BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 14 2013, pages 1734–1741
doi:10.1093/bioinformatics/btt277

Sequence analysis

Advance Access publication May 15, 2013

ReviSTER: an automated pipeline to revise misaligned reads to simple tandem repeats

Hongseok Tae, Kevin W. McMahon, Robert E. Settlage, Jasmin H. Bavarva and Harold R. Garner*

Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Simple tandem repeats are highly variable genetic elements and widespread in genomes of many organisms. Next-generation sequencing technologies have enabled a robust comparison of large numbers of simple tandem repeat loci; however, analysis of their variation using traditional sequence analysis approaches still remains limiting and problematic due to variants occurring in repeat sequences confusing alignment programs into mapping sequence reads to incorrect loci when the sequence reads are significantly different from the reference sequence.

Results: We have developed a program, ReviSTER, which is an automated pipeline using a 'local mapping reference reconstruction method' to revise mismapped or partially misaligned reads at simple tandem repeat loci. RevisSTER estimates alleles of repeat loci using a local alignment method and creates temporary local mapping reference sequences, and finally remaps reads to the local mapping references. Using this approach, ReviSTER was able to successfully revise reads misaligned to repeat loci from both simulated data and real data.

Availability: ReviSTER is open-source software available at http://revister.sourceforge.net.

Contact: garner@vbi.vt.edu

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

Received on October 22, 2012; revised on April 25, 2013; accepted on May 10, 2013

1 INTRODUCTION

With the advent of next-generation sequencing technologies, sequencing approaches became the most common strategy for variation discovery in a population (Mackay *et al.*, 2012). Next-generation sequencing technologies have enabled investigators to generate a huge amount of sequence data and to compare variants between different populations in a cost- and time-efficient way. Unfortunately, with their inherent sequencing errors and short sequence read lengths, data analysis for several kinds of repeat elements such as transposon elements and tandem repeats still remains limiting and problematic.

The quality of variation discovery relies on the mapping results of sequence reads to a reference. Incorrect mapping or non-mapping of sequence reads to the reference is mainly caused by inconsistency between sequences of the test subject and the reference, especially in repeat regions, and often results in invalid variant calling. Simple tandem repeat (STR) sequences in a test subject frequently show significant differences from a reference sequence (McIver *et al.*, 2011). This is likely due to mutation rates that have been estimated to be 10^{-2}–10^{-6} mutation events per generation (HanCock, 1999), which is much higher than mutation rate ($1.2 \times 10^{-8}\,\mathrm{bp}^{-1}\mathrm{generation}^{-1}$) for a SNP (single-nucleotide polymorphism) in a unique sequence (Scally and Durbin, 2012). In our study of Brucella genomes (Tae *et al.*, 2012b), we observed that mapping programs often assign high-quality scores to incorrectly mapped reads when two or more tandem repeat loci contain the same motif with different repeat lengths and their flanking sequences show high similarity. This is because mapping program parameters are normally set to minimize the number of mismatch or insertion/deletions (INDEL) bases in an alignment (Supplementary Figs S1 and S2). This mismapping leads directly to invalid variant calls in repeat loci because the variation calling programs rely only on the mapping quality scores to filter out false-positive variants from incorrectly mapped reads. In the human genome, we found that more than two-third of STRs are overlapping or near (within 50 nt) transposon elements (Supplementary Fig. S3). Notably, AT-rich STRs are often discovered near the 3′ ends of retrotransposons (Batzer and Deininger, 2002), which frequently results in the left or right flanking sequence of an STR being highly replicated, while the other flanking sequence is unique (Supplementary Fig. S4). The sequence reads mapped to the incorrect STR loci due to length variation of the STRs can be remapped if flanking sequences on one side of the STRs are unique and the correct lengths of the STRs in the sequenced sample are known.

Sequence reads are also often partially misaligned to a reference sequence if the reads contain INDEL variants and do not span enough of both flanking sequences of the locus. A few programs such as SMRA (Homer and Nelson, 2010) and GATK (McKenna *et al.*, 2010) realign sequence reads mapped to the INDEL variant loci to correct misalignment, but their performance is poor for the reads mapped to STR loci containing long INDELs. To correctly realign sequence reads at the INDEL variant loci, the programs require correctly mapped reads supporting the variants, but the reads containing tandem repeat variation often fail to be mapped to the correct loci and as a result the programs do not obtain sufficient read depth to support the correct variant calls.

Several programs, including Pindel (Ye *et al.*, 2009), Dindel (Albers *et al.*, 2011) and SOAPindel (Li *et al.*, 2013), have been developed to find insertions and deletions using remapping of

*To whom correspondence should be addressed.

unmapped reads or local alignment approaches, but they do not handle reads mapped to incorrect loci. Although lobSTR (Gymrek *et al*., 2012) has been developed to profile short tandem repeat loci using its own alignment approach, it is limited to only analyzing 2–6-mer motif repeat loci. To overcome the difficulty of variant calling at STR loci in a bacterial genome, we applied an iterative backbone remapping and assembly method to generate the genome sequence of bacterial field isolates and to call their correct variants (Tae *et al*., 2012b). The method has successfully detected INDEL variants including tandem repeat variants shorter than the read length, but is not applicable to eukaryotic genomes because many variants at diploid genomes are heterozygous. Analysis of heterozygous STR variants is even more challenging compared with SNPs or short INDELs because two different non-reference alleles are frequently discovered from an STR locus and the alleles at the STR locus can be different between individuals (Edwards *et al*., 1991).

