To search personal genomes for STRs, the most cost-efficient way would be to resequence an entire personal genome and to collect billions of short reads of ~100 bp in length using available high-throughput sequencers. However, the infeasibility of obtaining longer reads at reasonable cost might lead to the failure to detect important STRs because expandable repeats associated with diseases can sometimes be quite long (e.g., $(ATTCT)_n$, n = 800-4500 in SCA10, and $(CCTG)_n$, n = ~5000 in DM2), and are much longer than 100 bp, the typical length of short reads, making the identification and location of long STRs in a personal genome nontrivial.

Another serious problem is that STRs have several variants with many mutations. The spontaneous mutation rate of STRs, 3.78×10^{-4} to 7.44×10^{-2} in the human Y-chromosome (Ballantyne, *et al.*, 2010), is far higher than the rate of copy number variation (CNV), 1.7×10^{-6} to 1.2×10^{-4} (Lupski, 2007), and the reported average rate of *de novo* single-nucleotide variation (SNV), 1.18×10^{-8} (S.D. 0.15×10^{-8}) (Conrad, *et al.*, 2011) and 1.20×10^{-8} (Kong, *et al.*, 2012). The ultrahigh mutation rate of STRs is thought to be a major force driving genetic variation producing a variety of STRs with differences often specific to personal genomes. Therefore, detecting various STRs by processing billions of short raw reads is fundamental to the analysis of personal genomes.

Several software programs list STRs, such as Tandem Repeat Finder (TRF) (Benson, 1999), Mreps (Kolpakov, *et al.*, 2003), ATRHunter (Wexler, *et al.*, 2005), IMEx (Mudunuri and Nagarajaram, 2007) and T-reks (Jorda and Kajava, 2009) (for a recent review that compares these programs, see Lim, *et al.*, 2013); however, these conventional programs are designed to retrieve STRs from nearly complete or draft long genomes and are not intended for processing billions of short reads in a reasonable amount of time. Another problem involved in handling short reads is the difficulty of determining the accurate positions of STRs in the genome because reads filled with STRs are not included in the genome or often map to multiple locations. The problem is solvable in some cases when a flanking region around an STR in a read is long enough to map to a unique position (Fig. 1B). To resolve these special cases, Gymrek *et al.* developed the program lobSTR (Gymrek, *et al.*, 2012), which improves the efficiency of this process by selecting approximately 240,000 candidate regions harboring STRs in the human genome. Due to severe restrictions in potential STR regions, however, we might overlook novel STRs hidden in numerous short reads because known STRs associated with diseases are frequently much longer than 100 bp, the typical length of short reads produced by high-throughput sequencers (Fig. 1C).

Here, we propose a new, cost-efficient method for calculating a comprehensive collection of STRs that are longer than short reads by inspecting the frequency distribution of STRs in short reads. To approximate the locations of such STRs, we utilize paired-end sequencing to facilitate locating the opposite end of the read with the focal STR in a pair, thereby narrowing down the location of the focal STR. Finally, we present a statistical procedure for selecting STRs that are significantly expanded in the case sample.

2 METHODS

2.1 Nonredundant representation of STRs

Our goal was to enumerate all possible instances of STRs with 2-6-base-long repeat units efficiently. In general, our algorithm can detect repeat units of an arbitrary length without sacrificing computational time. We also present an example of disease-associated STRs with a 10-base repeat unit in SCA31 (Sato, *et al.*, 2009). Care is required to avoid double counting identical STR occurrences characterized by more than one STR pattern. To remove redundancy, the basic unit of an STR should be minimized; e.g., the repeat unit of ACACACAC is AC rather than ACAC. Another reduction method is to merge occurrences of the reverse complement of an STR into the set of the focal STR. Therefore, we call the repeat unit representative if it is not a repeat of a shorter unit and is the first lexicographical motif when all possible shifts of the motif and its reverse complement are considered. Supplementary Table S1 presents the numbers of representative repeat units with typical examples.

2.2 Efficient algorithm for listing approximate STRs in billions of short reads

STRs are inherently "approximate" in the sense that some unit occurrences are allowed to contain a small number of mutations (Ballantyne, *et al.*, 2010). Listing approximate STRs, however, becomes computationally intractable because its time complexity grows exponentially in the maximum number of allowed mutations (Domanic and Preparata, 2007; Pellegrini, *et al.*, 2010). We therefore use a heuristic approach to this problem. We first identify "exact" STRs with no mutations in each short read using an efficient $O(n \log n)$-time algorithm (Main, 1989), where n is the length of the read. A *repetition* is any nonempty string of the form $(p)_m\, q$, where p, a nonempty string, is called the unit of the repetition, $m \geq 2$, and q is a prefix of p. For example,

$$(CAG)_3CA = CAGCAGCAGCA$$

is a repetition of the form $(p)_m\, q$, where p = CAG, $m = 3$, and q =CA, a prefix of p. A repetition is *maximal* if it is not a proper substring of a repetition that has the same unit. For example, consider:

$$(CAG)_2CA(CAG)_2CA= CAGCAGCACAGCAGCA$$

$(CAG)_2CA$, a repetition with unit CAG, is maximal. In addition, the entire string is also a maximal repetition with unit $(CAG)_2CA$. Listing all maximal repetitions is sufficient to identify all occurrences of STRs. We performed the following steps to retrieve STRs from each read.

1. Enumerate all maximal repetitions in a read using Main's $O(n \log n)$-time algorithm, where n is the length of the read (Main, 1989). More precisely, in 1984, Main and Lorentz designed an algorithm for enumerating all repetitions of the form xx (Main and Lorentz, 1984). In 1989, Main modified the algorithm to calculate maximal repetitions accurately (Main, 1989), and this is the version that we used to implement our system.

