BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 12 2014, pages 1651–1659
doi:10.1093/bioinformatics/btu068

Genome analysis

Advance Access publication February 13, 2014

TSSV: a tool for characterization of complex allelic variants in pure and mixed genomes

Seyed Yahya Anvar[1,2,*], Kristiaan J. van der Gaag[1], Jaap W. F. van der Heijden[1], Marcel H. A. M. Veltrop[1], Rolf H. A. M. Vossen[2], Rick H. de Leeuw[1], Cor Breukel[1], Henk P. J. Buermans[1,2], J. Sjef Verbeek[1], Peter de Knijff[1], Johan T. den Dunnen[1,2] and Jeroen F. J. Laros[1,2,3]

[1]Department of Human Genetic, [2]Leiden Genome Technology Center, Leiden University Medical Center, Leiden, 2300 RC, The Netherlands and [3]Netherlands Bioinformatics Centre, Leiden, The Netherlands

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Advances in sequencing technologies and computational algorithms have enabled the study of genomic variants to dissect their functional consequence. Despite this unprecedented progress, current tools fail to reliably detect and characterize more complex allelic variants, such as short tandem repeats (STRs). We developed TSSV as an efficient and sensitive tool to specifically profile all allelic variants present in targeted loci. Based on its design, requiring only two short flanking sequences, TSSV can work without the use of a complete reference sequence to reliably profile highly polymorphic, repetitive or uncharacterized regions.

Results: We show that TSSV can accurately determine allelic STR structures in mixtures with 10% representation of minor alleles or complex mixtures in which a single STR allele is shared. Furthermore, we show the universal utility of TSSV in two other independent studies: characterizing *de novo* mutations introduced by transcription activator-like effector nucleases (TALENs) and profiling the noise and systematic errors in an IonTorrent sequencing experiment. TSSV complements the existing tools by aiding the study of highly polymorphic and complex regions and provides a high-resolution map that can be used in a wide range of applications, from personal genomics to forensic analysis and clinical diagnostics.

Availability and implementation: We have implemented TSSV as a Python package that can be installed through the command-line using pip install TSSV command. Its source code and documentation are available at https://pypi.python.org/pypi/tssv and http://www.lgtc.nl/tssv.

Contact: S.Y.Anvar@lumc.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 17, 2013; revised on January 6, 2014; accepted on January 27, 2014

1 INTRODUCTION

As a consequence of various mechanisms such as DNA recombination, replication and repair-associated processes, the spectrum of human genetic variation ranges from single nucleotide differences to large chromosomal events. Among the different types of genetic changes, repetitive DNA sequences show more polymorphism than single nucleotide variants (Conrad *et al.*, 2010; Hinds *et al.*, 2006; Iafrate *et al.*, 2004; Kidd *et al.*, 2008; Redon *et al.*, 2006; Sebat *et al.*, 2004; Tuzun *et al.*, 2005), and they are important in human diseases (Conrad *et al.*, 2010; de Cid *et al.*, 2009; Girirajan *et al.*, 2011; Hollox *et al.*, 2008; McCarroll *et al.*, 2009; Pinto *et al.*, 2010), complex traits and evolution (Mills *et al.*, 2011; Stephens *et al.*, 2011; Sudmant *et al.*, 2010). In particular, microsatellite variants, also known as short tandem repeats (STR), and their expansion/shortening have been linked to a variety of human genetic disorders (Mirkin, 2007; Pearson *et al.*, 2005; Sutherland and Richards, 1995), and have been used in genotyping (Kimura *et al.*, 2009; Weber and May, 1989) and forensic DNA fingerprinting studies (Kayser and de Knijff, 2011; Moretti *et al.*, 2001).

Because of the repetitive nature of STRs and often the low level of complexity of the DNA sequences in which they occur (Treangen and Salzberg, 2012), characterization of STR variability and understanding of their functional consequences are challenging (Weischenfeldt *et al.*, 2013). So far, sequencing-based strategies have focused on reads mapped to the reference genome and subsequent identification of discordant signatures and classification of associated STRs (Medvedev *et al.*, 2009; Mills *et al.*, 2011). Yet, the mainstream aligners, such as BWA (Li and Durbin, 2009) or Bowtie (Langmead and Salzberg, 2012), do not tolerate repeats or insertions and deletions (indels) as a trade-off of run time (Li and Homer, 2010). This limitation leads to ambiguities in the alignment or assembly of repeats which, in turn, can obscure the interpretation of results (Treangen and Salzberg, 2012). Moreover, the current human genome reference still remains incomplete and provides only limited information on expected and potentially uncharacterized STRs in different individuals (Alkan *et al.*, 2011; Iafrate *et al.*, 2004; Kidd *et al.*, 2008; Sebat *et al.*, 2004). Consequently, STRs are not routinely analyzed in whole-genome or whole-exome sequencing studies, despite their obvious applications and their role in human diseases, complex traits and evolution.

Here, we present a method for targeted profiling of STRs that reports a full spectrum of all observed genomic variants along with their respective abundance. Our tool, TSSV, can accurately

*To whom correspondence should be addressed.

profile and characterize STRs without the use of a complete reference genome, and therefore minimizes biases introduced during the alignment and downstream analysis. TSSV scans sequencing data for reads that fully or partially encompass loci of interest based on the detection of unique flanking sequences. Subsequently, TSSV characterizes the sequence between a pair of non-repetitive flanking regions and reports statistics on known and novel alleles for each locus of interest. We show the performance of TSSV on robust characterization of all allelic variants in a given targeted locus by its application in several case studies: forensic DNA fingerprinting of mixed samples by STR profiling, characterization of variants introduced by transcription activator-like effector nucleases (TALENs) in embryonic stem (ES) cells and detailed characterization of errors derived from a next-generation sequencing (NGS) experiment.

2 MATERIALS AND METHODS

2.1 TSSV algorithm

The algorithm expects a FASTA file containing sequencing data and a library containing a list of loci of interest that are described by two unique sequences flanking a target locus in the form of a simple regular expression. The description of targeted loci consists of a series of triplets (i.e. CTTA 2 5), each containing a sequence followed by two integers that denote the minimum and maximum number of times the preceding sequence is expected. The notation of expected alleles is then compiled into a regular expression that is used to distinguish between known and new alleles. It is important that a library that contains a description of loci of interest according to the aforementioned instruction should be customized and provided. TSSV reports an overview of marker pair alignments and a detailed description of the identified alleles and their respective frequency per strand. TSSV also provides supporting reads of each locus of interest in separate FASTA files.