Here, we describe ReviSTER (**Revi**se **S**imple **T**andem repeat **E**rror **R**eads), which is an automated pipeline using a 'local mapping reference reconstruction method' to revise mismapped (mapped to incorrect position) or partially misaligned (mapped to correct position but one of ends misaligned) reads at STR loci. It takes FASTQ-formatted files, a reference sequence file and a list file containing STR locations as inputs and uses BWA as an initial mapping program. It subsequently realigns reads unmapped by BWA (Li and Durbin, 2009) using BLAT (Kent, 2002), and conducts local assembly with all aligned reads to an STR locus. From the local assembled result, new local mapping reference sequences are generated, and all mapped reads containing more than one mismatch in their alignments and unmapped reads are mapped again to the new local mapping references. Reads mapped to the local mapping references are relocated to the original reference and compared with the original alignments to choose the best alignment. The performance of ReviSTER was compared with two mapping programs, including BWA and Bowtie2 (Langmead and Salzberg, 2012), and a realignment program, GATK (McKenna *et al*., 2010), using three different test datasets.

2 METHODS

2.1 Six steps to revise mismapping and misalignment

ReviSTER is a PERL program and uses three programs including BWA, BLAT and SAMTools (Li *et al*., 2009) installed in the user's environment. RevisTER takes FASTQ-formatted files, a FASTA-formatted reference sequence file and an optional file containing a list of STR locations as inputs (Supplementary Note 'ReviSTER manual'). If the STR location file is not submitted, it searches for STR loci from the reference file (Supplementary Methods). ReviSTER uses a novel approach called a 'local mapping reference reconstruction method' to revise mismapping and misalignment of sequence reads derived from the STR loci (Fig. 1). The whole process to revise the read alignments is composed of six steps. The first step is to map reads to a reference sequence using BWA. ReviSTER takes the '-n' option, which is used for BWA mapping to record multiple mapping candidates for reads derived from repeat sequences. Next, BLAT is used to remap unmapped reads to temporary reference sequences, which are extracted from the original reference sequence only around a given STR loci. Because BLAT generates many false alignments for a read, ReviSTER realigns them and chooses the best alignment from several alignment candidates (See 'Remapping unmapped reads using BLAT'). Third, ReviSTER employs a local assembly step using the reads mapped to each STR locus. It generates paths in a graph of reads overlapping at least 30 bases with each other, chooses a given number of paths corresponding to allele candidates, extracts sequences of the allele candidates and creates local mapping reference sequences containing the allele candidates (see 'Local mapping reference reconstruction using a local assembly'). In this step, sequence reads containing more than one mismatch/INDEL bases or showing abnormally long pair distances are saved in a separated file along with unmapped reads. Fourth, the reads saved in the separate file are mapped to the local mapping reference sequences by BWA (with the -n option). Fifth, mapping positions of a read on the local mapping reference sequences are converted to positions on the original reference. Then a mapping position with the most optimal pair distance and the lowest mismatch number is chosen among all mapping candidates identified in the first step and the fifth step (see 'Selecting an optimized mapping position'). The final step is to revise reads partially misaligned to STR loci, which is an independent process from the previous steps. Some reads may have been incorrectly aligned to the STR loci containing long INDELs and not be revised by the previous steps. The reads are realigned to another read that has been mapped to the same STR locus and sufficiently span the flanking sequences of the locus (see 'Realignment of partially misaligned reads').

2.2 Remapping unmapped reads using BLAT

When the allele length of an STR locus is different from a reference, sequence reads containing the allele often fail to be mapped because mapping programs give large negative alignment penalties for long INDELs. To remap the unmapped reads to STR loci, ReviSTER creates temporary reference sequences, which are extracted from the original reference around STR loci. Each temporary reference sequence includes 'max. read length-10' base left flanking sequence, STR sequence and 'max. read length-10' base right flanking sequence. BLAT [with option '-maxIntron = max (50, max. read length-15) '] is used to map the reads unmapped by BWA. BLAT searches multiple mapping position candidates for a single query and provides multiple alignment blocks for each mapping candidate. Because many STR loci share the same motifs and their flanking sequences often exhibit high similarities, many of the BLAT alignments are false positives. Additionally, if a read does not include sufficient flanking sequence length at both the ends of an STR locus, BLAT may produce a partial misalignment. ReviSTER realigns each alignment by assigning a low mismatch penalty to INDELs in the STR locus to correct the partial misalignment (Supplementary Fig. S5). It assigns 1 as the mismatch penalty score for any individual mismatch or INDEL bases in flanking sequences but does not assign a penalty of more than 2 to any INDEL event in the STR locus regardless its length. ReviSTER then chooses a mapping position candidate producing the lowest mismatch score for the read.

Because ReviSTER uses the sequences only around STR loci as mapping reference sequences, there is still a high probability that the positions of the reads mapped by BLAT are incorrect. This problem is reduced by the local mapping reference reconstruction and remapping steps as follows.

