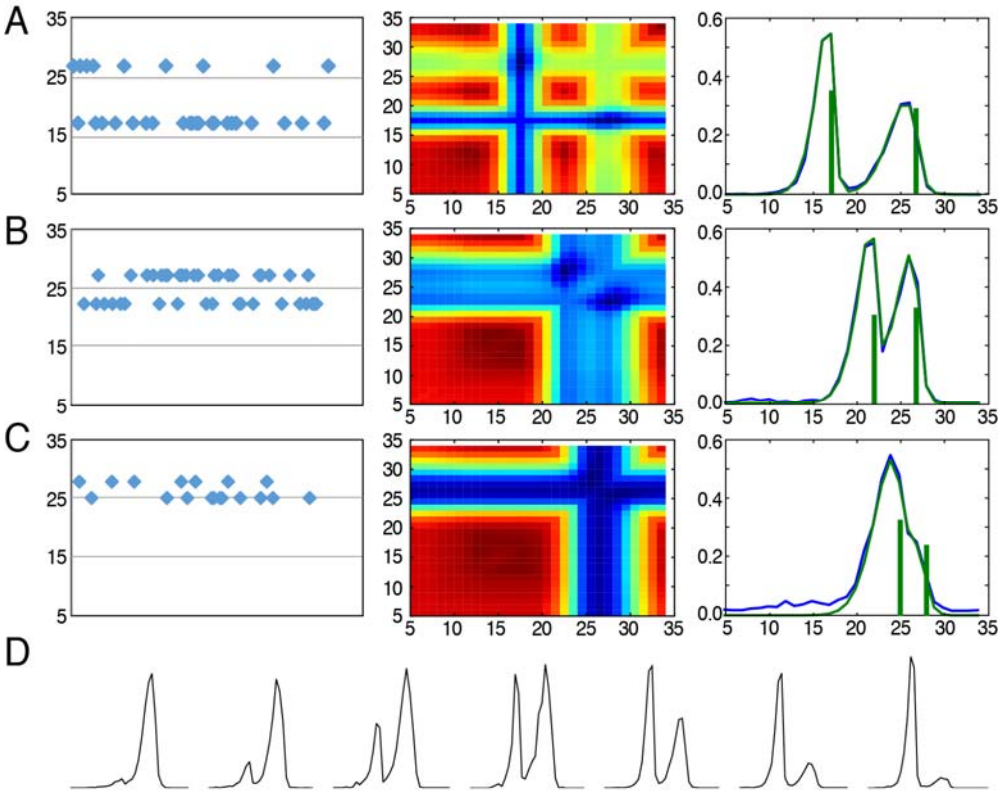

Figure 2. Biallelic genotyping using overlaid model histograms. Figure rows A, B and C show the successful genotyping of biallelic loci (AC repeats) within a SC population of H1 stem cells (Biezuner *et al.*, accepted). A, Recognizing overlapping alleles spanning 17 and 27 repeats, B, 22 and 27 repeats, 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.

Previously published STR genotyping tools (Gymrek et al. 2012; Highnam et al. 2013; Fungtammasan et al. 2015) faced two distinct problems when trying to genotype STRs from NGS data: mapping the reads to the reference genome with respect to the mutability of the STR part of the read and genotyping the often noisy populations of reads attributed to each locus. The tool "RepeatSeq"(Highnam et al. 2013), provided the first clear cut between the two problems and systematically compared multiple methods for mapping. In the dataset presented by Biezuner *et al.*, the mapping issue was 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 this data set, we can isolate the problem of genotyping stutter patterns and avoid a possible mapping bias. The dataset contains cells from a

controlled *ex vivo* cell lineage tree experiment, semi-automatically picked while documenting the sampling lineage of each analyzed cell. This known topology of individually analyzed SCs, provides a solid reference to measure any genotyping tools against. 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.

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

We define F as the *reference tree fitting*:

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

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.

In order to avoid mapping-related bias we have applied our proposed genotyping method on two histogram datasets that were calculated based on the *ex vivo* NGS data. To maintain simplicity, we only account for mono-allelic AC loci from the X chromosome of the cancerous cell line used in this experiment (human male DU145).