TSSV is an open source Python package that can be easily incorporated in any standard NGS pipeline. In addition, we have made the Python package *fastools* available at *https://pypi.python.org/pypi/fastools*. *fastools* offers a series of functions to manipulate, characterize, sanitize and convert FASTQ/FASTA files to other formats. Therefore, it can be used to convert FASTQ files to TSSV desired format (FASTA). For further information on usage and generated data see Supplementary Table S1.

2.1.1 Marker alignment Each pair of markers (unique flanking sequences) is aligned to the reads by using a semi-global pairwise alignment, a modified version of the Smith–Waterman algorithm (Smith and Waterman, 1981). The alignment matrix is initialized with penalties only for the aligned sequence and not for the reference sequence. By using this approach, we can use the alignment matrix to calculate the edit distance between the aligned sequence and all substrings of the reference sequence. Finally, TSSV uses the alignment matrix to select the rightmost alignment with a minimum edit distance. To guarantee symmetry with regard to reverse complement sequences, TSSV aligns the reverse complement of the right marker to the reverse complement of the reference sequence.

2.1.2 Allele identification Once TSSV successfully aligns a marker pair to either the forward or the reverse complement of the reference sequence, the region of interest is selected by extracting the sequence between the alignment coordinates, which is then converted to the forward orientation. The target variable sequence is then matched to the regular expression of the corresponding marker pair for classification as either a known or a new allele. In case of partial identification of markers (i.e. only the left or right marker of the pair is identified), the input sequence is flagged as having either no beginning or no end. The assessment of required runtime for TSSV to identify alleles in datasets with different sequencing depth is provided in Supplementary Figure S1. Each dataset is profiled to characterize 16 allelic STR structures. It should be noted that currently TSSV uses a single processor for the analysis.

2.1.3 Annotations Once a list of new alleles is constructed, TSSV uses a revised version of the *Mutalyzer* online service (Wildeman *et al.*, 2008; *https://mutalyzer.nl*) to describe all observed variants compared with the reference sequence. *Mutalyzer* provides a description of observed variants according to the Human Genome Variation Society format for sequence variant description. This can be used to provide an overview of most frequent mutations that are observed within each locus of interest.

2.1.4 Interpretation guidelines TSSV provides the frequency in which each allelic structure is observed on plus and minus strand. Based on the experimental design, the frequencies of allelic variants and the balance between supporting reads on the plus and minus strand can aid the identification of potential sequencing biases. Moreover, based on the choice of sequencing technology, homopolymers are prone to introducing artificial allelic structures, so it is advised, when possible, to allow for a tolerance of a few base difference in the homopolymer length while describing targeted loci. The estimation of a lower boundary for the identification of variant alleles is subject to the experimental design. Thus, sequencing of control samples, if possible, can aid a more reliable analysis by ruling out potential slippage and background noise.

2.1.5 Availability TSSV is available at *http://www.lgtc.nl/tssv* and *https://pypi.python.org/pypi/tssv*. It can also be installed through the command line: *pip install tssv*. All original datasets and the analysis results can be obtained from figshare (*http://www.figshare.com*): detection of STRs, SNPs and short indels (Anvar, 2013a), determining *de novo* structural variations (SVs) in TALEN-treated ES cells (Anvar, 2013b), characterization of STRs (Anvar, 2013c) and detection of systematic errors in PGM (Anvar, 2013c).

2.2 Library preparations and sequencing

STR PCR products for sequencing were generated using the *Powerplex® 16-kit* from *Promega* (commercial assay designed and optimized for fluorescent dye-based fragment analysis of STR loci) and were purified with Ampure XP beads according to manufacturer's protocol. Library preparation was performed using the Rapid Library Preparation Kit (Roche). Emulsion PCR and sequencing were performed on the FLX Genome Sequencer (454/Roche) according to the protocol provided by the manufacturer.

PCR products for sequencing of all other samples on the Personal Genome Machine (PGM, IonTorrent) were prepared using the Ion Plus Fragment Library Kit or amplicon fusion primers. Emulsion PCR was performed using the OneTouch (OT1, IonTorrent). Sequencing was performed according to LifeTech protocol using the Ion PGM™ 200 Sequencing Kit. PCR reaction was done in 10 µl containing 1× FastStart High Fidelity reaction buffer (Roche), 1.8 mM $MgCl_2$, 2% DMSO, 200 µM dNTPs, 0.5 U FastStart High Fidelity Enzyme Blend (Roche), 20 ng DNA, 300 nM universal barcoding primer, 300 nM reverse target primer and 30 nM forward target primer. After 10 min of initial denaturation at 95°C, 30 PCR cycles were performed at 20 s 95°C, 30 s 60°C and 40 s 72°C. Primer sequences are provided in Supplementary Table S2.

2.3 TALEN design and transfection

The TALENs -pair targeting intron 52 of the human *DMD* gene (*hDMD*) was designed using the TALEN toolbox described by Cermak *et al.* (2011). Next, hDMD/*mdx* ES cells (t Hoen *et al.*, 2008; Veltrop *et al.*, 2013) were transfected with the TALENs plasmids without any

homologous recombination vector. ES cells were routinely cultured on murine embryonic fibroblast (MEF) feeder cells in knockout DMEM supplemented with 2 mM L-glutamine, 1 mM sodium pyruvate, non-essential amino acids, 50 units of penicillin as well as streptomycin, 1000 units of leukemia inhibitory factor and 10% fetal bovine serum (FBS Gold, all from Life Technologies Ltd). Per TALEN, total of 750 ng in 1.5 µg of DNA was used to transfect 1 000 000 hDMD/*mdx* ES cells using Lipofectamin 2000 (Invitrogen). DNA-Lipofectamin 2000 suspension was prepared in serum and antibiotic-free medium according to the supplier's manual Cells were incubated for 30 min in suspension with the DNA-Lipofectamine mixture and then plated in two 9 cm culture dishes coated with MEF in regular ES culture medium. ES cells were cultured for a week, and DNA was isolated from a pool of 1500 ES clones. This DNA was then prepared for sequencing using IonTorrent PGM according to the instrument guidelines.

