

A Novel Method for Identifying Length Variations of Short Tandem Repeats Based on Next Generation Sequencing and Its Application in Human Genetic Disease Research*

YAN Zhang-Ming¹⁾, WANG Yao²⁾, LIU Ke³⁾, XIANG Shu-Nian¹⁾, SUN Zhi-Rong^{1)**}

¹⁾MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China;

²⁾The First Affiliated Hospital of Nanchang University, Nanchang 330006, China;

³⁾Department of Statistics, University of California, Berkeley, CA 94720, USA)

Abstract Next generation sequencing (NGS) technologies boosted genomic and medical research, particularly for identification of disease-causing variants. Although most types of genetic variants could be identified through NGS data analysis, there are still some limitations, such as length variations of short tandem repeats (STRs). Many genetic diseases are known to be caused by expansions of STRs, especially neurological disorders, such as Huntington disease. However, almost none of existing tools could detect STRs expanded longer than sequencing read length based on NGS. To break through the limitation, we developed a novel method for detecting length variations of STRs and estimating the length of expansions based on paired-end NGS. We applied our method in a clinical study of motor neuron disease using whole-exome sequencing and successfully identified a disease-causing expansion of STR. Our method firstly used special features of depth of read coverage at STRs to address the variant calling problem. It has widely application value in human genetic disease research and inspirational value in developing new NGS data processing tools.

Key words next generation sequencing, short tandem repeats, length variation, motor neuron disease

DOI: 10.16476/j.pibb.2016.0144

Next-generation sequencing (NGS) technologies have revolutionized the approaches and strategies used in genomic and medical research. In particular, in the era of precision medicine, personal genomics provides the most valuable information for clinical diagnosis and treatment. Various methods and tools have been developed to detect different types of genetic variations from NGS data, such as single-nucleotide variations (SNVs), indels and copy number variations. Such tools are constantly evolving and have accelerated the search for disease-causing genes and variations.

However, some limitations in detecting variations from NGS data remain. One of the limitations is the identification and analysis of long expansions of short tandem repeats (STRs). STRs are DNA repeats of 2-6-nucleotide stretches. When the change in length is

minuscule (several base pairs), it can be identified as an indel by certain variant-calling tools. If the change is large, some special tools such as lobSTR can detect it^[1]. However, if the length of the STR variation is greater than the sequencing read length (typically about 100 bp), almost none of existing methods can be used to identify them^[2]. Although a tool called TRhist can analyze such variations, it uses hybrid sequencing data of both NGS and third-generation sequencing^[3]. The analysis of length variations of STRs, especially

*This work was supported by grants from The National Natural Science Foundation of China (31171274) and National Basic Research Program of China (2012CB725203).

**Corresponding author.

Tel: 86-10-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

Received: April 22, 2016 Accepted: May 13, 2016

those longer variations, is important because many genetic diseases are known to be caused by expansions of STRs, especially neurological disorders [4-5]. For example, Huntington disease is caused by expansion of CAG repeats in HTT gene [6]. When the CAG repeat count is more than 40, it will cause Huntington disease. Such long STR variation in Huntington disease is hard to be identified from NGS data using existing variant calling methods.

To overcome the limitation of STR analysis, we developed a novel method to identify large length variations in STRs, referred to as long STR variations. Our method firstly used a special feature of depth of read coverage using paired-end sequencing to detect long STR variations. We also used the depth information to estimate length of long STR variation based on regression model. To validate the new method, we applied it in a study of familial motor neuron disease (MND) using whole-exome sequencing. Motor neuron disorders (MNDs) are a clinically and pathologically heterogeneous group of neurologic diseases characterized by progressive degeneration of motor neurons [7]. It is hard to make a precise diagnosis from the clinical syndrome of MND. Using our new method, we successfully identified disease-causing long STR variation at androgen receptor gene (AR) and specifically diagnosed the MND as spinobulbar muscular atrophy (SBMA) [8]. The

application of our method in MND study showed its effectiveness and widely usage value in human genetic disease research. Moreover, the idea behind our method might inspire development of novel NGS data analysis tools.