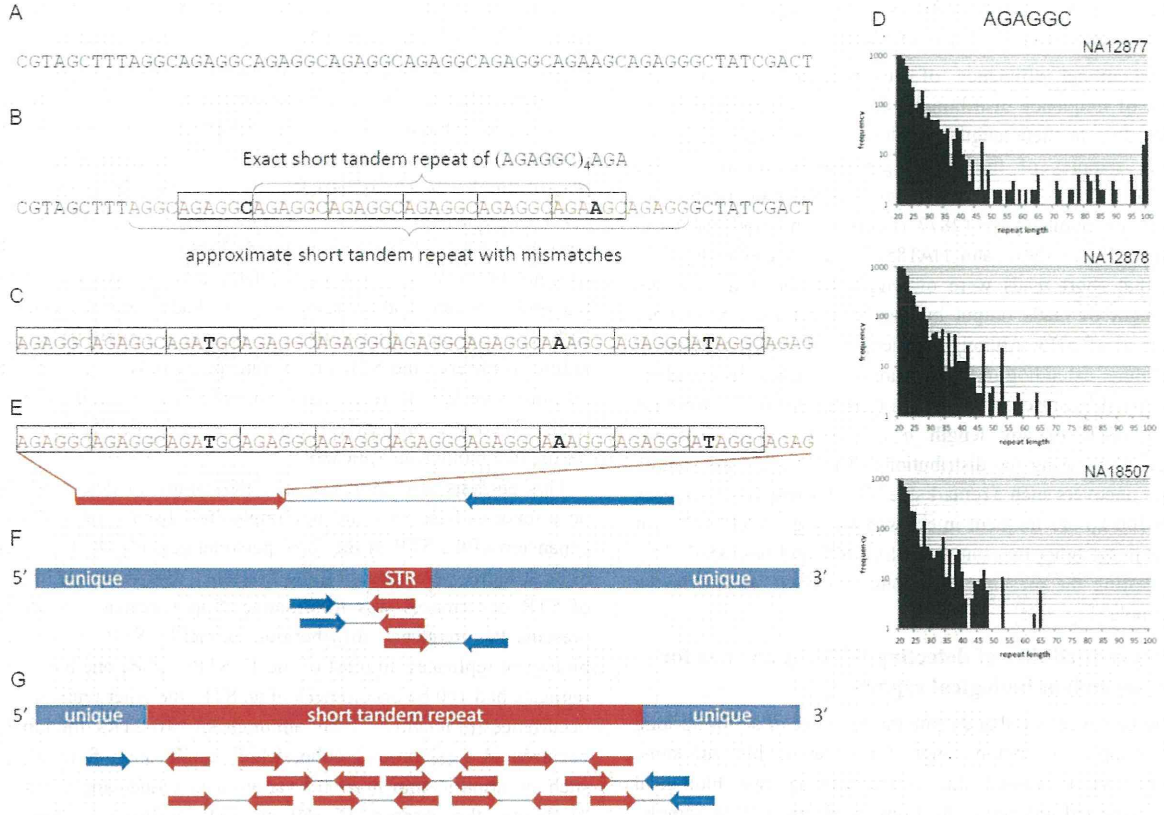
A
CGTAGCTTTAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAAGCAGAGGGCTATCGACT
B
Exact short tandem repeat of (AGAGGC)$_4$AGA
CGTAGCTTTAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAAGCAGAGGGCTATCGACT
approximate short tandem repeat with mismatches
C
AGAGGCAGAGGCAGATGCAGAGGCAGAGGCAGAGGCAGAGGCAAAGGCAGAGGCATAGGCAGAG
E
AGAGGCAGAGGCAGATGCAGAGGCAGAGGCAGAGGCAGAGGCAAAGGCAGAGGCATAGGCAGAG
F
5′
unique
STR
unique
3′
G
5′
unique
short tandem repeat
unique
3′
D
AGAGGC
NA12877
NA12878
NA18507
frequency
repeat length
1000
100
10
1
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Fig. 1. Sensing and locating short tandem repeats (STRs) in short reads. (A) An original short read. (B) An approximate STR (AGAGGC)*n* (*n*=6) in the short read. The central four copies of AGAGGC are an exact STR with no mutations, while the flanking copies contain the mutations shown in bold letters. If one of the regions (black) surrounding the STR aligns in a unique position, the STR can be located in the genome. (C) A read occupied by an approximate STR. (D) Sensing STRs from frequency distributions of (AGAGCC)*n* in NA12877 (father of the HapMap CEU trio), NA12878 (mother), and NA18507 (an African male). The x-axis is the lengths of STR occurrences detected in a read, and the y-axis is the frequency of reads containing STR occurrences of the length indicated on the x-axis. Note that 100-bp-long STR occurrences are frequent in NA12877, while no STR occurrences of length >70 bp are observed in samples NA12878 and NA18507. (E) When a read is filled with an STR (red), we attempt to anchor the other end read (blue) to a unique position unambiguously. (F, G) An STR is located easily if its location can be sandwiched using information on paired-end reads. The length of an STR of length < 100 bp is easily estimated (F), while determining the length of a much longer STR is nontrivial (G). We need to use third-generation sequencers, such as PacBio RS, with the capability of reading DNA fragments having a length of thousands of bases.

2. For each maximal repetition *Y*, identify the minimum unit *U* such that *U* is not a repetition and *Y* is a concatenation of multiple occurrences of *U* and a prefix of *U*. For example, when *Y* = (CAG)$_6$CA, *U* = CAG.
3. An approximate repetition is a substring such that its alignment with repetition (U)$_m$ is decomposed into series of exact matches of length |U| or more, and neighboring series must have only one mismatch, one insertion, or one deletion between them in the alignment, where |U| indicates the length of U. We calculate an approximate repetition by extending a maximal (exact) repetition in both directions in a greedy manner. For example, given

   CGCCCGCAGCGCAT(CAG)$_6$CATCAGGGA,

   we can extend repetition (CAG)$_6$CA to the underlined substring,

   CGCCCGCAGC-GCA**T**(CAG)$_6$CA**T**CAGGGA,

   where bold letters represent mismatches and "–" indicates a deletion. In this way, we retrieve an approximate STR that is not necessarily an exact repeat of the minimum unit *U*, but may contain mismatches and indels.
4. A read may contain multiple overlapping STRs with the same unit. If two overlap, eliminate the shorter one. If both are of the same length, select one arbitrarily.

The algorithm is able to process ten million reads of length 100 bases in ~1700 s on a Xeon X5690 with a clock rate of 3.47-GHz (Supplementary Fig. S1). As the computational time is proportional to the number of reads, ~47 hours is required to process 1 billion 100-bp reads, confirming the practicality of the method for processing real human resequencing data.

2.3 Sensing expanded STRs by analyzing the frequency distributions of STRs

The computational efficiency of our program facilitates the generation of frequency distributions of all approximate STRs in reads according to their lengths, as illustrated in Figure 1D. We used three samples of the whole genome resequencing data downloaded from http://www.illumina.com/platinumgenomes/ with accession numbers NA12877 (father of the HapMap CEU trio), NA12878 (mother), and NA18507 (an African male). We assumed that short reads were of length 100 bp, which is the typical length of reads output by cost-efficient high-throughput sequencers as of 2013. Although the length will likely increase in the near future, extending our procedure to process longer reads is straightforward because our algorithm runs in O(n log n)-time for processing reads of any length n as stated in the previous subsection. Comparing the distributions of more than one sample sometimes uncovers such a remarkable STR for which occurrences of length 100 bp are frequent in one sample (e.g., NA12877), but are absent in the other two samples, NA12878 and NA18507 (Fig. 1D), suggesting the presence of a long AGAGGC repeat in the former sample (Fig. 1D).