3 RESULTS

3.1 Characterization of STRs

We tested the performance of TSSV in characterizing known STRs from Roche/454 targeted sequencing data of 16 STR loci, amplified in a multiplex reaction. To demonstrate the added value of TSSV over mainstream aligners, we generated four sequencing libraries of which two consisted of pure individual samples and two mixtures in the ratios of 50:50 and 90:10 with comparable depth of coverage (Supplementary Table S3). A full spectrum of STR structures and their abundance was generated after a semi-global alignment of the 25 bp flanking regions adjacent to the STR structure, with tolerance of up to three mismatches (Fig. 1A). On average, 8% of reads remained uncharacterized, mostly because the sequences did not cover both flanking reference sequences or that sequences contained too many mismatches for regions that are required for identification of unique flanking reference sequences (Supplementary Table S3). The PCR product used for preparing the sequencing libraries were generated using the *Powerplex 16-kit* from *Promega*, which is an assay designed and optimized for fluorescent dye-based fragment analysis of STR loci. This resulted in a strong imbalance in sequencing yield between STR markers with different dyes in the fragment analysis (Supplementary Table S3). Thus, we restricted the analysis to the three markers with highest coverage (D3S1358, TH01 and D13S317). Frequencies of the observed alleles were interpreted to distinguish actual alleles from slippage artifacts (Supplementary Tables S4–S6).

For D3S1358 ($TCTA_1TCTG_{1\text{-}3}TCTA_{12\text{–}13}$), TSSV robustly identified the STR structure associated with each of the samples, with >91% of reads supporting the presence of two STR alleles (Fig. 1B). In addition, TSSV could pick up a minor frequency (7.25%) for alternative STR structures, in which the DNA amplicons show false STR structures because of DNA polymerase slippage during the amplification (Ellegren, 2004; Hauge and Litt, 1993). Despite the presence of PCR amplification artifacts, the major and minor STR structures in balanced and more extreme mixtures (50:50 and 90:10, respectively) could accurately be identified by TSSV (Fig. 1B and Supplementary Table S4).

We next explored whether TSSV can correctly detect alleles of more complex cases differing based on STR length ($ATCT_{12}ATCA_2$ and $ATCT_{11}ATCA_3$) or composition ($CATT_9$ and $CATT_3CAT_1CATT_6$) as well as mixtures that shared one

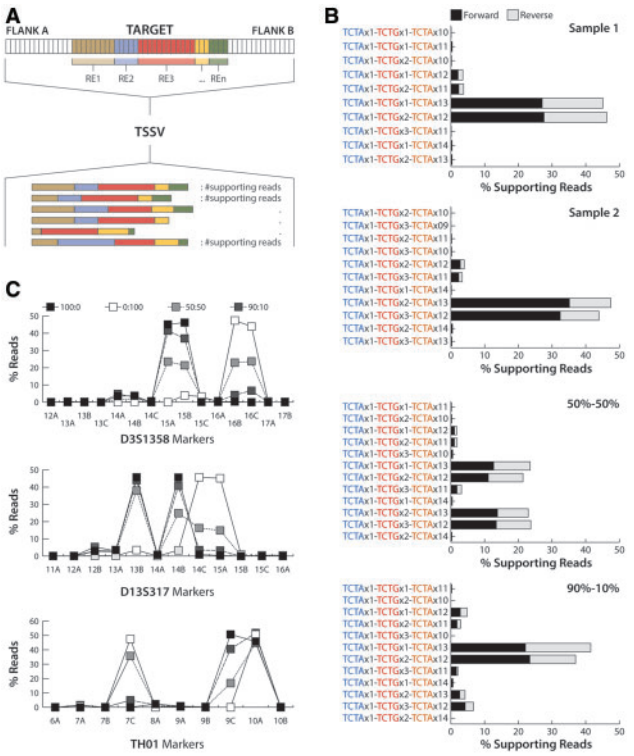

Fig. 1. Characterization of allelic STR structures in samples and their mixtures of differing ratios (**A**) Schematic representation of STR structure identification and quantification. After proper alignment of two flanking sequences, TSSV performs a strand-specific classification and quantification of repetitive elements (RE) that constructs a given STR-structure. (**B**) The number of sequencing reads that support the presence of different allelic D3S1358 STR structures on both strands. Pure samples and their mixtures in two different ratios are presented separately. (**C**) The proportion of reads that support different allelic STR structures for three most abundant markers (D3S1358, D13S317 and TH01). STR markers differ in complexity based on STR length or composition as well as mixtures in which one allelic STR structure in shared

allelic STR structure ($CATT_3CAT_1CATT_6$). Markedly, TSSV could correctly detect, characterize and quantify reads supporting all STR alleles, including mixtures with only 10% representation of the minor alleles (i.e. D3S1358 markers) and more complex mixtures (TH01 markers) where a single STR allele is shared (Fig. 1C, Supplementary Tables S4–S6 and Supplementary Figs S2–S4). Results of the remaining STR markers are provided in supplementary materials, Supplementary Tables S7–S17.

3.2 Determining *de novo* structural variations in TALEN-treated cells

TALENs have shown promising potential in site-specific genome editing (Boch, 2011; Cermak *et al.*, 2011; Miller *et al.*, 2011; Zhang *et al.*, 2011). Their modular structure enables simple construction of TALENs that can specifically recognize virtually any DNA sequence of interest. On delivery of a TALENs-pair, a double strand break is introduced that is repaired by non-homologous end-joining, introducing a large variety of