1 Methods

1.1 Detection of long STR variation

Long STR variation refers specifically to STRs that have expanded longer in the personal genome than in the reference genome and the sequencing read. This kind of STR variation could not be directly detected by existing NGS data processing tools [2]. To address this problem, we carefully studied the characteristics of sequencing reads mapping in the STR region and finally designed a novel strategy for detecting long STR variation using only paired-end sequencing data. This strategy was based on the feature of depth profile in the STR region. The design and concept of our strategy are illustrated in Figure 1, which shows an STR region of the reference genome that is expanded much longer in the personal genome. Furthermore, the expanded STR in the personal genome is longer than the sequencing read length. Hence, it is a typical long STR variation. Read pairs of paired-end genome or exome sequencing were randomly distributed in the personal genome, such as read pairs A, B and C in Figure 1a. However, when mapping these read pairs to

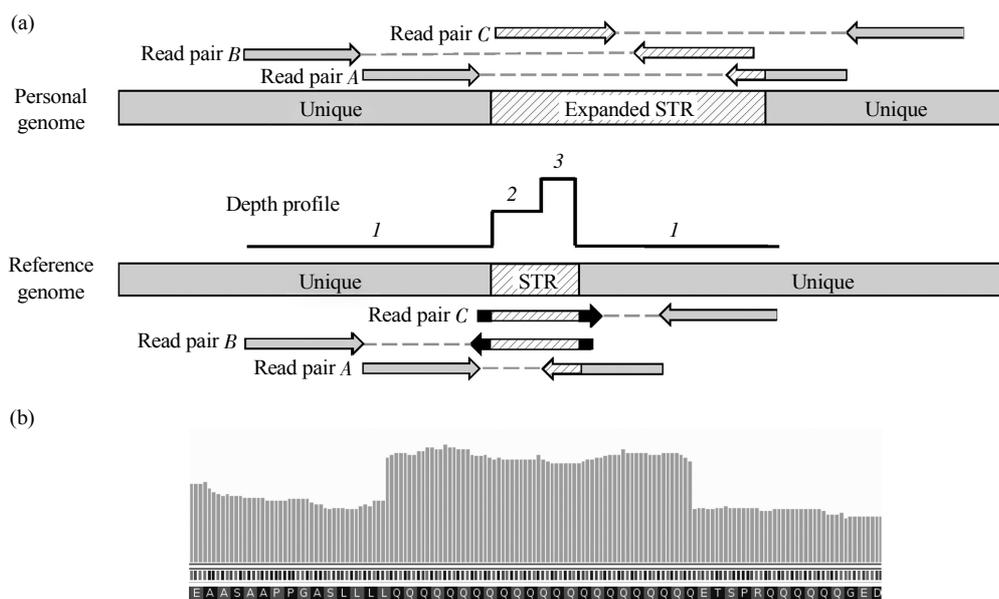


Fig. 1 Illustration of depth shift at the STR region

(a) Schematic diagram of reads mapping in the region of long STR variation. The slash and grey represents STR and unique region, respectively. Three read pairs are connected with dashed lines. The depth profile is generated based on the three read pairs mapping. (b) A depth shift phenomena in a real paired-end genome sequencing data.

the reference genome, an interesting phenomenon was observed. Reads completely in the expanded STR region (in slash in the figure) could be mapped to the reference genome using the information of their paired reads in a unique sequence region. However, as they were repeat sequences, they could not be mapped unambiguously in the STR region. As a result, they piled up in the shortened STR region. Finally, we found a sudden level shift in read depth in the STR region compared to the unique sequence regions on both sides, which is called depth shift. Figure 1b shows a real depth shift in a paired-end genome sequencing data, which is consistent with our inference.

Therefore, long STR variation could be identified by detecting a depth shift. For computational implementation, we converted this problem to outlier detection of time-series data. The left part of Figure 2 shows the procedure. First, we extracted the depth profile of the STR region (length L_0 in reference genome) and upstream and downstream regions with sequencing read length l . Then we got the depth profile y in region $l+L_0+l$. Next, we calculated the difference between adjacent positions to generate a difference profile dy with $dy_i=y_{i+1}-y_i$. If depth shift existed, there would be two significant outlier points in the

difference profile corresponding to the boundaries of depth shift. We used the tsoutliers R package to detect and test these outliers^[9]. A significant likelihood of the presence of outliers indicated a long STR variation.

Exome-wide screening of long STR variation was implemented by using the annotated human STRs database downloaded from a "simple repeats" track of the UCSC Table browser^[10-11]. We set the period of STR at less than 7 and the match percentage at more than 80%, and we filtered out STRs not located in the exon region of known genes. As a result, 2 858 STRs would be tested, more than 98% of which were shorter than 100 bp in the reference genome. The depth profiles of only these STRs were extracted and analyzed.