2.3 Local mapping reference reconstruction using a local assembly

Most mapping programs assign higher penalty to a gap opening than a mismatch. This frequently leads to the partial misalignment of a read mapped to the STR locus. As the length of an INDEL at the locus gets longer, there is less chance for the read containing the INDEL to be correctly aligned. Also, when two or more STR loci contain the same motif repeat sequences and their variants are significantly different from the reference, mapping programs tend to map a read to a position producing the smallest number of mismatch/INDEL bases instead of mapping

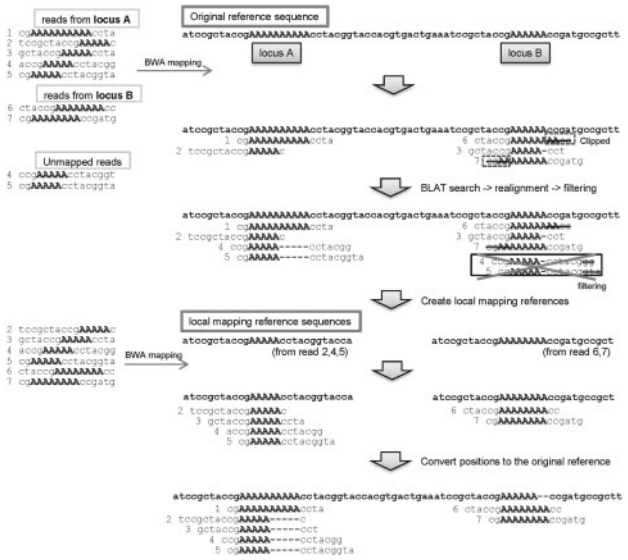

Fig. 1. Local mapping reference reconstruction. A local mapping reference reconstruction method is used to revise mismapped or partially misaligned reads at STR loci. It initially maps reads to a reference using BWA and uses BLAT to map reads unmapped by BWA. From the mapping results, it conducts local assembly using all aligned reads to an STR locus to create local mapping reference sequences. Then all mapped reads containing more than one mismatch in their alignments and unmapped reads are mapped again to the local mapping references. Reads mapped to the local mapping references are relocated to the original reference and compared with the original alignments to choose the best alignments

it to a position requiring an alignment containing a long INDEL. This problem is frequently observed at the STR loci of which flanking sequences show high similarities to each other. If flanking sequences at both sides of an STR locus show high similarities to that of other loci, it is difficult to correct a mismapped read derived from one of the STR loci showing the similarities. But if at least one side of the flanking sequences is unique, the mapping position of reads mismapped to another similar locus can be corrected.

The local mapping reference reconstruction method was developed to correct mismapping and partial misalignment of reads derived from the STR loci by creating temporary mapping reference sequences containing the alleles of each locus in the sequenced genome. To find the alleles at a locus, ReviSTER uses a local assembly approach (Supplementary Fig. S6). It first collects all reads mapped to an STR locus and creates a graph from the reads overlapping at least 30 bases with one another without any mismatch. Then it selects a path passing through the highest number of reads among all paths including at least three reads, spanning at least 15 bases of both flanking sequences of the STR and containing at least one read including at least five bases of both flanking sequences (Supplementary Note 'Required sequence read coverage for different lengths of simple tandem repeats'). From the original reference, a partial sequence from 'max. read length-10' base upstream to 'max. read length-10' base downstream of the STR locus is extracted and the STR sequence in the partial sequence is replaced by the allele sequence in the selected path. If the length of the allele in the selected path is at least three bases different from the reference, the partial sequence is saved as a local mapping reference sequence for the next step. When the genome of the sequenced sample is diploid, another path containing the second allele from the graph is selected. If the length of the second allele is at least three bases different from the reference and the first selected allele, another local mapping reference sequence is created. The number of local mapping reference sequences (i.e. the number of possible alleles) for a single locus is decided by the '-p [ploidy of the target genome]' parameter.

All sequence reads containing more than one mismatch/INDEL base or showing abnormally long pair distances (default greater than mode distance × 2) along with originally unmapped reads are remapped to the local mapping references by BWA (with the -n option, default 10). Because the positions of local mapping reference sequences on the original reference are known, the mapping positions of reads on the local mapping reference sequences can be easily converted to positions on the original reference. Then the mapping positions of remapped reads are compared with their original mapping positions.

2.4 Selecting an optimized mapping position

Mapping programs first consider the alignment scores (decided by number of mismatches/INDELs) and the pair distance at each position to select mapping positions for the reads in a pair. If they do not find an appropriate alignment pair within a given distance, they map the two reads individually to the positions producing the highest alignment scores. Because STRs are highly variable, they often show significant differences from the reference. If an STR locus has a different length from the reference, mapping programs may not consider the locus as a mapping position candidate of a read that is derived from the locus because an alignment score of the read at that position may be low. Instead, they may map the read to another STR locus that also has high similarity to the correct locus. To correct the mismapped reads, ReviSTER reconstructs a local mapping reference sequence containing the allele length in the sequenced sample for the possible correct locus and the read can then be mapped to the local mapping reference producing a high alignment score.

To select optimized mapping positions for reads in a pair, ReviSTER collects all mapping position candidates from 'XA:' tags (added by BWA to show alternative mapping candidates) in the original mapping data and the local mapping data, and compares all possible alignment pairs in a certain distance (default less than or equal to mode distance of the original mapping data × 2) after it converts the mapping positions on a local mapping reference sequence to positions on the original reference. It then chooses an alignment pair with the smallest number of mismatch/INDEL bases for both alignments. When a sequence read is generated by a single-end sequencing method, only the number of mismatch/INDEL bases is considered to select the mapping position [The information in the 'XA:' tags are not used to revise single-end mapping (SE)]. When an alignment on a local mapping reference is selected for a read, a new alignment is estimated from the alignment information between the local mapping reference sequence and the original reference.