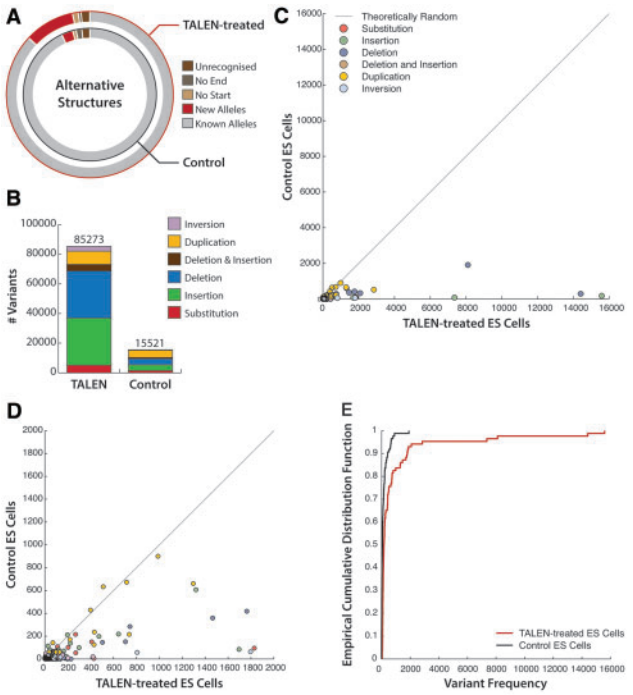

Fig. 2. Variant characterization and quantification of TALEN-treated and Control ES Cells. (**A**) Basic statistics of the TSSV analysis. Pie charts show the proportion of sequencing reads that support the presence of new alleles in TALEN-treated (outer circle) or control (inner circle) ES Cells. *No Start* and *No End* fragments represent reads in which one of the flanking sequences was not recognized. (**B**) Total number of variants in TALEN-treated ES Cells and Controls, grouped by type. (**C**) Comparative analysis of the number of occurrences for individual variants in both samples. Data points are colored based on the type of variation. (**D**) Zoomed in scatter plot for variants with frequencies lower than 2000. (**E**) Empirical cumulative distribution of variant frequencies for TALEN-treated (red) and Control ES Cells (black). Kolmogorov–Smirnov test was performed to assess if two distributions are significantly different

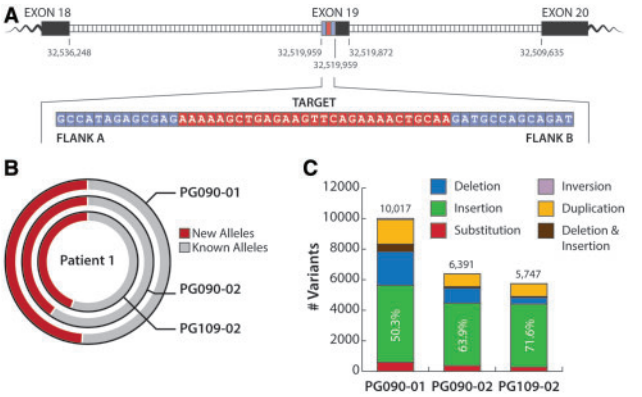

Fig. 3. Identification of mutations within exon 19 of DMD gene. (**A**) Schematic representation of the locus of interest for resequencing, the design of unique flanking sequences (blue), and the targeted region (red) to be profiled using TSSV. (**B**) Pie charts show the proportion of reads that support the presence of new alleles (red) in sequencing library of patient 1. Pie charts represent different sequencing runs (PG090 or PG109) or the base-calling algorithm used during the primary analysis (01 or 02). The two most outer pie charts are sequencing reads from the same PGM IonTorrent run processed using two different versions of base-calling algorithm. The most inner pie chart represents an independent run of the same library. (**C**) Number of observed variants separated by variation type. Percentages show the proportion of insertion events from the total number of variants in each set

mutations (Supplementary Fig. S5). Because the method lacks a positive selection procedure, the applicability depends largely on its efficacy. We used TSSV to estimate the efficiency of genome editing in ES cells from a mouse model with the *hDMD*, stably integrated in the mouse genome (t Hoen *et al.*, 2008), and determine the utility of an assembled TALEN pair (Supplementary Table S18) in introducing mutations within targeted intron 52 of the *hDMD* (Supplementary Fig. S6).

For 100 000 TALENs-transfected and non-transfected (control) ES cells, a 135 bp fragment encompassing the entire targeted locus was PCR amplified and sequenced using the IonTorrent PGM (Supplementary Table S19). The targeted locus was covered over 450 000 times, which allows for precise detection and characterization of any variant present. From the control ES cells, we determined a background of 3.1% of reads that contain at least one mismatch, derived from sequencing errors and potential spontaneous mutations (Fig. 2A and Supplementary Table S19). In TALENs-treated ES cells, the rate of sequencing reads that contain at least one mismatch was 11.4%, almost 4-fold higher than controls (Fig. 2A). The majority of mutations introduced by TALENs pair were small insertions and deletions (75.6%; excluding duplications) (Fig. 2B), which is consistent with the expected type of variants introduced by TALENs (Cermak *et al.*, 2011). The frequency in which individual variants occurred was specific to TALEN-treated ES cells, even for those that were observed with very low frequency (Fig. 2C and D). However, we observed a few mutations that were not specific to TALEN-treated ES cells (Fig. 2D). These were mainly duplications that arose from inaccurate detection of homopolymer stretches. Overall, TSSV results indicate significant enrichment ($P = 2.85E\text{-}09$; Kolmogorov–Smirnov test) of variants in TALEN-treated ES cells as compared with controls (Fig. 2E). Furthermore, TSSV reported a list of the most frequent variants and cleavage sites, majority of which were either exclusive to TALEN-treated ES cells or with over 3-fold higher frequency in TALEN-treated ES cells than controls (Supplementary Fig. S7). The IonTorrent variant caller (version 3.2) did not report any variant because of the nature and frequency of variants introduced by TALENs.