2.4 Reproducibility of detecting STR expansions for independent biological replicates

One might be concerned that despite the presence of a 100-bp-long STR in a sample, our method might fail to report this with some probability. We examined this concern using two biological replicates collected independently from an identical DNA sample. The two replicates were independent datasets of 100 bp reads sequenced from the same DNA sample, NA12878, using an Illumina HiSeq2000 (Supplementary Table S2). One dataset was collected by DePristo *et al.* (DePristo, *et al.*, 2011) and the other dataset was downloaded from Illumina's platinum genome web site (http://www.illumina.com/platinumgenomes/). We applied our method to both biological replicates (Supplementary Table S2) and examined whether 100 bp occurrences of individual STRs were present simultaneously in both. We identified 60 STRs with 100 bp occurrences in one (n = 13, 21.7%) or both (n = 47, 78.3%) replicates of NA12878 (Supplementary Table S3). Of the 13 STRs with no counts in one replicate, 12 had one or two occurrences in the other replicate, and the remaining one had four in the other. If an STR occurrence in the genome is short (*e.g.*, 100bp in length), failure to observe the STR has a high probability (e.g., 50% for 50-fold coverage of reads assuming the random collection of reads). Therefore, our method outputs essentially consistent results for the two biological replicates.

This analysis also indicated that the failure to detect 100 bp occurrences of an STR did not imply the absence of a 100 bp expansion of the STR in the focal personal genome. To be certain of its absence, we examined if the frequency distribution of lengths of STR occurrences was informative. Supplementary Figure S2 presents the frequency distributions of the 13 STRs in the two biological replicates. In most of the 13 STRs, when one biological replicate had 100 bp occurrences of an STR, the other replicate had occurrences of length > 90 bp, although for two STRs, the longest occurrences were around 60bp, which might stem from factors such as amplification bias and variation in sequencing coverage. Therefore, the absence of >60 bp STR occurrences does not necessarily deny the existence of 100 bp expansions of the STR in the genome.

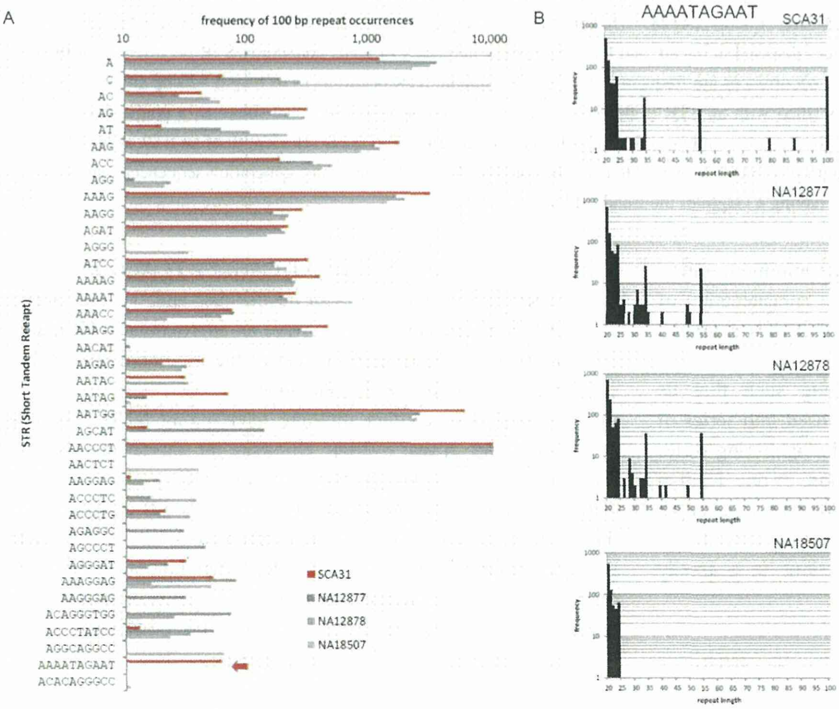

Fig. 2. Sensing expanded STRs associated with SCA31. (A) Frequencies of 100 bp STRs that have more than ten occurrences in one of either SCA31, NA12877, NA12878, or NA18507. For example, the arrow in the second lowest row shows that the (AAAATAGAAT) repeat is expanded only in SCA31. Our *ab initio* procedure analyzes this bar chart and selects STRs that are significantly abundant in the case sample (e.g., SCA31) but absent in all of the control samples. The bar chart is also useful for confirming the abundance of (AATGG) and (AACCCT) repeats, equivalent to the (GGGTTA) repeat, where the former and latter motifs are known to be enriched in centromeres and telomeres, respectively. (B) Frequency distributions of the (AAAATAGAAT) repeat. SCA31 has many 100 bp occurrences, while no occurrences of length >55 bp were observed in NA12877, NA12878, and NA18507.

2.5 Locating long expansions of STRs in the human genome

The genomic positions of each uncovered STR in a read remain to be determined. The problem is solvable if one of the two regions flanking an STR maps to a unique position (Fig. 1B), the method used in lobSTR (Gymrek, *et al.*, 2012). Otherwise, we attempt to use information on paired-end reads, the two ends of an identical DNA fragment such that their typical average length ranges from 300 to 350 bp with an average standard deviation of ~10%. When one end-read is filled with an STR, we test whether the other end maps to a unique position in the genome using the Burrows-Wheeler Alignment Maximal Exact Matches algorithm (BWA-MEM), a tool for aligning reads with the genome (Li, 2013). If the test is successful, we can approximate the position of the STR from the location of the other end (Fig. 1E). An STR can be located if its location can be sandwiched using information on paired-end reads (Fig. 1F,G). An STR shorter than 100 bp is easier to determine (Fig. 1F), while estimating the lengths of longer STRs becomes more difficult (Fig. 1G). We will discuss this issue later.