1.2 Length estimation of long STR variation

STR length varies in different individuals, and it is an important factor in determining disease^[5-6, 8]. Thus, length estimation of long STR variation is essential. After detecting long STR variation, we used a novel method to estimate the length of it based on a regression model of depth profile.

The whole procedure is illustrated in the flow chart of Figure 2. For each detected long STR variation, we extracted all the mapped reads

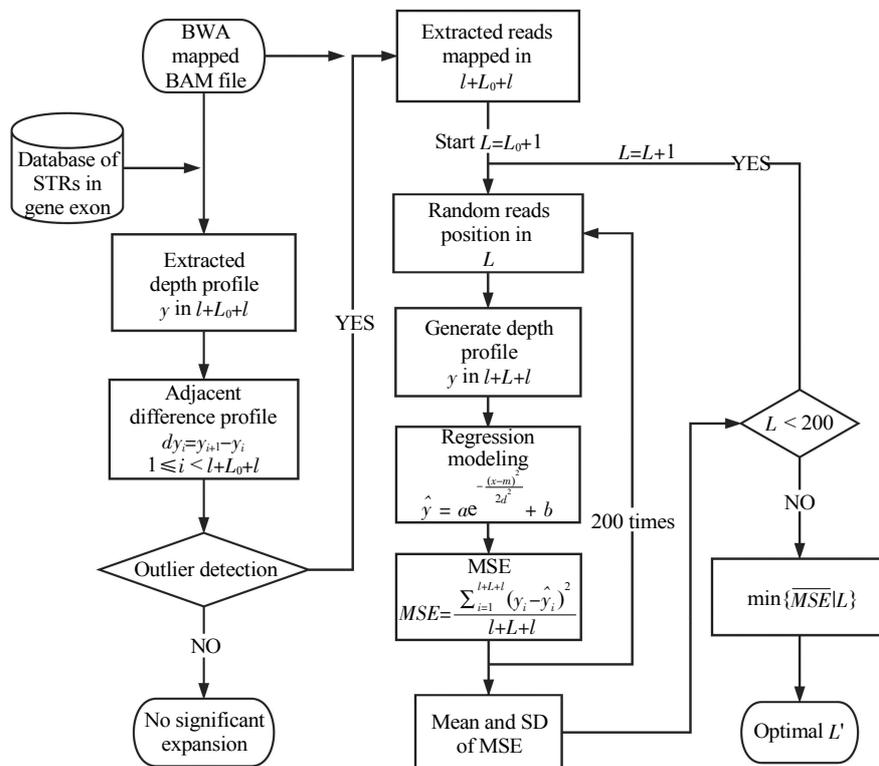


Fig. 2 Flowchart of our method of detecting long STR variation and estimating its length

The procedure of our method starts from mapped reads file in BAM format. The left part of the flowchart is the method of detecting long STR variations. The right part illustrates the method of estimating the length of long STR variation.

overlapping to an extended region of the STR in the reference genome, which was composed of an STR region and upstream and downstream regions of the read length. The extended region is denoted as a region of $l+L_0+l$, where l is the read length and L_0 is the original length of the STR in the reference genome. There are two rational assumptions. First, the sequencing reads are locally randomly distributed. Therefore, for an assigned length of STR variation L , we could randomly distribute all the reads piled up in the STR region L and generate a new depth profile of the $l+L+l$ region. The second assumption is that the depth profile is continuous, allowing us to build a regression model based on the new depth profile. For whole genome sequencing, it could be a linear regression model. Here we discuss whole-exome sequencing. Because of the coverage bias in whole-exome sequencing, we built a regression model based on the Gaussian function, which is as follows:

$$\hat{y} = ae^{-\frac{(x-m)^2}{2d^2}} + b \quad (1)$$

Where a, b, d, m are regression parameters, x is the relative position, and \hat{y} is the predicted depth. Then, we calculated mean square error of residuals as follows:

$$MSE = \frac{\sum_{i=1}^{l+L+l} (y_i - \hat{y}_i)^2}{l+L+l} \quad (2)$$

Where MSE is the mean square error, i is the relative position in the $l+L+l$ region, and y_i, \hat{y}_i are the real and predicted depth of position i , respectively. For the actual length of long STR variation L^* , the depth

profile in the $l+L^*+l$ region would have the best fit, *i.e.*, minimum mean square error. As a result, the problem becomes an easy optimization problem of finding the optimal L to satisfy:

$$\min\{MSE|L\} \quad (3)$$

As we randomly distributed the reads in the L region, there would be a bias in calculating the mean square error. Thus, for each L , we repeated the procedure 200 times. We calculated the mean value and standard deviation of MSE . Therefore, the final optimization problem is:

$$\min\{\overline{MSE}|L\} \quad (4)$$

Finally, the estimated length of long STR variation is the optimal L with the minimum \overline{MSE} , denoted as L' . Besides, we preferred to determine the confidence interval of optimal L' . As we had mean and standard deviation of MSE corresponded to each L , so we used the t -test to test the significant difference of MSE between optimal L' and other L . The L without significant difference of MSE from that of optimal L' comprised the confidence interval.