3.3 Detection of systematic errors in PGM IonTorrent

During the targeted IonTorrent resequencing of exon 19 of the *DMD* gene (X-chromosome) in five male patients and a female carrier, we observed a number of shared and unexplained heterozygous variants given that male patients have only one X-chromosome and *DMD* gene does not locate within pseudo-autosomal regions. We used TSSV to provide a high-resolution map of all sequence variants as a way to understand the origin of these artifacts (Fig. 3A). To assess the reproducibility of our findings, we performed two independent IonTorrent PGM

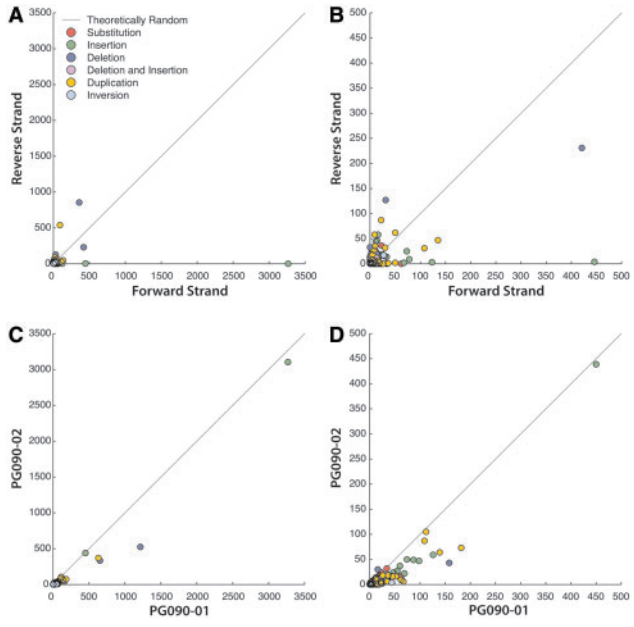

Fig. 4. Comparative analysis of observed mutations. (**A**) The total number of occurrences for variants to be observed on plus or minus strands is compared. Data points are colored based on the variation type. (**B**) Zoomed in scatter plot for variants with strand-specific frequency <500. (**C**) Sequencing data from the same sequencing run (PG090) are assessed for frequency of observed variants after the use of two different versions of base-calling algorithm. (**D**) Zoomed in scatter plot for variants with frequency <500 compared between two datasets generated using different base-caller algorithms

sequencing runs (PG090 and PG109). Two different versions of the IonTorrent base-calling algorithm were used for PG090 (versions 2.2 and 3.0) while PG109 was only processed by version 3.0 (sequencing run was carried out after the upgrade of the IonTorrent Suit). The three datasets enabled us to investigate potential artifacts derived from sequencing and/or different base-calling algorithms. Our first observation indicated a significant decrease in the total number of reads (average of 11.3 and 13.3% in respective to different runs and base-calling algorithms) that were recognized per individual (Supplementary Table S20). We also noticed a significant difference in the fraction of reads per dataset (44.3, 40.3 and 48.7%) that were reported as new alleles, having at least one mismatch with the reference sequence (Fig. 3B).

We observed a significant reduction of variants (36.2%) after adoption of the version 3.0 base-caller, mainly affecting the level of deletions and duplications calls (Fig. 3C). This prominent decrease (68.3 and 48.9%) arises from improvement of the algorithm in determining the length of homopolymer stretches. Notably, the majority of other variants were single nucleotide insertions (excluding duplications and indels) that remained at a comparable rate across different datasets (Fig. 3C). Next, we assessed the strand specificity of the variants based on the sequencing direction. Interestingly, while the majority of variants showed a similar frequency in both directions, the most frequent variants showed a clear imbalance between forward and reverse strand (Fig. 4A and B). The observed strand-specific bias was reproducible and was not influenced by the software version, as it was observed in all three datasets (Fig. 4C and D and Supplementary Figs S8–S11).

To study the possible nucleotide-specific biases, we quantified the frequency of all calls that predominantly occurred on one strand. Despite slight variation, substitutions were observed on both strands at a comparable rate. However, in each dataset, the majority of substituted bases were 'A's (59.2, 61.3 and 64.9%) and 'T's (28.1, 23.0 and 20.5%) that were predominantly substituted to 'G' and 'C', respectively (Fig. 5A). Insertions were primarily observed on the forward strand (94%, on average) while 'A' remained as the most affected base (77.7%, on average) across all samples (Fig. 5B and Supplementary Figs S9–S11).

We also observed a slight enrichment of deletions and duplications on the reverse strand that were more pronounced in PG109-02 (Fig. 5C and D). Consistently, the most affected base was 'A', which was mainly the result of under- or over-calling of 'A' homopolymers. We used TSSV to report a list of most occurring variants across different samples. A single 'A' nucleotide insertion at cycle 52 was by far the most predominant variant that occurred exclusively on the forward strand (Fig. 6A). In fact, irrespective of co-occurrence of this insertion with any other variants, the new observed sequence remains strand specific (Fig. 6A). This cannot be explained from a biological standpoint and can only arise from a sequencing error. Moreover, we did not observe any variation after sequencing the same library with Sanger sequencing (Fig. 6B), ruling out the possibility of artifacts introduced by sample preparation and PCR amplification.

3.4 Comparative analysis of TSSV performance

To our knowledge, lobSTR (Gymrek *et al.*, 2012), STRait Razor (Warshauer *et al.*, 2013) and RepeatSeq (Highnam *et al.*, 2013) are currently the most recent and frequently used STR profiling tools. STRait Razor has limited functionality and only provides an estimated copy number of major STR units. Therefore, we could only compare the performance of TSSV only with lobSTR and RepeatSeq. As lobSTR relies on alignment of sequencing reads to the predefined and indexed STR reference sequences, lobSTR outperformed TSSV in recognizing partial reads, containing only one of the two flanking sites required by TSSV (Supplementary Table S21). Concordantly, lobSTR accepted 1288 reads for the D3S1357 STR locus that were not reported by TSSV. However, TSSV performed significantly better on more complex STR loci (Supplementary Tables S21–S22). Across all four datasets, TSSV identified on average 2471 and 2353 reads in excess of what was recognized by lobSTR for the D13S317 and TH01 STR loci, respectively. This difference is mainly derived from increasingly problematic alignments in lobSTR that is also reflected in inaccurate estimation of STR copy number for TH01 and D13S317 markers in pure samples (Supplementary Table S22). In addition, lobSTR does not provide information on allelic STR structure, as it only reports the copy number of the major and uninterrupted STR unit and ignores the information from other variable elements or variants outside the STR itself. Consequently, lobSTR failed to accurately detect the presence of mixed simples even in cases in which samples were mixed 50:50 (Supplementary Table S22).