2.6 TRhist: a tool for sensing and locating STRs from billions of short reads

To assist in the correct positioning of STRs, for a read with an STR instance, our program outputs the repeat unit, length of the STR, number of mutations in the STR, flanking regions surrounding the STR, and other paired-end read. With this information, the user can align the flanking regions and other end read to the reference to locate the STR in the genome. Our TRhist program is available at http://trhist.gi.k.u-tokyo.ac.jp/.

2.7 SMRT^{TM} sequencing of expanded STRs

Successful identification of an accurate position for one end provides useful input for other analytical methods, such as repeat-primed PCR (Warner, *et al.*, 1996) and SMRT^{TM} sequencing (Eid, *et al.*, 2009; Loomis, *et al.*, 2013), to estimate or determine long expansions of STRs. In particular, SMRT^{TM} sequencing is capable of reading DNA fragments of average length ~5 kb (Fig. 1G). Using this emerging technology, Loomis *et al.* reported the first sequence, 750 CGG repeats, for fragile X syndrome (Loomis, *et al.*, 2013). Using SMRT^{TM} sequencing, we amplified the repeat region associated with SCA31 using PCR primers

1.5k-ins-F (5′- ACTCCAACTGGGATGCAGTTTCTCAAT-3′) and

1.5k-ins-R (5′- TGGAGGAAGGAAATCAGGTCCCTAAAG-3′).

We will describe the analysis in the Results. PCR was performed in a final volume of 50 µl containing 0.2 µM of each primer, 200 µM of each dNTP, 1 mM MgCl2, 1.25 U of PrimeSTAR HS DNA polymerase (Takara Bio, Otsu, Japan), and 100 ng of genomic DNA. The PCR profile comprised an initial denaturing at 95°C for 5 min followed by 30 cycles at 95°C for 20 s and 68°C for 8 min. The PCR product was purified on 0.8% agarose gels and converted to the proprietary SMRTbell™ library format using an RS DNA Template Preparation Kit 2.0 (Pacific Biosciences, Menlo Park, CA). Briefly, the PCR product was end-repaired, and hairpin adapters were ligated using T4 DNA ligase. Incompletely formed SMRTbell^{TM} templates were degraded with a combination of exonuclease III and VII. The resulting DNA templates were purified using SPRI magnetic beads (AMPure; Agencourt Bioscience, Beverly, MA). Annealing was performed at a final template concentration of 5 nM, with a 20-fold molar excess of sequencing primer. The annealing reaction was carried out for 2 min at 80°C with slow cooling to 25°C. Annealed templates were stored at -20°C until polymerase binding. The DNA polymerase enzymes stably were bound to the primed sites of the annealed SMRTbell^{TM} templates using the DNA Polymerase Binding Kit 2.0 (Pacific Biosciences). SMRTbell^{TM} template (3 nM) was incubated with polymerase in the presence of phospholinked (Pacific Biosciences) nucleotides for 4 h at 30°C. Following incubation, the samples were stored at 4°C. Sequencing was performed within 36 h of binding. Samples were sequenced using commercial sequencing chemistry. Sequencing data were collcted on a PacBio RS (Pacific Biosciences) for 90 min. Given PacBio RS filtered subreads, we used the SMRT Pipe, P_ErrorCorrection module to generate corrected reads. Subsequently, we assembled these corrected reads using RS_CeleraAssembler to obtain contigs.

3 .RESULTS

Here, we demonstrate the utility of an *ab initio* procedure for sensing, locating, and sequencing STRs that are significantly expanded in the case sample.

Locating candidate STR positions

Select positions where STR occurrences are expanded significantly in the case sample in the following manner:

1. Locate occurrences of each candidate STR in both the case and control samples by anchoring paired-end reads such that one end has a $\geq$50 bp occurrence of the STR and the other end maps to a unique position.
2. Group paired-end reads anchored in a neighborhood (within ~300 bp, the average insert size of paired-end reads) into one cluster (Fig. 3).
3. In each cluster, generate the frequency distribution of STR occurrences according to their lengths ranging from 50 to 100 bp (Fig. 3). If an STR in the cluster is significantly longer than 100 bp, the frequency of 100 bp occurrences in reads, denoted by f_{100}, becomes significantly greater than the frequencies of those shorter than 100 bp (Fig. 3B). We test this hypothesis statistically by checking if f_{100} is an outlier in the frequency distribution with the Smirnov-Grubbs' test. We calculate the t-score, $(f_{100} - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the frequency distribution, respectively, and obtain the probability (p-value) that the t-score exceeds a threshold according to the Smirnov-Grubbs' test. For example, the p-value is $<5 \times 10^{-9}$ when the t-score is >5.27.

4. We consider approximately 10 million nonoverlapping regions of length 300 bp (the average insert size of paired-end reads) in the human genome. We perform multiple hypothesis testing using the Bonferroni correction to test if each 300 bp region has a significant STR expansion in the case sample at a significance level of 5% divided by 10 million (i.e., 5×10^{-9}). We select positions such that $p < 5 \times 10^{-9}$ in the case sample but no 100 bp STR occurrences are present in any of the control samples. We can relax the condition to consider more candidates with less evidence.

Sequencing candidate STR positions

SMRT$^{\text{TM}}$ sequencing of expanded STRs is performed using information on the boundaries of individual STR positions.

3.1 A rare STR significantly expanded in the case sample

To demonstrate the effectiveness of this approach, we first examined a well-characterized case sample, SCA31 (Sato, *et al.*, 2009), which contains long expansions of two STRs, (AAAATAGAAT) repeat and (AATGG) repeat, in the introns of genes BEAN1 and TK2 (Chr.16 66,524,303 in hg19), where the reference genome has an (AAAAT) repeat.

We resequenced the genome of a sample from an individual whose parent is a case of SCA31 using an Illumina HiSeq2000 (Supplementary Table S4). All primary sequencing data of the SCA31 sample will be made available under controlled access through the DNA Databank of Japan (DDBJ; accession number DRA000898). We examined whether we could find these STRs with no prior information. We applied the *ab initio* procedure to SCA31 as the case sample, and NA12877, NA12878, and NA18507 as control samples (Fig. 2A). Our procedure detected only one STR; AAAATAGAAT ($p = 1.07 \times 10^{-19}$).