2.5 Realignment of partially misaligned reads

The local mapping reference reconstruction method requires high-quality sequence reads because it connects reads overlapping at least 30 bases with one another without any mismatch to create a graph. If an STR locus does not satisfy these conditions, a graph for the locus is not created and no read alignment is revised. To revise partially misaligned reads at the STR locus, an additional step is required. ReviSTER first collects all mapped reads overlapping with the STR sequence and separates them into three different groups. The reads in the first group cover the whole STR sequence and span at least seven bases of both flanking sequences of the STR. The reads in the second group contain at least one mismatch in their alignments and do not span or span less than six bases of the left flanking sequence. The third group has the same conditions as the second group except that they do not span or span less than six bases of the right flanking sequence. The other reads are excluded from the local realignment. From the reads in the first group, template sequences, which

include read sequences corresponding to the reference from seven bases upstream to seven bases downstream of the STR and are not duplicated, are created. And each read in the second group is aligned to the template sequences from seven bases upstream of the 3′ end of the STR. Then, a template sequence producing the smallest number of mismatch bases in their alignment is selected as an alignment template for the read and the old alignment for the read is replaced by the new alignment estimated from the alignment template. The reads in the third group are also aligned to the template sequences corresponding to the reference from the 5′ end to seven bases downstream of the STR and their alignments are revised.

This final local alignment step can be replaced by GATK using '-GATK' option.

2.6 Test sets

To evaluate the performance of ReviSTER, we created a simulated dataset and also used sequence datasets from *Brucella suis* 1330 and human genomes. The first dataset was simulated data created as sequence reads of a single individual diploid genome. To evaluate the performance of ReviSTER for sequence reads mapped to STR loci near replicated regions such as transposon elements, we created five subsets including different references with $N = \{1, 3, 10, 15, 20\}$ by placing the same 400 base sequences on the left sides of N STR loci consisting of the same motifs (Supplementary Fig. S7A). For each subset, the reference sequence was created from human chromosome 1 (build 37) after removing all repeat sequences identified by RepeatMasker (http://repeatmasker.org). Then 31 types of 1–8-mer motifs were used to generate $31 \times 3 \times N$ STR loci with 15–84 bases (3–55 repeats of a motif in the sequence) in length, each of which was inserted every 1400 bases (400 bases as a replicated sequence on the left side and 1000 bases as a unique sequence on the right side of the STR sequence). To create sequence reads, two allele sequences for each STR locus, of which lengths were randomly selected within 12–120 bases, were created and inserted into 210 base DNA fragment sequences completely covering the STR sequence (Supplementary Fig. S7B). The 210 base DNA fragment sequences containing each allele were created every two bases on the reference sequence. While the DNA fragment at the left-most side spanned the left two bases of the right flanking sequence of the STR locus, the fragment at the right-most side spanned the right two bases of the left flanking sequence. Then 100 base sequences at both the ends of the fragment sequences were used as paired-end sequencing reads (Supplementary Fig. S7C). About 3% of alleles were longer than 90 bases, lengths of which are difficult to analyze with 100 base sequence reads.

To test the performance with haploid genomes, the second dataset was created from the sequencing data (SRA: SRA056338) (sequenced by Illumina 76 cycle paired-end protocols) of *B.suis* VBI22 (Tae *et al.*, 2012a), which is a gram negative bacteria having two chromosomes. Its genome has 11 loci containing 8-mer tandem repeats (more than or equal to three motif copies), which are highly variable. The left flanking sequences of five loci among them contain highly similar 90 base sequences (Supplementary Fig. S2). Because the average sequence coverage of the original raw dataset was approximately 1000×, we reduced the number of reads by randomly selecting 8 213 000 sequencing read pairs (16 426 000 reads). The genome sequence of *B.suis* 1330 (Tae *et al.*, 2011) was used as a reference sequence and an STR list file was created with eleven 8-mer STR loci from the reference. Because the completed genome sequences of both strains have been published, the allele lengths of each STR locus in both genomes are known. The other STR loci containing 1–7-mer motifs were also compared but there were no loci showing differences between the two genomes except three G-homopolymer loci. But the G-homopolymer loci were not included in the list file because the sequence coverage for the loci was low.

The third dataset was created from the four exon-targeted sequence sets of human genomes, HG00641 (SRA: SRR107085), HG01105 (SRA: ERR034575), NA06994 (SRA: SRR070528, SRR070819) and NA19153 (SRA: SRR070660, SRR070846), downloaded from the 1000 genome project Web site (http://www.1000genomes.org). Exon-targeted sequence data were selected to obtain sufficient read coverage at STR loci, and the human genome NCBI build 37 was used as a reference sequence for mapping. To create a list of STR loci, TRF v4.04 (Benson, 1999) (with '2 7 5 80 10 14 6' options) was used to search for repeat sequences including incomplete repeat sets. When two STR loci were within 30 bases, they were merged into one locus; and loci containing at least 10 bases for 1-mer, 12 bases for 2-mer and 15 bases for 3,4,5,6-mer of pure repeat sequences (resulting 1 418 122 loci); and shorter than 90 bases were selected (1 373 574 loci). Then the 30 base flanking sequences on both sides of each locus were extracted and mapped as single-end reads by BWA to the human reference sequence. Among the selected STR loci, the loci of which at least one flanking sequence was assigned the highest BWA mapping score (Phred score 37) were selected, which resulted 1 362 903 loci. Among them, 8652 loci overlapping with exon regions (Supplementary Table S6), which had been used for exon targeted sequencing in the 1000 genome project, were obtained.