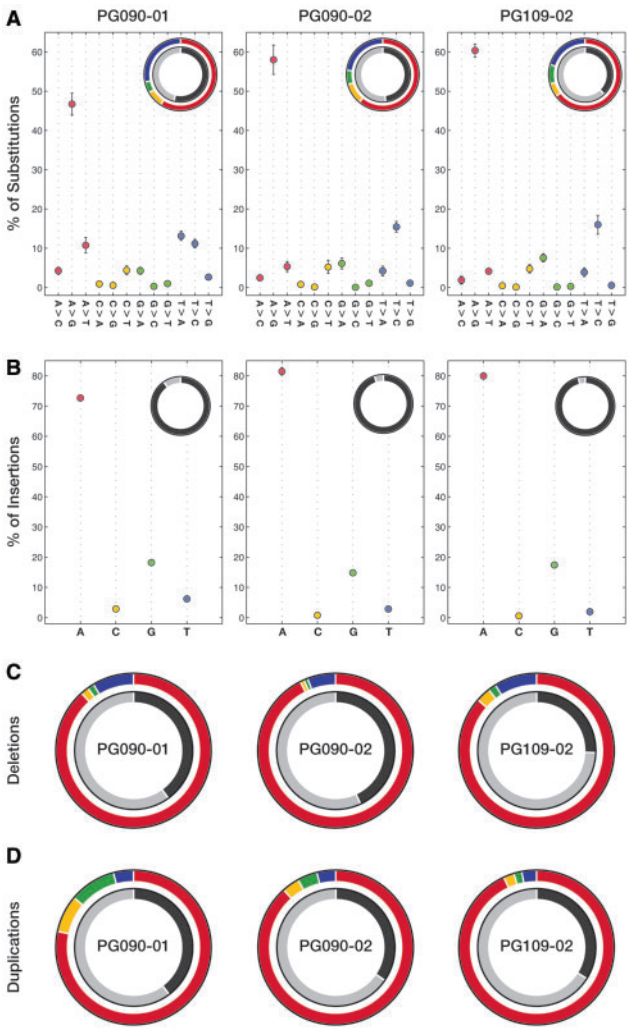

Fig. 5. Strand-specific and nucleotide-dependent variation patterns in targeted sequencing data. **(A)** Breakdown of substitution type frequencies across all samples. Pie charts depict substitution events per nucleotide (outer circle) and strand (inner circle) for each dataset. The outer pie charts illustrate the proportion of substitutions based on the preferred nucleotide to which substitutions are made. A, C, G and T nucleotides are reflected in red, yellow, green and blue colors, respectively. Black and gray colors represent the plus and minus strands, respectively. **(B)** Breakdown of insertion frequencies per nucleotide across all samples. Pie charts represent the proportion of insertion events per nucleotide (outer circle) and strand (inner circle) for each dataset. **(C)** The fraction of deleted bases is shown in pie charts based on the nucleotide (outer circle) and strand (inner circle) for each dataset. **(D)** The amount of duplications per nucleotide (outer circle) and strand (inner circle) is presented for three datasets used in this study

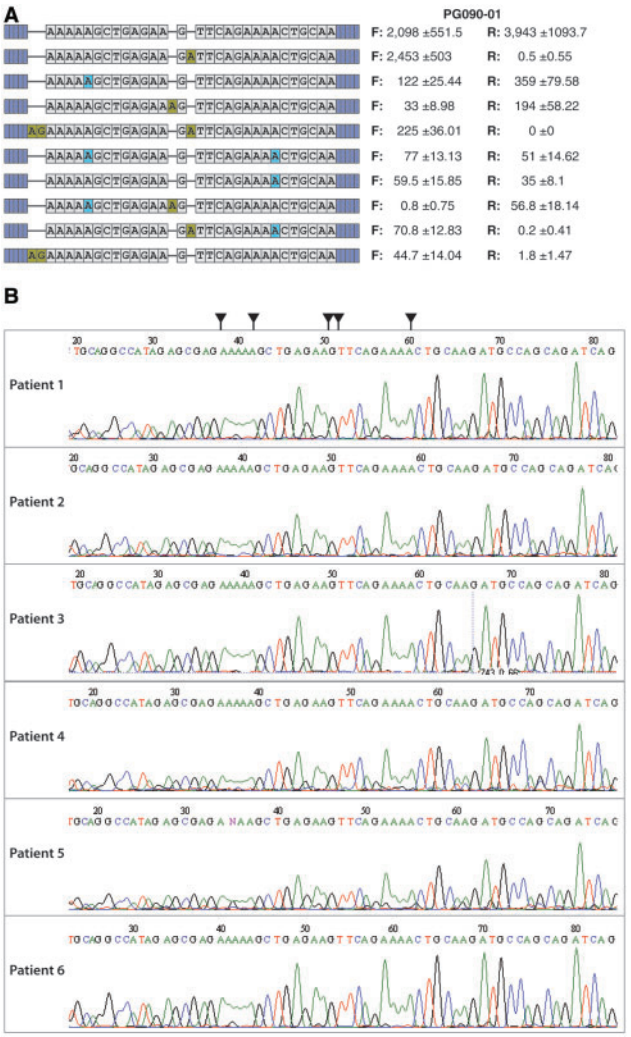

Fig. 6. Most abundant sequences and validation by Sanger sequencing. (**A**) A list of most abundant sequences across different samples. The observed occurrence of each new allele is quantified per strand along with corresponding standard deviation. (**B**) The result of Sanger sequencing of the target locus is presented per sample. Arrows mark the location in which mutations are reported based on the PGM IonTorrent sequencing data

Although the information on strand specificity of the aligned reads is present in the alignment file, unlike TSSV, lobSTR does not provide the frequency in which each STR structure is observed. This is an important measure to detect inconsistencies and to rule out potential artifacts. RepeatSeq requires aligned data and uses predefined regions to characterize observed STR alleles. Thus, reads were mapped to the reference genome (hg19) using GS mapper, specifically designed for 454 sequencing data. RepeatSeq reported results for only one STR locus (D8S1179), despite sufficient coverage for a number of STR loci in the BAM file. After manipulating the region descriptions, we could not improve the efficiency of RepeatSeq in identifying the targeted STR loci. Thus, the result of RepeatSeq could not be used for a conclusive comparison with TSSV.