Figure 2B shows the frequency distributions of the (AAAATAGAAT) repeat, supporting the presence of long occurrences of the STR in SCA31 and the absence of long occurrences of length >60 bp in the other control samples. Supplementary Figure S3A shows the distributions of the (AATGG) repeat, but the difference between SCA31 and the other samples was unclear because the (AATGG) repeat is enriched in human centromeres (Grady, *et al.*, 1992). Therefore, our *ab initio* analysis suggests that long occurrences of the (AAAATAGAAT) repeat characterize SCA31, consistent with reported observations (Sato, *et al.*, 2009). Arguably, we could detect the (AAAATAGAAT) repeat as an approximate (AAAAT) repeat because the last half, AGAAT, is identical to (AAAAT), except for the second base G; therefore, we analyzed the frequency distribution of the (AAAAT) repeat to determine the remarkable expansion of the (AAAAT) repeat in SCA31. This failed due to numerous long instances of the (AAAAT) repeat in all samples

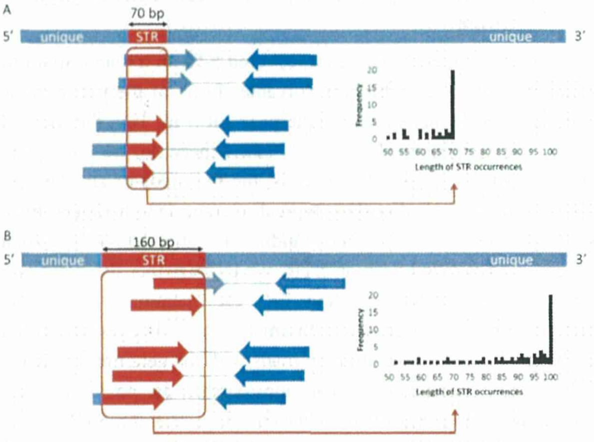

Fig. 3. Select positions where STR occurrences are expanded significantly (A) We generate the frequency distribution of lengths of STR occurrences in paired-end reads. This picture shows the case of a 70-bp-long STR. The histogram of the frequency distribution peaks at 70 bp. (B) When the STR is 160 bp long, the distribution has a significant peak at 100 bp. We test if the peak is a significant outlier in the frequency distribution using the Smirnov-Grubbs' test.

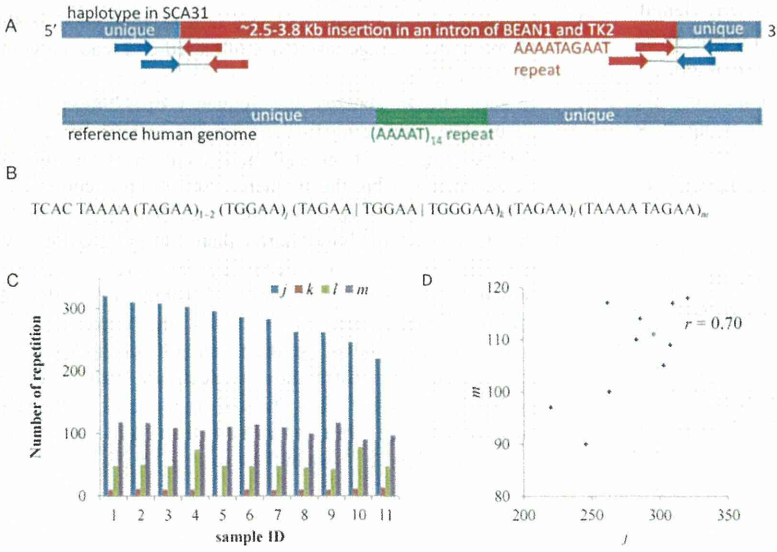

Fig. 4. Locating and sequencing expanded STRs associated with SCA31. (A) A real example from SCA31. One haplotype contains a ~2.5-3.8 kb insertion at Chr.16 66,524,303 in hg19 in an intron of BEAN1 and TK2. The right boundary of the insertion could be identified using paired-end reads with AAAATAGAAT repeats at their left ends and uniquely mapped reads at their right ends. The lower bar illustrates the reference genome (hg19) with an AAAAT repeat. (B) A form of expanded repeat associated with SCA31 samples. The values of *i*, *j*, *l* and *m* vary in the individual SCA31 samples. (C) We determined the values of *i*, *j*, *l* and *m* in eleven SCA31 samples using SMRT$^{\text{TM}}$ sequencing. This shows that ~90% of the repeat expansion are (TAGAA)$_j$ and (TAAAA TAGAA)$_m$. (D) The values of *j* and *m* are positively correlated (r = 0.70). These two values are the determinants of the instability of the repeat expansions in SCA31.

(Supplementary Fig. S3B). This example indicates the importance of looking at STRs of repeat units longer than 2-6-base units, to determine expansions of STRs associated with cases.

We also examined the frequency distributions of other well-characterized repeats, such as the (GGGTTA) repeat in telomeres (Supplementary Fig. S3C), (CAG) repeat encoding poly-glutamine stretches in protein coding regions (La Spada, *et al.*, 1991; The Huntington's Disease Collaborative Research Group, 1993; Walker, 2007; Supplementary Fig. S4A), (CCTG) repeat associated with myotonic dystrophy type 2 (DM2; Liquori, *et al.*, 2001; Supplementary Fig. S4B), and (ATTCT) repeat associated with spinocerebellar ataxia type 10 (SCA10; Matsuura, *et al.*, 2000; Supplementary Fig. S4C). For the last three repeats, no significant differences were detected between SCA31 and the three control samples, suggesting that these three repeats are not associated with SCA31.

Using paired-end reads with AAAATAGAAT repeats at their 5'-ends and uniquely mapped reads at their 3'-ends, we could determine the 3'-end of the insertion. Figure 4A shows how we locate a ~2.5-3.8 kb insertion of the repeat associated with the SCA31 sample (Sato, *et al.*, 2009).

We sequenced the repeat region in eleven SCA31 samples using SMRT$^{\mathrm{TM}}$ sequencing. We designed a pair of PCR primers around the candidate repeat region in the SCA31 sample the right boundary of which could be determined. As illustrated in Figure 4A, we could identify the left boundary in the reference genome because the left ends of many paired-end reads mapped to the upstream region of the left boundary, while the right ends did not. We could sequence the candidate repeat region. Supplementary Table S6 presents the statistics of filtered subreads, corrected subreads, and assembled contigs. Previously, Sato *et al.* estimated a 2.5-3.8 kb insertion of the following form for an SCA31 sample (Sato, *et al.*, 2009):