The fourth dataset was composed of two sequence datasets from blood and saliva samples from a single human individual (SRA: SRR345592, SRA: SRR345593) (Illumina 101 cycle paired-end sequencing), which contained 1 499 021 500 and 1 692 395 618 reads, respectively. We used 1 362 903 STR loci for this test.

For the final test to compare the genotyping performance before and after ReviSTER revised misaligned reads, NCBI release_5_30 of *Drosophila melanogaster* was used as a reference genome sequence and simulated sequence reads for the *Drosophila* reference sequence were generated by pIRS (Hu *et al.*, 2012) (Supplementary Results 'Genotyping test with simulated data generated by pIRS from the *Drosophila* reference' for detail).

3 RESULTS

BWA (v0.6.1), Bowtie2 (v2.0.0) and GATK (v1.6–9) were used for the performance comparison. For BWA and Bowtie2, default options were used to map sequence reads to reference sequences. For GATK, which realigned the results of two mapping programs, options '-T IndelRealigner -targetIntervals' with interval files that contained positions of target STR loci were used. ReviSTER used BWA (v0.6.1) and BLAT (v34) as initial mapping programs.

3.1 Performance test with simulated data

Because STR loci are frequently discovered at heads or tails of long repeat sequences such as transposon elements, we created five subsets containing different reference sequences with $N = \{1, 3, 10, 15, 20\}$ by placing the same 400 bases on the left sides of N STR loci to test ReviSTER's performance for the STR loci near replicated regions (Supplementary Fig. S7). Paired-end sequence reads of a single individual diploid genome were created for each reference (See 'Test sets') and mapped by ReviSTER (with option '-n 10 -p 2'), BWA (two different mapping with default options and '-e 90 -E 1' options) and Bowtie2. The BWA options '-e 90 -E 1' allow alignments with long gap extension. Then mapped reads using BWA and Bowtie2 were realigned by GATK. To test the performance of the programs with single-end sequence reads, reads were also mapped as single-end reads by the programs and their mapping positions were compared. Because the exact alignment position of each read was known, the aligned reads were categorized into four groups including 'unmapped', 'mismapped', 'partially misaligned' and 'correctly aligned'. When both end positions of a read alignment were

not correct, it was counted as 'mismapped', and when only one of the end positions of a read alignment was not correct, it was counted as 'partially misaligned'.

ReviSTER achieved the highest overall accuracy in all subsets. Especially when N (the replicated number of left flanking sequences of STR loci) was equal to or smaller than 10 (input value of ReviSTER set by the -n option). It showed >90% correct alignments in paired-end mapping (PE) results, while BWA/GATK, BWA with '-e80 E1' and Botiwe2/GATK showed only 53.7, 76.6 and 66.4% correct alignments for the subset with $N = 10$, respectively (Fig. 2A and B). When N was greater than 10, the correct alignment percentage of ReviSTER dropped to ~80% but it was still higher than that of other programs for the same subsets. The accuracy could be improved by increasing the '-n' option values (Supplementary Fig. S8). In the SE test, ReviSTER had the highest correct alignment rate (92.5%) for the subset with the unique left flanking sequences ($N = 1$), while BWA/GATK, BWA/e90E1 and Botiwe2/GATK showed 65.1, 78.9 and 71.2% correct alignments, respectively, for the same data. For high N (>10), the correct alignment percentage of ReviSTER dropped below 70%. ReviSTER, BWA/GATK, BWA/e90E1 and Botiwe2/GATK showed 65.6, 36.8, 48.7 and 43.7% correct alignments, respectively, when N was 20 because many reads could be aligned to multiple locations without any mismatches, which then could not be corrected without PE information.

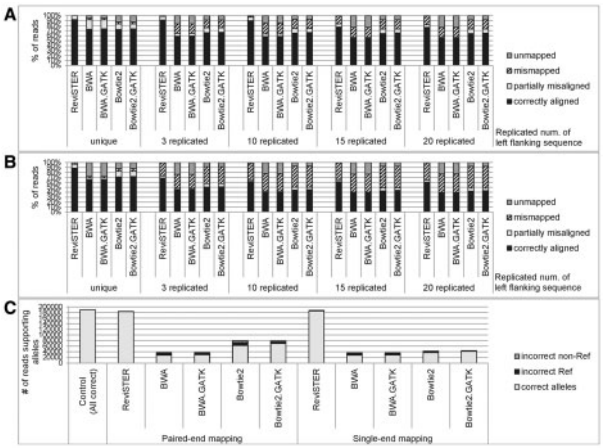

Fig. 2. Performance test with simulated data. (**A**) For each subset, the same 400 base sequence was placed on $N = \{1, 3, 10, 15, 20\}$ STR loci as their left flanking sequences and $31 \times 3 \times$ N STR loci were created. Sequence reads with 100 bases in length were mapped by PE methods. ReviSTER obtained >90% correct alignments for the subsets with $N \leq 10$, while the other programs had <70% of correct alignments for the subset with $N = 10$. (**B**) The same data as used in (**A**) was then processed using SE methods. Because several loci have common left flanking sequences that are 400 bases in length, many reads could be mapped to multiple places when N > 1. ReviSTER showed the highest accuracy in the comparison. (**C**) For the subset with $N = 20$, the number of reads supporting correct alleles, incorrect reference alleles and incorrect non-reference alleles were counted from the PE and SE results regardless their original mapped positions. ReviSTER successfully aligned reads 96.3 and 97.6% of the time that supported correct alleles in the PE and SE tests, respectively, while the other programs properly aligned reads <55% of the time, thus supporting correct alleles