All the computation processes were implemented with R (<http://www.R-project.org>)^[12].

2 Materials and data processing

2.1 A case of familial MND

A case of familial MND was investigated in this work. Figure 3 shows the pedigree of the family, in which three men, II -3, II -9 and III -1, were affected. The average age of disease onset was 40.7 years (43, 40 and 39 years, respectively). Lower limb weakness was the common presenting symptom in all the three patients, followed by upper limb weakness, tremors,

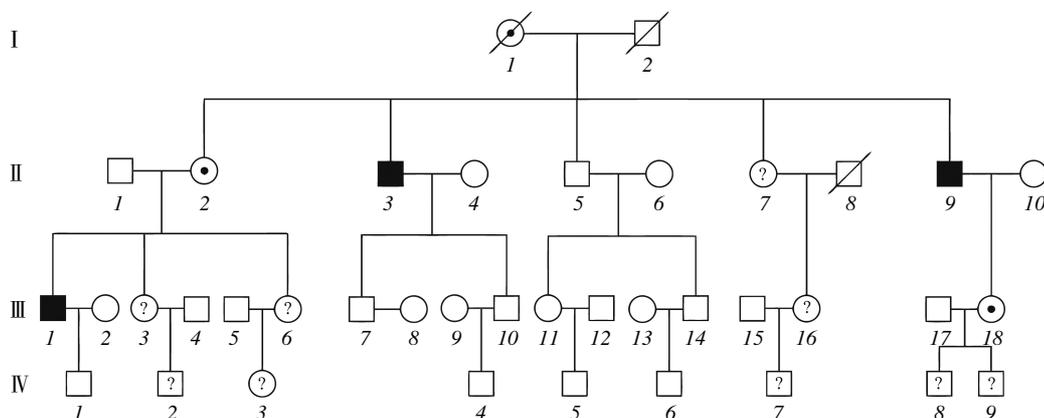


Fig. 3 Pedigree of the case of familial MND

There are four generations in the pedigree. Square and circle represent male and female, respectively. Three male patients are represented as black squares. The black dot in the middle of a circle means female carrier. The question mark indicates unknown genotype.

and bulbar weakness. The onset of weakness was asymmetrical, and weakness occurred on the dominant side first. Before onset, they had experienced hand tremors for about ten years, and two patients (II -9 and III -1) underwent muscle biopsy and received a general diagnosis of motor neuron disease.

2.2 Whole-exome sequencing of the family

The genomes of 12 individuals of the family (II -1, II -2, II -3, II -5, II -7, II -9, III -1, III -3, III -4, III -18, and IV -2) were sequenced using whole-exome sequencing. Each participant provided written informed consent for participating in the study. Genomic DNA was extracted from peripheral blood leukocytes. The DNA samples were randomly fragmented by Covaris into fragments of 250 bp in average length (200~300 bp). The exome sequencing libraries were prepared using Roche NimbleGen SeqCap EZ. Each library was sequenced on a HiSeq2000 platform with 2 × 90-bp paired-end read length and 100× average depth of read coverage.

2.3 Variant calling

All the whole-exome sequencing data were mapped to the GRCh37/hg19 reference genome (obtained from the UCSC Genome Browser) using the BWA-MEM algorithm with default parameters^[13-14]. The mapped reads were saved in BAM format. The mapped BAM files of whole-exome sequencing were processed for variant calling and filtering with the best practice pipeline of the Genome Analysis Toolkit (GATK, version 3.3)^[15]. GATK could detect SNVs and indels. SnpEff and SnpSift were used to annotate and filter the genetic variants with dbSNP, dbNSFP, and other human variant annotation databases^[16-18]. SnpSift was used to select suspect variants based on variant type, population frequency, predicted impact, and other information.

lobSTR (version 3.0.3) was used to profile STRs of all the 12 sequenced individuals^[1]. lobSTR could only detect and profile STRs shorter than the sequencing read length (90 bp in our data). We compared the profiled STRs between patients and healthy people in the family and filtered out those STR variations that also existed in healthy people.