TCAC TAAAA (TAGAA)$_i$(TGGAA)$_j$
(TAGAA | TGGAA | TGGGAA)$_k$
(TAGAA)$_l$(TAAAA TAGAA)$_m$,

where (TAGAA | TGGAA | TGGGAA)$_k$ is a series of k occurrences of TAGAA, TGGAA, and TGGGAA. In their sample, they determined that i = 2, k = 10, and l = 46, but left j and m undetermined because both appeared to be extremely long. In our eleven SCA31 samples, we could determine the values of j and m. We found that the numbers of individual repeats varied markedly (i = 1~2, j = 220~321, k = 9~13, l = 42~78, and m = 90~118) and the insertion size ranged from 2,350 to 3,088 b (Fig. 4C, Supplementary Table S5), demonstrating the instability of the STR expansion in SCA31. In particular, two STRs, (TAGAA)*j* and (TAAAA TAGAA)*m*, form ~90% of the entire repeat expansion, and the values of *j* and *m* are positively correlated (correlation coefficient r = 0.70), implying that these two values are the determinants of the instability of the repeat expansions in SCA31 (Fig. 4D). In all samples, the repeat expansion was present in one allele, but was absent in the other. Note that the numbers of STR units might not be exact because PCR for repeat regions can introduce more replication errors than those produced by bacterial DNA replication (Loomis, *et al.*, 2013).

3.2 Rare STRs significantly expanded in the case sample

We also applied our procedure to the SCA31 data, and examined common STRs, AAAG, ATCC, AAAAG, AATAG, and AATGG, present in both the case and control samples but significantly expanded in the case sample. We identified STR expansions at eleven genomic locations that were significantly expanded in the case sample ($p < 5 \times 10^{-9}$, Supplementary Figure S5). We then used SMRT$^{\mathrm{TM}}$ sequencing to confirm the four expanded STRs in the case sample that were significantly longer than the corresponding STR occurrences in the reference genome (Figure 5, Supplementary Figure S6). No false-positive expansions were found in this experiment, suggesting that the false-positive rate of the procedure is generally quite low.

4 DISCUSSION

STRs in personal genomes remain largely uncharacterized. We proposed a novel method for listing long approximate STRs with mutations in personal genomes utilizing a massive number of short reads of length ~100 bp. Here, we discuss some situations in which detecting a long expansion of STRs specific to disease samples is inherently problematic. As genomic regions of GC content > 70% are difficult to cover with an ample number of Illumina reads, our method is unlikely to detect long expansions of STRs with high GC contents. STRs in reads originating in centromeres, telomeres, or retrotransposons are too numerous to map to unique genomic positions. As illustrated in Supplementary Figure S3, massive numbers of long expansions of these STRs can be found in any sample.

We also presented an *ab initio* procedure for detecting significant expansions of STRs in case samples that are absent in control samples via comparisons between the frequency distributions of STRs in case and control samples. We demonstrated the potential applicability of this method using three publicly available control samples. To exploit this approach, however, constructing a large-scale database of the frequency

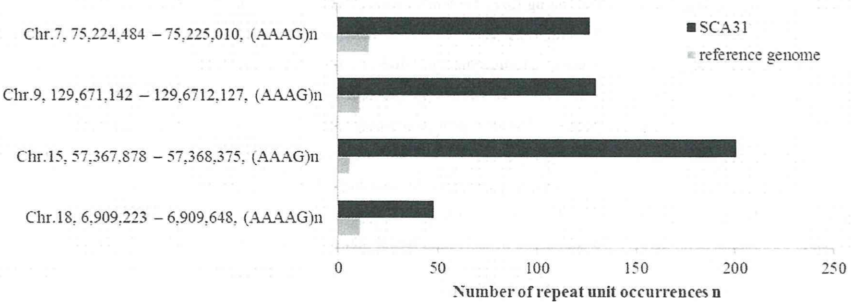

Fig. 5. Sizes of the common STRs, (AAAG)*n* and (AAAAG)*n*, at four genomic positions in the SCA31 sample and reference genome. Note that individual STR occurrences are significantly expanded in the SCA31 sample. The PCR primers used for amplifying individual regions and the sequences of amplicons can be found in Supplementary Figure S6.

distributions of STRs collected from a number of control samples is necessary.

The variety of expanded STRs of length >1 kb in disease remains unexplored. Also, examining whether expansions of STRs are more pronounced in germline and somatic cells would be intriguing. Thus, after locating STRs, sequencing expanded STRs is a promising direction of study. For this purpose, SMRT™ sequencing enables the sequencing DNA fragments averaging ~5 kb long. Indeed, using SMRT™ sequencing, we were able to determine a divergent set of 2,3-3.1 kb STR sequences in eleven SCA31 samples, showing the instability of STR expansions. Analysis of the stability of STR expansions in germline and somatic cells of a specific disease might eventually lead to the recognition of a functional role of STRs.

In the near future, the typical lengths of short reads in the majority of commercial sequencers should increase to 150–500 bases. Our method is ready to process longer reads in a straightforward manner. Furthermore, our method was designed so that it could output STRs of repeat units of any length, and we presented an illustrative case in which detecting STRs of a 10-base repeat unit from an SCA31 sample was essential. Our program will serve as a valuable tool for discovering unknown STRs in a variety of diseases, even with future advances in sequencing technology.

ACKNOWLEDGEMENTS

We are grateful to an anonymous reviewer for suggesting a way of detecting expansions of common STRs and for many valuable comments on the manuscript.

Funding: This work was supported in part by Grant-in-Aid for Scientific Research on Innovative Areas [22129008, 221S0002] from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and by the Global COE program (Deciphering Biosphere from Genome Big Bang) from the MEXT.

REFERENCES

Ballantyne, K.N., *et al.* (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications., *American journal of human genetics*, **87**, 341-353.

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

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.

Conrad, D.F., *et al.* (2011) Variation in genome-wide mutation rates within and between human families., *Nature genetics*, **43**, 712-714.

DeJesus-Hernandez, M., *et al.* (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS, *Neuron*, **72**, 245-256.

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

Domanic, N.O. and Preparata, F.P. (2007) A novel approach to the detection of genomic approximate tandem repeats in the Levenshtein metric, *J Comput Biol*, **14**, 873-891.

Eid, J., *et al.* (2009) Real-time DNA sequencing from single polymerase molecules, *Science*, **323**, 133-138.

Grady, D.L., *et al.* (1992) Highly conserved repetitive DNA sequences are present at human centromeres, *Proc Natl Acad Sci U S A*, **89**, 1695-1699.

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

Jorda, J. and Kajava, A.V. (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm, *Bioinformatics*, **25**, 2632-2638.

Kobayashi, H., *et al.* (2011) Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement, *Am J Hum Genet*, **89**, 121-130.