Many reads could be mapped to multiple locations, when N STR loci had the same motifs and left flanking sequences. To measure the performance for alignments to the STR loci instead of the mapping positions, we ignored the original positions of reads and counted the numbers of read alignments supporting correct or incorrect allele lengths for the subset with $N = 20$. To filter out read alignments not completely covering STR loci, only read alignments spanning at least two bases for both flanking sequences were used for the comparison. We compared the INDEL lengths in the alignments to the reference (ignoring mismatches) and counted the numbers of read alignments supporting 'correct', 'incorrect reference' and 'incorrect non-reference' allele lengths of STR loci. ReviSTER generated 96.3 and 97.6% of read alignments supporting correct alleles of the STR loci from PE and SE, respectively, while BWA/GATK, BWA/e90E1 and Botiwe2/GATK generated 15.5, 52.1 and 36.6% from PE and 15.1, 51.7 and 21.8% from SE, of which read alignments supported correct alleles, respectively (Fig. 2C). Most incorrect alignments in BWA/GATK and Botiwe2/GATK results supported incorrect reference alleles rather than non-reference alleles because standard mapping programs prefer to minimize the number of mismatch or INDEL bases in an alignment.

BWA mapping with a lower gap extension penalty showed higher accuracy than BWA mapping with a default penalty at STR loci, but the lower gap extension penalty may cause inaccurate read mapping to non-STR loci or slow performance if a target sequence is a whole genome (Supplementary Table S1). We also tested BWA mapping with a lower gap penalty to remap initially unmapped reads instead of BLAT in the ReviSTER pipeline, but observed lower accuracies with it (Supplementary Fig. S9).

3.2 Haploid genome sequence data

To test the performance with haploid genomes, we used the sequencing data (SRA: SRA056338) of *B.suis* VBI22 (Tae *et al.*, 2012a) sequenced by the standard Illumina 76 cycle paired-end protocols and a genome sequence of *B.suis* 1330 (Tae *et al.*, 2011) as a reference. Because completed genome sequences of both strains have been published, the allele lengths of each STR locus in both genomes are known. For the comparison, 11 STR 8-mer loci were targeted. Because most loci contain incomplete repeat units, the lengths of the loci are not multiples of eight. Sequence reads were mapped by ReviSTER, BWA and Bowtie2, and the alignments generated by BWA and Bowtie2 around targeted STR loci were realigned by GATK. To measure the performance of the programs, we compared the INDEL lengths in the alignments to the reference (ignoring mismatches) and counted the numbers of read alignments supporting 'correct', 'incorrect reference' and 'incorrect non-reference' allele lengths of target STR loci (Table 1) (we ignored mismatches in alignments). Only read alignments spanning at least two bases of both flanking sequences of each locus were counted. ReviSTER aligned 1325 reads correctly to the target loci, while BWA/GATK and Bowtie2/GATK aligned 687 and 700 reads correctly. BWA had low rates of read alignments completely covering STR loci. Many alignments generated by BWA did not span at least two bases of both flanking sequences and were not included in the read counts because BWA clipped ends of alignments when it

Table 1. The number of read alignments supporting correct, incorrect reference and incorrect non-reference alleles of eleven 8-mer STR loci

Chromosome	Position	Length in 1330	Difference in VBI22	Number of reads supporting allele lengths								
				ReviSTER			BWA/GATK			Bowtie2/GATK		
				TP	FP Ref	FP NonRef	TP	FP Ref	FP NonRef	TP	FP Ref	FP NonRef
1	62810	49	0	113	0	0	114	0	0	114	0	6
1	64726	91	−40	78	0	0	0	0	0	0	0	0
1	87040	30	0	182	0	2	182	0	2	182	0	2
1	438870	44	−16	254	0	0	135	0	0	0	0	3
1	736270	45	−8	160	0	0	115	0	0	151	0	0
1	736377	31	+24	73	0	5	0	0	1	0	0	91
1	1399533	32	+16	96	0	1	0	0	1	0	15	74
1	73008	23	+8	141	0	0	86	0	0	141	0	0
2	73070	89	−16	0	0	0	0	0	0	0	0	0
2	749707	35	+8	113	1	0	55	1	0	112	1	0
2	976068	35	+16	115	0	2	0	0	2	0	26	38
Total				1325	1	10	687	1	6	700	42	214

Note: Sequence reads of *B.suis* VBI22 were aligned to the genome sequence of *B.suis* 1330 by BWA, Bowtie2 and ReviSTER. The read alignments spanning at least two bases of both flanking sequences of the STR loci were counted.
TP: True positive, read alignments supporting correct allele lengths.
FP: False positive, read alignments supporting incorrect allele lengths.
Ref: Reference allele lengths.
NonRef: Non-reference allele lengths.

detected abundant mismatches at the regions. Bowtie2 aligned reads incorrectly most frequently among three programs because it gave high strength to aligned bases instead of clipping the mismatch abundant alignment ends. ReviSTER did not detect the correct allele from only one locus because that allele length in *B.suis* VBI22 was 73 bases, which was close to the raw read length (76 bases).