4 DISCUSSION

In the past decade, advances in sequencing technologies as well as computational analysis tools have enabled the study of genomic variations to dissect the mechanisms by which they exert their function in the case of human diseases, evolution and other complex traits. Despite this unprecedented progress, structural

variations and repetitive DNA sequences (such as STRs) or coupling of *de novo* mutations present major obstacles for accurate and reliable allelic analysis (Alkan *et al.*, 2011; Gymrek *et al.*, 2012; Kidd *et al.*, 2008; Treangen and Salzberg, 2012; Weischenfeldt *et al.*, 2013). In particular, most computational tools are not ideal to identify STRs because of biases introduced during alignment as well as strong reliance of algorithms on coverage depth or the presence of split-reads. Here, we present a method (TSSV) that provides a high-resolution map of allele-specific genomic variants within targeted loci of interest. Our approach does not rely on the use of a complete reference sequence to reliably profile highly polymorphic sequences (such as STRs) or uncharacterized variants at a single-nucleotide resolution. However, it does require two unique flanking sequences that harbor the region of interest to identify supporting reads. We assess the performance of TSSV on profiling known allelic STR structures across pure samples from a single individual as well as mixed samples with variable abundance. Of 16 allelic STR structures that were targeted for sequencing, six STR loci were sufficiently covered so that the associated allelic STR structures could be reliably resolved. The strong imbalance between yield of STR markers is because of the assay (designed and optimized for fluorescent dye-based fragment analysis of STR loci) used for preparing the sequencing library. We show that sensitivity of TSSV in determining allelic STR structures exceeds mixtures with only 10% representation of minor alleles and more complex mixtures in which a single STR allele is shared. The lower boundary of detecting minor allele frequencies is subject to experimental design and the complexity of the targeted locus that may result in variable rate of slippage and background noise. Our detailed analysis of three STR loci provides significant insights into forensic DNA fingerprinting of mixed samples while it confirms the feasibility of TSSV to profile causal allelic expansion of triplet, tetranucleotide or more complex repeat structures in variety of human disorders (Brook *et al.*, 1992; Dere *et al.*, 2004; Kremer *et al.*, 1991; Mahadevan *et al.*, 1992; Mirkin, 2007; Pearson *et al.*, 2002; Verkerk *et al.*, 1991).

Second, we sought to profile and annotate the full spectrum of *de novo* mutations introduced by TALENs that specifically target intron 52 of hDMD in mouse ES cells. The applicability of designed TALENs to introduce mutations in a targeted locus largely depends on its efficacy because this method lacks a selection procedure. Detected TALEN-specific editing events were almost exclusively insertions and deletions that fit the expected mutation profile of TALENs (Cermak *et al.*, 2011). Although it has recently been reported that TALENs induce insertions at a much lower frequency than deletions (Kim *et al.*, 2013), we have observed an extremely balanced rate of insertion and deletion events (37.26% versus 37.20%, respectively). Nevertheless, TALENs-induced deletions tend to affect more bases than insertions. We show that TSSV can resolve difficult-to-call editing events that affect the length of homopolymers based on the variant frequency in TALEN-treated ES cells versus controls. Moreover, the result of TSSV analysis of TALEN-treated and control ES cells suggests that observed *de novo* structural variants are predominantly caused by initiation of a double-strand break that is repaired by non-homologous end-joining mechanism and are not the result of sequencing errors. Notably, the IonTorrent variant caller failed to identify any of the observed variants because of their complexity, and therefore does not provide any information on *de novo* allelic structures that were introduced.

As laboratories begin to generate deep coverage sequencing data to identify low frequent mutations (i.e. cancer genomics), the robustness and accuracy of NGS technology and library preparation methods has become vital (Costello *et al.*, 2013). After running TSSV on a third dataset to identify potential causal mutations in samples from five DMD patients and one female carrier, we observed numerous systematic errors introduced by the IonTorrent PGM sequencer or the base-calling algorithms. The number of sequencing reads that support the presence of a new allele was in excess of 45% while no mutation was found after Sanger sequencing of the same libraries. Moreover, the amount of allelic discordant reads were unexpected and could not be biologically explained as five out of six samples were derived from male patients who are expected to have only one copy of the X-chromosome. Across all samples, the majority of detected variations were single nucleotide insertions (~62%), excluding duplications, that were mostly the result of a single 'A' insertion (78%). Surprisingly, insertions were predominantly specific to the plus strand (94%) that can be the result of flow order in specific sequence contexts. Although the second base-caller improved the deletions and duplications rates that were derived from over- or under-calling of homopolymers, the insertion rates remained unchanged. We further observed a preference for erroneous substitution events that were more pronounced in the second base-caller. However, we were unable to identify motifs that may be associated with observed biases. We argue that the result of TSSV analysis and its ability to provide a high-resolution map of variants ever more highlights the importance of robust and vigorous assessment of downstream analysis as we generate volumes of sequencing data to identify rare mutations and in the advent of NGS in clinical diagnosis.

To demonstrate the added value of TSSV over mainstream STR profiling tools, we ran lobSTR (Gymrek *et al.*, 2012) and RepeatSeq (Highnam *et al.*, 2013) on four samples used for resolving allelic STR structures. Because RepeatSeq hardly reported any STR markers, the performance of TSSV could only be compared with that of lobSTR. We show that TSSV robustly and accurately resolved allelic STR structures with differing complexity. TSSV outperformed lobSTR in reporting the accurate copy number of major STR unit while it provides additional information on allelic STR structures and their strand-specific frequencies. Notably, TSSV excelled in resolving complex mixtures, whereas lobSTR failed to differentiate STR structures associated with different samples, and therefore produced unreliable and inaccurate estimations. Although lobSTR performs well on genotyping diploid samples, there is a clear need for tools to resolve mixtures with differing level of complexity and abundance.

Currently, the major limitation of TSSV is the sequencing read length because the detectable allelic structures are restricted to those that can entirely be covered by a single read. Thus, we envision that the immediate developmental outlook for TSSV can be the inference of allelic locus structure by local assembly of partial reads (reads with only one recognizable flanking region) combined with the comparative analysis of coverage of targeted loci and flanking regions. Furthermore, the promise of

novel sequencing technologies (such as *Pacific Biosciences RS II*), and therefore significant increase in read length will aid the study of larger structural variations.