Kolpakov, R., Bana, G. and Kucherov, G. (2003) mreps: Efficient and flexible detection of tandem repeats in DNA, *Nucleic Acids Res*, **31**, 3672-3678.

Kong, A., *et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk, *Nature*, **488**, 471-475.

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.

La Spada, A.R., *et al.* (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy, *Nature*, **352**, 77-79.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv preprint arXiv:1303.3997*.

Lim, K.G., *et al.* (2013) Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance, *Brief Bioinform*, **14**, 67-81.

Liquori, C.L., *et al.* (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9, *Science*, **293**, 864-867.

Loomis, E.W., *et al.* (2013) Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene, *Genome Res*, **23**, 121-128.

Lupski, J.R. (2007) Genomic rearrangements and sporadic disease., *Nature genetics*, **39**, S43-47.

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

Main, M.G. (1989) Detecting Leftmost maximal periodisities, *Discrete Applied Mathematics* **25**, 145-153.

Main, M.G. and Lorentz, R.J. (1984) an O(n log n) algorithm for finding all repetitions in a string, *Journal of Algorithms*, 422-432.

Matsuura, T., *et al.* (2000) Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10, *Nat Genet*, **26**, 191-194.

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

Mudunuri, S.B. and Nagarajaram, H.A. (2007) IMEx: Imperfect Microsatellite Extractor, *Bioinformatics*, **23**, 1181-1187.

Orr, H.T. (2011) FTD and ALS: genetic ties that bind, *Neuron*, **72**, 189-190.

Pellegrini, M., Renda, M.E. and Vecchio, A. (2010) TRStalker: an efficient heuristic for finding fuzzy tandem repeats, *Bioinformatics*, **26**, i358-366.

Renton, A.E., *et al.* (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD, *Neuron*, **72**, 257-268.

Sato, N., *et al.* (2009) Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n, *Am J Hum Genet*, **85**, 544-557.

Sherman, S.L., *et al.* (1985) Further segregation analysis of the fragile X syndrome with special reference to transmitting males, *Hum Genet*, **69**, 289-299.

The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group, *Cell*, **72**, 971-983.

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.

Walker, F.O. (2007) Huntington's disease, *Lancet*, **369**, 218-228.

Warner, J.P., *et al.* (1996) A general method for the detection of large CAG repeat expansions by fluorescent PCR, *J Med Genet*, **33**, 1022-1026.

Wexler, Y., *et al.* (2005) Finding approximate tandem repeats in genomic sequences., *Journal of computational biology : a journal of computational molecular cell biology*, **12**, 928-942.

Wojciechowska, M. and Krzyzosiak, W.J. (2011) Cellular toxicity of expanded RNA repeats: focus on RNA foci, *Hum Mol Genet*, **20**, 3811-3821.

A Simple but Powerful Heuristic Method for Accelerating k-Means Clustering of Large-Scale Data in Life Science

Kazuki Ichikawa and Shinichi Morishita

Abstract—K-means clustering has been widely used to gain insight into biological systems from large-scale life science data. To quantify the similarities among biological data sets, Pearson correlation distance and standardized Euclidean distance are used most frequently; however, optimization methods have been largely unexplored. These two distance measurements are equivalent in the sense that they yield the same k-means clustering result for identical sets of k initial centroids. Thus, an efficient algorithm used for one is applicable to the other. Several optimization methods are available for the Euclidean distance and can be used for processing the standardized Euclidean distance; however, they are not customized for this context. We instead approached the problem by studying the properties of the Pearson correlation distance, and we invented a simple but powerful heuristic method for markedly pruning unnecessary computation while retaining the final solution. Tests using real biological data sets with 50-60K vectors of dimensions 10–2001 (~400 MB in size) demonstrated marked reduction in computation time for $k = 10$-500 in comparison with other state-of-the-art pruning methods such as Elkan's and Hamerly's algorithms. The BoostKCP software is available at http://mlab.cb.k.u-tokyo.ac.jp/~ichikawa/boostKCP/.

Index Terms—Bioinformatics, clustering, mining methods and algorithms, optimization

1 Introduction

Clustering, an unsupervised learning algorithm to group data into similar categories, has been widely used to gain insights into biological systems from large-scale biological data, such as gene expression data monitored by microarrays [1], [2], [3], [4], histone modifications [5], [6], [7], [8], [9], [10], [11], [12], [13], and nucleosome positioning [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. A variety of clustering algorithms, such as hierarchical clustering, k-means clustering, self-organizing map (SOM), and principal components analysis (PCA), have been used (for review, see [25]). Of these, k-means clustering is the most widely used to process large-scale data sets, in part because the computational complexity of hierarchical clustering is quadratic or higher in the number of data points, while k-means clustering algorithms have lower computational complexity [26]. Accelerating k-means clustering algorithms is still necessary to process the growing volume of biological data due to the recent progress in data collection by next-generation sequencing.

The basic concept of k-means clustering is simple.

1) It first selects k cluster centroids in some manner. The behavior of the algorithm is highly sensitive to the initial selection of k initial centroids, and many efficient initialization methods have been proposed to calculate better k centroids [26], [27], [28], [29], [30], [31], [32], [33]. In this study, we use the initialization method proposed by Bradley and Fayyad [31], since it consistently performs better than the other methods in terms of several criteria according to the recent report by Celebi et al. [26].
2) Subsequently, k-means clustering repeats the process of assigning individual points to their nearest centroids and updating each of k centroids as the mean of points assigned to the centroid until no further changes occur on the k centroids [34].

Quantifying the same data points is essential. Various measures are available, such as Euclidean distance, Manhattan distance, Pearson correlation distance, and Spellman rank correlation. Of these, Euclidean distance and Pearson correlation distance have been widely used for large-scale biological data processing [3], [4], [24], [35], [36]. Euclidean distance is sensitive to scaling, while correlation is unaffected by scaling. Precisely, given two data of high dimension such that their patterns are quite similar but their scales are different, Euclidean distance is not suitable for measuring the similarity. To avoid this problem, standardized Euclidean distance, which is not sensitive to scaling, is frequently used [3], [36], [37], [38], [39], [40].

Of note, standardized Euclidean and Pearson correlation distances are equivalent in the sense that both yield the same k-means clustering result for identical sets of k initial centroids because the standardized Euclidean distance is proportional to the square root of the Pearson correlation distance [3], [40], and the two distances always produce consistent orderings. Thus, optimization methods designed to calculate one distance are applicable to the other.