3.3 Human exome sequencing data

Exon-targeted sequence reads of four human genomes (HG00641, HG01105, NA06994 and NA19153) were downloaded from the 1000 genome project site (http://www.1000genomes.org). To compare the alignment results from before and after revision by ReviSTER, we mapped reads using BWA with an option -n 10 (default for the other options) using two different mapping methods including PE and SE. Then ReviSTER used each BWA alignment file (SAM) as an input to revise reads mapped to 8652 STR loci.

Approximately 94% of the targeted STR loci were completely covered by at least two reads, which spanned at least five bases of both flanking sequences of the loci (We analyzed only reads spanning at least five bases of both flanking sequences to reduce false positives in this test). Because allele candidates for most loci showed few differences from reference alleles (Supplementary Fig. S10), ReviSTER did not revise most read alignments for the alleles. On average 24 516 reads at 1507 loci from PE results and 15 041 reads at 1415 loci from SE results were revised, respectively (Table 2; for additional exome data, Supplementary Tables S2 and S3). ReviSTER revised more read alignments when PE was used. The numbers of affected loci and revised read alignments were proportional to the length of sequence reads.

Because it was difficult, given that there is no 'gold standard', to distinguish true-positive alleles from false-positive alleles for STR loci from human genome sequence data, we assumed that the results of PE approaches would contain more potentially true positives than those from SE. We compared the numbers of reads, which were aligned by ReviSTER SE and BWA SE and concordant with non-reference allele candidates detected by ReviSTER PE and BWA PE for NA19153 to validate improvement of alignments after ReviSTER's revision of data from SE approaches (Fig. 3). While 78 672 reads aligned by ReviSTER SE supported non-reference allele candidates detected by both of ReviSTER PE and BWA PE, only 71 339 reads aligned by BWA SE supported the same allele candidates. The numbers of allele candidates commonly detected by different mapping approaches are shown in the Supplementary Figure S11.

3.4 Two sequencing datasets from a single human individual

We used two sequencing datasets from blood and saliva samples from a single human individual. The reads were aligned to the human genome reference NCBI build 37 by BWA and realigned by GATK. Among 1 362 903 STR loci (see 'Test set' in Methods), 1 349 170 and 1 347 548 loci were completely covered by 38 150 693 reads from the blood sample and 39 927 488 reads from the saliva samples, respectively. Most of the reads (28 836 900 reads from blood and 30 238 042 from saliva) supported reference alleles. To test the performance of ReviSTER, we compared the consistencies of GATK genotyping results in

Table 2. Summary of ReviSTER revision for four exon targeted sequence reads of four human genomes

Sample ID	Read length	Total read number	Mapping method	Insertion size, mode	Number of reads on STRs		Number of loci completely covered by ≥2 reads		Affected loci	Revised read alignments
					BWA	ReviSTER	BWA	ReviSTER		
HG00641	76 bp	307 906 150	Paired-end	152 bp	1 903 773	1 905 839	8436	8436	1889	37 516
			Single-end		1 625 502	1 639 230	8324	8324	1866	26 604
HG01105	90 bp	110 373 536	Paired-end	180 bp	1 038 773	1 039 284	8309	8309	1602	23 213
			Single-end		926 617	936 699	8226	8226	1461	11 204
NA06994	100 bp	99 362 536	Paired-end	200 bp	868 411	868 835	8192	8192	1213	17 207
			Single-end		799 467	806 409	8133	8133	980	7650
NA19153	100 bp	115 318 754	Paired-end	200 bp	891 598	892 191	8245	8245	1325	20 127
			Single-end		812 140	821 151	8168	8169	1354	14 704

Note: The exon-targeted sequence reads of four human genomes were mapped by BWA to the NCBI build 37 human genome reference sequence and their alignments were revised by ReviSTER. The same sequence reads were mapped by two different methods including PE and SE for comparison.

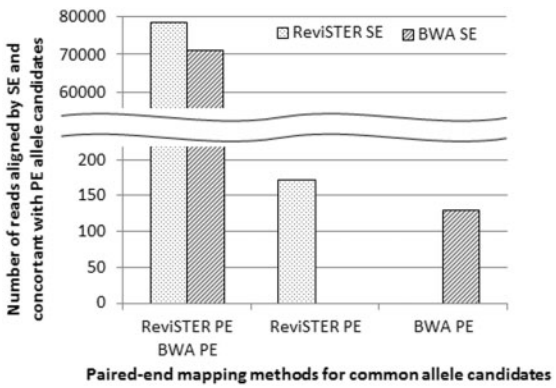

Fig. 3. Numbers of reads, aligned by two SE methods, supporting non-reference allele candidates commonly detected by two different PE methods. Because allele candidates of PE methods have higher reliability than that of SE, the numbers of reads, which were aligned by ReviSTER SE and BWA SE and concordant with non-reference allele candidates detected by ReviSTER PE and BWA PE for human exome sequence (NA19153), were compared with validate improvement of alignments after ReviSTER's revision of data from SE approaches

two different samples before and after ReviSTER revision. While GATK genotyped non-reference alleles from 277 540 loci consistently in the two matched samples before revision, it could genotype non-reference alleles from 303 995 loci consistently after revision (Supplementary Fig. S12).