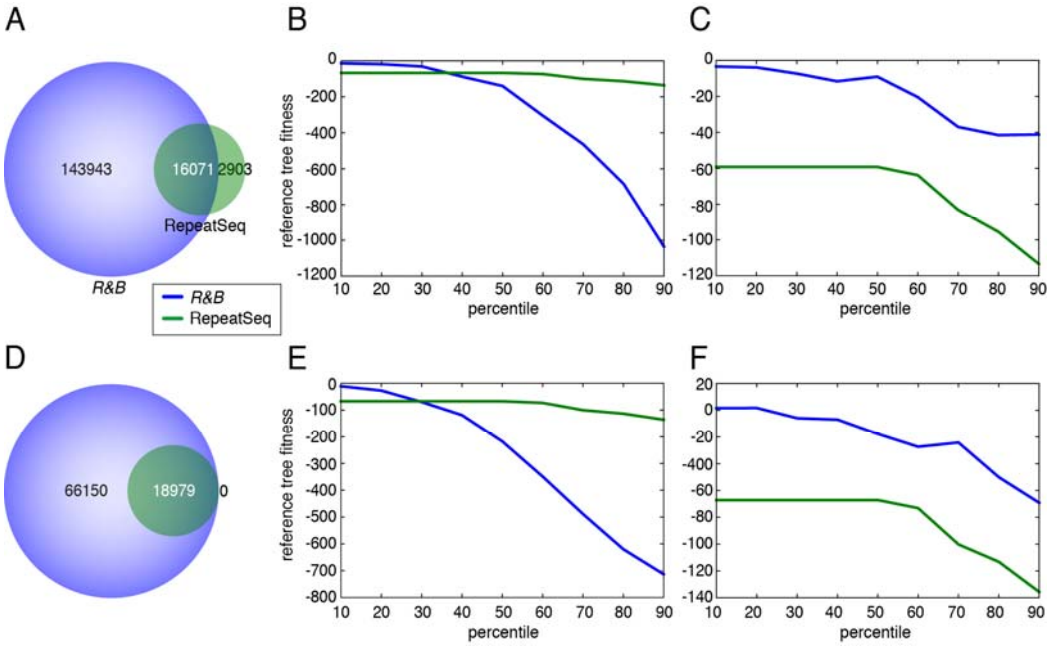

Figure 3. Genotyping results. Comparison of the proposed genotyping method, hereby *R&B*, with the RepeatSeq genotyping tool(Highnam et al. 2013). In the first row (A, B, C) we compare our genotyping methods with RepeatSeq native histograms: *R&B* with the data from the *ex vivo* paper (Biezuner *et al*., accepted) and RepeatSeq with its default mapping parameters. In the second row (D, E, F) we ran both genotyping methods on RepeatSeq's native histograms. In the first column (A, D) we view the sets of cell/locus combinations each genotyping method provided. In the second column (B, E) we plot the *reference tree fitness* scores as a function of the confidence score threshold for each genotyper. In the third column (C, F) we show the same score/threshold behavior but this time only for the cell/locus combinations that intersect with both genotypers' results.

Our results show that RepeatSeq discards over 70% of the loci it correctly mapped with over 30 reads (Figure 3D) and that the possible set of loci that can be mapped using the exhaustive method by *R&B* is over twice as big (Figure 3A). Both genotyping methods, *R&B* and RepeatSeq, provide a measure of confidence together with each locus it attempts to genotype. While these confidence metrics are 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. We first assessed the absolute performance of each tool by scoring its top % best genotyping attempts against the known tree topology (Figure 3B,E). We show that for the more confident percentiles, the *R&B* method provides genotypes that corresponds better to the tree's topology. The confidence metric for *R&B* is plainly the correlation of the predicted histogram with the reads population. To compare the lower confidence genotyping attempts of both tools despite the large difference in the number of attempts, we compared only the

cells/loci combinations where both tools provide a genotyping attempt (Figure 3C,F). Here we can see that across all confidence levels, when both tools attempt to provide a genotype, the *R&B* attempts are on average more in line with the known tree topology.

Discussion

STR usage in scientific research is increasing. High throughput sequencing opens a new frontier for STR science, both for basic(Willems et al. 2014; Gymrek et al. 2016) and for applicative research(Churchill et al. 2016; Kim et al. 2016). 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(Gymrek et al. 2012; Highnam et al. 2013; Fungtammasan et al. 2015). 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(Fungtammasan et al. 2015). 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(Mertes et al. 2011). 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(Huse et al. 2007) and inferred estimation of such noise from old Illumina platforms(Albers et al. 2011). 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(Ellegren 2004).

We provided two 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. These experiments also demonstrate 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 better enzymes for WGA and PCR for the purpose of STR analysis.

(2) 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(Highnam et al. 2013) utilizing NGS data from SC targeted enrichment data, originated from an *ex vivo* controlled cell lineage tree (Biezuner *et al.*, accepted). 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.

We acknowledge that the bioinformatics improvement we supply here is the stutter model itself, where in shotgun sequencing, highly accurate mapping tools already exist. Nevertheless, we recommend this model as an integrative step for STR noise analysis, specifically when sequenced samples undergo extensive amplification. The tolerance of our model in the analysis of noisy STR signal allows for a more flexible

experimental design and opens the gate for highly mutable STR sequencing research. In future work we will attempt to model mono repeats using a similar calibration method.