- *The authors are with the Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-0882, Japan. E-mail: {ichikawa, moris}@cb.k.u-tokyo.ac.jp.*

Manuscript received 29 Dec. 2013; accepted 24 Jan. 2014. Date of publication 12 Feb. 2014; date of current version 4 Aug. 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2306200

Despite the importance of the Pearson correlation and standardized Euclidean distances for machine learning, optimization methods customized for these distances are largely unexplored. In general, several efficient k-means clustering algorithms have been proposed for processing Euclidean distances by utilizing the triangle inequality [41], [42], [43] or by analyzing the correlation coefficient between the centroids [44]. Thus, we can use optimization methods for the Euclidean distance to yield a k-means clustering result based on the standardized Euclidean distance that is in agreement with that based on the Pearson correlation distance [3].

We instead examined the properties of the Pearson correlation distance and devised a simple and novel method for avoiding unnecessary computation in order to boost k-means clustering using the Pearson correlation distance. We demonstrate that our method outperforms pruning method applications using the Euclidean distance [41], [42], [43] compared with those that use the standardized Euclidean distance. Our method has been best optimized for k-means clustering using the standardized Euclidean and Pearson correlation distances.

2 METHODS

We first introduce the definition of Pearson's correlation coefficient.

Definition. *To measure the distance between two d dimensional vectors $x = (x[1], \dots, x[d]), y = (y[1], \dots, y[d])$, we define Pearson's correlation coefficient:*

$$\rho(x, y) = \frac{1}{d}\sum_{i=1}^{d}\left(\frac{x[i]-\overline{x}}{\sigma_x}\right)\left(\frac{y[i]-\overline{y}}{\sigma_y}\right),$$

where $\overline{x}$ denotes the average of $x[1], \dots, x[d]$, and σ_x is the standard deviation, defined as $\sqrt{\sum_{i=1}^{d}(x[i]-\overline{x})^2/d}$. Let x denote the length, defined as $\sqrt{\sum_{i=1}^{d} x[i]^2}$.

Note that Pearson's correlation coefficient ranges from -1 to 1, i.e., $-1 \le \rho(x, y) \le 1$. The Pearson's correlation coefficient $\rho(x, y)$ itself does not serve as a distance because when x and y are more similar to each other, $\rho(x, y)$ becomes larger and approaches 1 rather than 0.

Definition. *[45] The Pearson correlation distance $dis(x, y)$ is defined as $1 - \rho(x, y)$.*

The Pearson correlation distance approaches 0 when x and y are similar. In contrast, when x and y are more dissimilar, the Pearson's correlation coefficient decreases to -1, and the Pearson correlation distance between x and y increases approaching 2. The range of the distance is $0 \le \mathrm{dis}(x, y) \le 2$. The Pearson correlation distance violates the triangular inequality.

Example. When $x_1 = (9, 3, 1), x_2 = (3, 1, 9)$, and $x_3 = (1, 3, 9)$, we have $\mathrm{dis}\,(x_1, x_2) = 1.5$, $\mathrm{dis}\,(x_2, x_3) = 0.115$, and $\mathrm{dis}(x_1, x_3) = 1.846$, which do not meet the triangular inequality:

$$\mathrm{dis}(x_1, x_2) + \mathrm{dis}(x_2, x_3) \ge \mathrm{dis}(x_1, x_3)$$

We illustrate here two examples that clarify how the Pearson correlation distance differs from the Euclidean distance.

Example. When $x_1 = (1, 3, 9), x_2 = (0.9, 0.3, 0.1)$, and $x_3 = (0.1, 0.3, 0.9)$ x_1 and x_3 have similar patterns, but their scales are different, while x_2 and x_3 have dissimilar patterns, yet their Euclidean distance is smaller than the distance between x_1 and x_3. Indeed, we have:

$$\mathrm{dis}(x_1, x_3) = 0 < 1.84615 = \mathrm{dis}(x_2, x_3),$$

while

$$||x_1 - x_3|| = 8.58545 > 1.13137 = ||x_2 - x_3||.$$

The next example illustrates the discrepancy between the Pearson correlation distance and the "normalized" Euclidean distance.

Example. When $x_1 = (0.1, 0.3, 10), x_2 = (0.1, 1, 10)$, and $x_3 = (0.1, 0.1, 1)$, Pearson correlation distances meet

$$\mathrm{dis}(x_1, x_3) = 0.00016 < 0.00338 = \mathrm{dis}(x_2, x_3)$$

implying that x_3 is more similar to (correlated with) x_1. In contrast, the normalized Euclidean distance yields the opposite ordering:

$$\left|\left|\frac{x_1}{||x_1||} - \frac{x_3}{||x_3||}\right|\right| = 0.11304 > 0.08920 = \left|\left|\frac{x_2}{||x_2||} - \frac{x_3}{||x_3||}\right|\right|.$$

We next define the standardized Euclidean distance.

Definition. *Let* $\mathrm{dis\_SE}(x, y)$ *denote*

$$\sqrt{\sum_{i=1}^{d}\left(\frac{\mathrm{x}[i]-\overline{x}}{\sigma_x} - \frac{y[i]-\overline{y}}{\sigma_y}\right)^2},$$

the standardized Euclidean distance between two d dimensional vectors x and y.

The square root of the Pearson correlation is proportional to the standardized Euclidean distance.

Proposition. *[3], [40]*

$$\sqrt{2d}\sqrt{\mathrm{dis}(x, y)} = \mathrm{dis\_SE}(x, y).$$

The Pearson correlation distance and the standardized Euclidean distance produce consistent orderings; namely, for any x_1, y_1, x_2, y_2,

$$\mathrm{dis}(x_1, y_1) \le \mathrm{dis}(x_2, y_2),$$

if and only if

$$\mathrm{dis\_SE}(x_1, y_1) \le \mathrm{dis\_SE}(x_2, y_2).$$

We note here that the Pearson correlation distance and its square root are largely different. For example, $\sqrt{\mathrm{dis}(x, y)} = 0.4$ when $\mathrm{dis}(x, y) = 0.16$, and $\sqrt{\mathrm{dis}(x, y)} = 1.3$ when $\mathrm{dis}(x, y) = 1.69$. In general, two proximal (distal, respectively) points of the Pearson correlation distance <1 (>1) become more distant (closer) according to the square root of the Pearson correlation distance.

Next, we outline Lloyd's algorithm, which implements k-means clustering. Given n points in d dimensional space, a k-means algorithm starts with selecting k initial centroids, $\{c_p | p = 1, \dots, k\}$, in some way. It then repeats the following