3.5 Genotyping performance before and after ReviSTER revision

To compare genotyping performance before and after ReviSTER revising misaligned reads, GATK was used to genotype INDELs in STR loci with simulated sequence reads generated by pIRS (Hu *et al.*, 2012) from the *Drosophila* reference sequence (Supplementary Results 'Genotyping test with simulated data generated by pIRS from the *Drosophila* reference' for detail). In this test, we obtained improved genotyping results after ReviSTER revision (before 75.5%, after 84.1% of correct calls). The profiling results from lobSTR for target STR loci were also used for the comparison. lobSTR showed poor results (2.8% of correct calls) using default conditions. lobSTR's high false-negative rate is consistent with other published results (Highnam *et al.*, 2012). Because all loci in our test set contain at least one non-reference allele, we found that lobSTR performed better at calling reference alleles than non-reference alleles.

4 DISCUSSION

Here we presented ReviSTER as an automated pipeline to revise mismapped or partially misaligned sequence reads to STR loci, which frequently have high variation rates resulting in misalignment of reads. The STR loci are often found near the beginning or end of long repeat sequences such as transposon elements, which results incorrect mapping of reads because of high similarity found in the flanking sequence. ReviSTER creates local mapping references for each STR locus from a graph generated from mapped reads and then remaps the sequence reads to the local mapping references. Using this approach, we were able to successfully correct reads misaligned to the STR loci from BWA alignments for both simulated data and real data. Overall, we observed that ReviSTER could align 2 × more reads to support correct alleles of STR loci for haploid genomes than the other programs.

In application to the human genome data, we used indirect comparison methods to evaluate ReviSTER's performance because it is difficult to identify true alleles of STR loci in a given human genome. Many STR in the human genome are parts of transposon elements, which may be distributed in more than several tens of thousands of copies (Supplementary Figs S3 and S4). In such cases, mapping programs perform poorly to initially map sequence reads to correct loci. ReviSTER, like other mapping software, will find it difficult to correctly identify

allele candidates and realign reads for such loci. This issue, common to for other mapping and genotyping programs, may be reduced as advanced sequencing technologies begin producing longer reads than current technologies.

With the traditional resequencing approach, many researchers consider mismapping of reads a difficult problem yet to be resolved, while a few methods have been developed to revise partially misaligned reads. ReviSTER is the first program to revise not only partially misaligned reads but also mismapped reads to STR loci. Like other programs, its performance can be improved in the future. First, its performance is greatly affected by the initial mapping results from BWA and BLAT. In our comparison, we observed that Bowtie2 showed better performance than BWA. ReviSTER currently uses BWA for the initial mapping (because of the 'XA' tags. The information in the tags are used only to revise PE), but an additional module to use other mapping programs with better performance will be implemented in near future. Second, in the local mapping reference reconstruction step, the current version of ReviSTER searches for only exact match of two reads to connect them in a graph to reduce the possibility of creating a graph supporting false-positive alleles. This may result in not creating any graph with reads containing sequencing errors. Time efficiency is also a limitation of the current version (Supplementary Table S5). It has been implemented in PERL, so we would expect significant speedup in the next version to be implemented in C or C++.

The significant advantage of ReviSTER is fault tolerant automation, which is one of the most important functions for complex pipelines. ReviSTER has been designed to be fully automated and to easily recover an analysis after a process interruption by a system failure such as a power failure. When the analysis process is restarted, ReviSTER begins the process from the unfinished step if users type the same command as the original run.

ACKNOWLEDGEMENTS

This work was supported by the Medical Informatics and Systems Division director's funds, and the 1000 Genomes Project Dataset Analysis Grant from the National Human Genome Research Institute of the National Institute of Health [grant number 1U01HG-005719-01]. DAC (Data Analysis Core) at the Virginia Bioinformatics Institute helped sequence data analysis.

Conflict of Interest: H.G. is a co-founder of GENOMEON, a company which is attempting to commercialize microsatellite discoveries.

REFERENCES

Albers,C.A. *et al.* (2011) Dindel: accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973.

Batzer,M.A. and Deininger,P.L. (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.*, **3**, 370–379.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Edwards,A. *et al.* (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.*, **49**, 746–756.

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

HanCock,J.M. (1999) Microsatellites and other simple sequences: genomic context and mutational mechanisms. In: Goldstein,D.B. and Schlötterer,C. (eds) *Microsatellites: Evolution and applications*. Oxford University Press, New York, pp. 1–9.

Highnam,G. *et al.* (2012) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.*, **41**, e32.

Homer,N. and Nelson,S.F. (2010) Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol.*, **11**, R99.

Hu,X. *et al.* (2012) pIRS: profile-based Illumina pair-end reads simulator. *Bioinformatics*, **28**, 1533–1535.

Kent,W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

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. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,S. *et al.* (2013) SOAPindel: efficient identification of indels from short paired reads. *Genome Res.*, **23**, 195–200.

Mackay,T.F. *et al.* (2012) The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, **482**, 173–178.

McIver,L.J. *et al.* (2011) Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics*, **97**, 193–199.

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

Scally,A. and Durbin,R. (2012) Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.*, **13**, 745–753.

Tae,H. *et al.* (2011) Revised genome sequence of *Brucella suis* 1330. *J. Bacteriol.*, **193**, 6410.

Tae,H. *et al.* (2012a) Complete genome sequence of *Brucella suis* VBI22, isolated from bovine milk. *J. Bacteriol.*, **194**, 910.

Tae,H. *et al.* (2012b) Improved variation calling via an iterative backbone remapping and local assembly method for bacterial genomes. *Genomics*, **100**, 271–276.

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.