2.4 Genetic linkage analysis

Considering pedigree information would help us to find the underlying genetic variants of familial disease, we applied genetic linkage analysis to the whole-exome sequencing data to detect the chromosomal location of the disease genes. First, we

used the linkdatagen tool to generate LINKAGE style files from whole-exome sequencing data for MERLIN, which is a popular pedigree analysis package^[19-20]. HapMap Phase II SNPs were used as genetic map positions required for linkage analysis^[21]. Then, the non-parametric linkage analysis of MERLIN was implemented on the prepared LINKAGE style files. The pedigree information is provided in Figure 3. Finally, the results of linkage analysis were summarized as genetic linkage regions.

3 Results

3.1 Genetic linkage analysis results

We performed genetic linkage analysis in the family under study using whole-exome sequencing data and pedigree information. Only one linkage region was consistent with the disease inheritance pattern in the family, which is chrX: 47444985-142121729. The disease has a typical X-linked recessive inheritance. The disease-related linkage region was used to filter variants, as well as long STR variations.

3.2 Variant calling by GATK and lobSTR

We performed a standard variant-calling process of whole-exome sequencing data using GATK and obtained the genotype of each individual. We filtered out variants having allele frequencies over 0.001 annotated in the dbSNP database. Variants existing in the healthy individuals of the family or absent in the patients were also filtered out. Finally, only 15 missense SNVs remained as suspect variants related to the disease. However, there was no clue about how these genes contributed to MNDs.

lobSTR was used to profile the STRs of each individual. We compared the results of lobSTR between healthy individuals and patients in the family. No significant difference was found.

As a result, conventional variant calling approaches of whole-exome sequencing data failed to reveal any credible disease-causing gene in this case of familial MND.

3.3 Long STR variation analysis

STRs are important and related to many diseases, but lobSTR could only detect STR variations shorter than read length. To better investigate the STR variations in the whole-exome sequencing data of the family studied, we applied our novel method to analyze long STR variations.

3.3.1 Detection of long STR variations

Through exome-wide screening of long STR variations, we found only one depth shift located in the disease-related linkage region, and it existed in all three patients (II -3, II -9 and III -1) and carrier II -2 but absent in all other healthy individuals. In Figure 4a, we plotted the depth shift signal detected by tsoutliers in patient III -1, which has quite significant effects. This depth shift indicated a long STR variation of the

AR gene, located in chrX: 66765159-66765227. For intuitively viewing the depth shift, the read coverage displayed in the Integrative Genomics Viewer is shown in Figure 4b^[22]. We showed the trio of II -1, II -2, and III -1. III -1 was a patient and had the most significant depth shift. In healthy individual II -1, there was no depth shift. II -2 (mother of III -1) was a carrier and had a moderate depth shift because one of her AR gene alleles had long STR variation.