Methods

Controlled amplification 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. Internal 3-mer internal barcodes were inserted to allow for cross-contamination detection between samples.

Experimental procedure

T_1 (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℃ for 16 hours, followed by inactivation at 80℃ for 20 minutes. 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’-CTACACGACGCTCTTCCGATCTTCCTAATCTTACGCGGCCATAAC-3’ and Control_Rev:

5'-CAGACGTGTGCTCTTCCGATCATGGACAGTCTTTAAGAGCCCATC-3'(IDT), at a concentration of 1µM each. PCR reactions and purifications were performed as described in (Biezuner *et al.*, accepted): In summary, a 1st PCR of 30 cycles PCR reaction is performed in the AA chip. Following sample harvesting, purification and dilution 1:100, a 2-step 2nd PCR of 17 cycles (5 cycles with annealing temperature of 55ºC + 12 cycles with annealing temperature of 70ºC) is performed to generate a dual indexed sequencing library (note that the "Outer primers" sequences were as described for the 2nd PCR primer sequences in (Biezuner *et al.*, accepted)). The 1st PCR (in the AA chip) is done using the manufacture recommended enzyme: FastStart High Fidelity PCR System, dNTPack (Roche) while the 2nd 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 2nd PCR, each sample was purified using SPRI beads. T_2 PCR was performed by using 0.1ng-1ng of each STR plasmid as a template. Samples were processed in accordance with the T_3 2nd 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 (200-500bp) using a 2% agarose BluePippin gel cassette (Sage Science) utilizing an upgraded software that avoids blue light exposure after markers detection. Products were concentrated again (Minelute) and were sequenced by a 2X220bp sequencing (Miseq, Illumina) using a manufacture recommended sequencing primers (R1, Index) and custom R2 primer 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3' (HPLC grade, IDT).

Enzyme comparison and serial dilution validation

5 high fidelity PCR enzymes were used in this study: The two 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 performed as in the 2-step 2nd PCR reaction of step T_3, namely 20µl reactions in a 96-well format, with real time amplification tracking using SYBR green I. The following exceptions were considered: 1) Activation, elongation and final elongation were adjusted to fit each enzyme's recommended protocol. 2) Annealing

temperature from the 6th amplification step and on was according to each enzyme's elongation temperature. 3) PCR reaction was stopped when amplification reached a plateau. 4) Due to failure of dNTPack to amplify using the standard 2-step PCR protocol, we applied the same program as being performed in the 1st PCR of T_3 (in the AA chip). 5) Reactions mixes were according to manufacturer's protocols, with primer concentrations of 0.3-0.5µM, with the exception of dNTPack, which composition was according to Fluidigm's recommended reaction mixture with primer concentration of 0.1µM each and a final volume of 10.6µl.

The template for each PCR was 2µl 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 3 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.

Acknowledgements

We thank O. Bechar for the prompt design of figures. This research was supported by The European Union grants: ERC-2008-AdG (Project No: 233047) and ERC-2014-AdG (Project No: 670535); by The Israel Science Foundation grants: Individual Research Grant (Grant No: 456/13) and Joint Broad-ISF Research Grants: 422/14 and 2012/15; by The German Research Foundation DFG SH grant 867/1-1 and by The Kenneth and Sally Leafman Appelbaum Discovery Fund. E.S. is the Incumbent of The Harry Weinrebe Professorial Chair of Computer Science and Biology.

Disclosure Declaration

The authors declare no competing financial interest.

References

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.

Byrd C, Ohtsuka E, Moon MW, Khorana HG. 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.

Churchill JD, Schmedes SE, King JL, 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.

Eftedal I, Schwartz M, Bendtsen H, Andersen AN, 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.

Ellegren H. 2004. Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics* **5**: 435-445.

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

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

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ et al. 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* **48**: 22-29.

H. Byrd R, Lu P, Nocedal J, Zhu C. 1994. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* **16**: 19.

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.

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

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. In *Genome Biol*, Vol 8, p. R143.

Jones E, Oliphant T, Peterson P. 2001. SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/

Kim EH, Lee HY, Yang IS, Jung SE, Yang WI, Shin KJ. 2016. Massively parallel sequencing of 17 commonly used forensic autosomal STRs and amelogenin with small amplicons. *Forensic Sci Int Genet* **22**: 1-7.

Mertes F, ElSharawy A, Sauer S, van Helvoort J, van der Zaag P, Franke A, Nilsson M, Lehrach H, Brookes A. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics* **10**: 374-386.

Mirkin SM. 2007. Expandable DNA repeats and human disease. *Nature* **447**: 932-940.

Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. 2014. Microsatellite instability detection by next generation sequencing. *Clin Chem* **60**: 1192-1199.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*.

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.

Subramanian S, Mishra RK, Singh L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**: R13.

Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res* **24**: 1894-1904.