Advances in sequencing technologies and computational analysis algorithms in unraveling genetic variations from SNPs and indels to CNVs (Chen *et al*., 2009; DePristo *et al*., 2011; Goya *et al*., 2010; Koboldt *et al*., 2009; McKenna *et al*., 2010; Ye *et al*., 2009) have facilitated the study of experimental data on an unprecedented scale to better understand the functional consequences of genetic variations. TSSV complements the existing tools by aiding the study of unknown, uncharacterized or highly polymorphic and repetitive short structural variations that can be used in a wide range of applications, from personal genomics to forensic analysis and clinical diagnostics.

ACKNOWLEDGEMENTS

The authors thank Dr. Annemieke Aartsma-Rus for her constructive feedbacks and discussions. S.Y.A. and K.J.vdG. performed the analyses. S.Y.A., K.J.vdG., M.H.A.M.V., J.S.V., P.dK. and J.F.J.L. designed the study. S.Y.A., K.J.vdG., J.W.F.vdH. and J.F.J.L. were involved in developing the tool. M.H.A.M.V., R.H.A.M.V., R.H.dL., C.B. and H.P.J.B. performed the wet-lab experiments and sequencing. J.T.dD., P.dK. and J.F.JL. coordinated the study. S.Y.A. drafted the manuscript that was subsequently revised by all co-authors.

Funding: This work was partially supported by the Centre for Molecular Systems Biology (CMSB), Duchenne Parent Project (the Netherlands) and a grant from the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands.

Conflict of Interest: none declared.

REFERENCES

Alkan,C. *et al*. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet*., **12**, 363–376.

Anvar,S.Y. (2013a) Allele-specific characterization of STR structures in pure and mixed forensic samples using TSSV, http://dx.doi.org/10.6084/m9.figshare.757791.

Anvar,S.Y. (2013b) Characterization of DeNovo structural variations induced by TALENs targeting hDMD in mouse ES cells using TSSV, http://dx.doi.org/10.6084/m9.figshare.757790.

Anvar,S.Y. (2013c) Characterizing IonTorrent PGM Error Profiles using TSSV, http://dx.doi.org/10.6084/m9.figshare.757792.

Boch,J. (2011) TALEs of genome targeting. *Nat. Biotechnol*., **29**, 135–136.

Brook,J.D. *et al*. (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell*, **69**, 385.

Cermak,T. *et al*. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res*., **39**, e82.

Chen,K. *et al*. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Conrad,D.F. *et al*. (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Costello,M. *et al*. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res*., **41**, e67.

de Cid,R. *et al*. (2009) Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet*., **41**, 211–215.

DePristo,M.A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*., **43**, 491–498.

Dere,R. *et al*. (2004) Hairpin structure-forming propensity of the (CCTG.CAGG) tetranucleotide repeats contributes to the genetic instability associated with myotonic dystrophy type 2. *J. Biol. Chem*., **279**, 41715–41726.

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

Girirajan,S. *et al*. (2011) Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet*., **7**, e1002334.

Goya,R. *et al*. (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.

Gymrek,M. *et al*. (2012) lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res*., **22**, 1154–1162.

Hauge,X.Y. and Litt,M. (1993) A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Hum. Mol. Genet*., **2**, 411–415.

Highnam,G. *et al*. (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res*., **41**, e32.

Hinds,D.A. *et al*. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet*., **38**, 82–85.

Hollox,E.J. *et al*. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet*., **40**, 23–25.

Iafrate,A.J. *et al*. (2004) Detection of large-scale variation in the human genome. *Nat. Genet*., **36**, 949–951.

Kayser,M. and de Knijff,P. (2011) Improving human forensics through advances in genetics, genomics and molecular biology. *Nat. Rev. Genet*., **12**, 179–192.

Kidd,J.M. *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.

Kim,Y. *et al*. (2013) TALENs and ZFNs are associated with different mutation signatures. *Nat. Methods*, **10**, 185.

Kimura,M. *et al*. (2009) Rapid variable-number tandem-repeat genotyping for Mycobacterium leprae clinical specimens. *J. Clin. Microbiol*., **47**, 1757–1766.

Koboldt,D.C. *et al*. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

Kremer,E.J. *et al*. (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science*, **252**, 1711–1714.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, **9**, 357–359.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform*., **11**, 473–483.

Mahadevan,M. *et al*. (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science*, **255**, 1253–1255.

McCarroll,S.A. *et al*. (2009) Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nat. Genet*., **41**, 1341–1344.

McKenna,A. *et al*. (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**, 1297–1303.

Medvedev,P. *et al*. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.

Miller,J.C. *et al*. (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol*., **29**, 143–148.

Mills,R.E. *et al*. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Mirkin,S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.

Moretti,T.R. *et al*. (2001) Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. *J. Forensic Sci*., **46**, 647–660.

Pearson,C.E. *et al*. (2002) Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. *Nucleic Acids Res*., **30**, 4534–4547.

Pearson,C.E. *et al*. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet*., **6**, 729–742.

Pinto,D. *et al*. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Sebat,J. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
Stephens,P.J. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.
Sudmant,P.H. *et al.* (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641–646.
Sutherland,G.R. and Richards,R.I. (1995) Simple tandem DNA repeats and human genetic disease. *Proc. Natl Acad. Sci. USA*, **92**, 3636–3641.
t Hoen,P.A. *et al.* (2008) Generation and characterization of transgenic mice with the full-length human DMD gene. *J. Biol. Chem.*, **283**, 5899–5907.
Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
Tuzun,E. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
Veltrop,M.H.A.M. *et al.* (2013) Generation of embryonic stem cells and mice for duchenne research. *PLoS Currents Muscular Dystrophy*, **1**.
Verkerk,A.J. *et al.* (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905–914.
Warshauer,D.H. *et al.* (2013) STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci. Int. Genet.*, **7**, 409–417.
Weber,J.L. and May,P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, **44**, 388–396.
Weischenfeldt,J. *et al.* (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, **14**, 125–138.
Wildeman,M. *et al.* (2008) Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum. Mutat.*, **29**, 6–13.
Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
Zhang,F. *et al.* (2011) Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.*, **29**, 149–153.