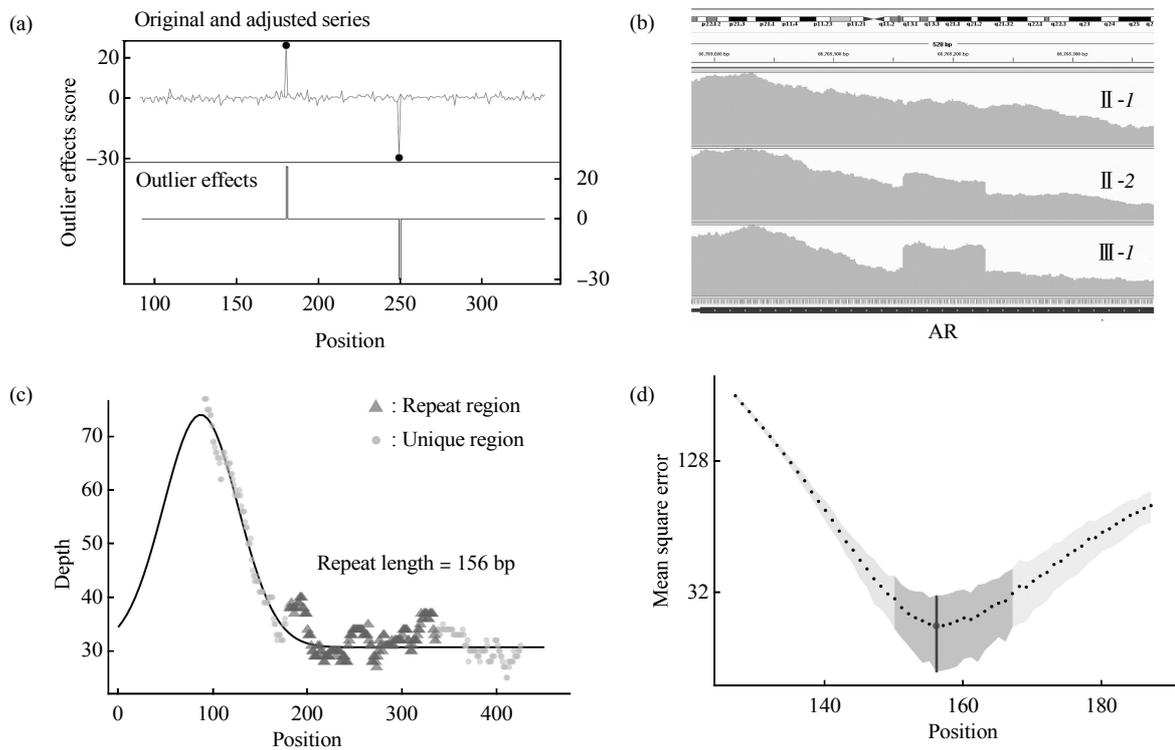


Fig. 4 Results of long STR variation analysis in the MND case

(a) Outlier signals of adjacent difference profile of depth in patient III -1. Up is the adjacent difference profile of depth and two black points indicate the detected significant outliers. The bottom is the outlier effects. The two outliers represent the boundaries of long STR variation. (b) Depth profiles of II -1, II -2 and III -1 displayed in IGV. (c) A regression model of depth profile with an assigned 156 bp long STR variation. Dark gray triangles are residuals of depth profile in long STR variation region. Light gray dots represent residuals in unique sequence regions of read length on the both sides. (d) The curve of mean square error (log2 scale) versus length of long STR variation. *MSE* is shown in log2 scale. The black dot indicates the mean *MSE* with each value of the length of long STR variation and the gray shade represents the range of standard deviation. Dark gray shade indicates the confidence interval of length estimation of long STR variation. The vertical black line shows the length with the minimum mean *MSE*.

3.3.2 Length estimation of long STR variation

Furthermore, we used our novel method to estimate the length of long STR variation. Figure 4c depicts an instance of regressed Gaussian function and residuals of depth profile with the best estimation of long STR variation in patient III -1. Figure 4d shows the optimization process of mean square error (MSE) with varying lengths of long STR variation. The length of long STR variation was assigned from 120 bp to 200 bp.

For each assigned value, we generated 200 random distributions of mapped reads in STR region and built regression model based on depth profile of each distribution. Moreover, we calculated MSE of each regression model and its residuals. Then we got mean MSE and standard deviation of MSE at each value of assigned length of long STR variation. The best estimation of long STR variation is the length with the minimum mean MSE. As shown in Figure 4d, the

best estimation of long STR variation in patient III-1 is 156 bp. The dark gray area represents the confidence interval of the estimated length in patient III-1, which was 151~166 bp. This estimation was consistent with the results of the other two patients II-3 and II-9, which were 149~168 bp and 153~165 bp. Besides, the estimated length of carrier II-2 was 115~140 bp with the best estimation of 123 bp. As II-2 was heterozygous, so the estimated length was an average effect of 2 different alleles, simplified as the mean length of the two alleles.

3.3.3 Experimental validation

We used agarose gel electrophoresis and Sanger sequencing to experimental validate the long STR variation. Firstly, we did PCR to amplify the androgen receptor (AR) GCA repeated region with primers: forward, 5' GCC TGT TGA ACT CTT CTG AGC 3', and reverse, 5' GCT GTG AAG GTT GCT GTT CCT C 3'. Then the PCR products were subjected to electrophoresis and Sanger sequencing.

Figure 5a shows the agarose gel electrophoresis results of II-1, II-2, and III-3. It is evident that patient III-1 had longer GCA repeats than II-1. There were two bands in the gel electrophoresis of II-2 because she had two alleles of different length. Figure 5b shows the Sanger sequencing results of III-1, having 51 GCA repeats, 153 bp in total. The other two patients had the same length with III-1. Sanger sequencing results of three patients were consistent with the estimation of our novel computational method. Besides, as PCR tended to amplify shorter sequences, the length of the repeats in II-2 could not be determined directly by Sanger sequencing. However, using the Sanger sequencing results of II-3, we could infer that II-2 had one allele of the same length with II-3, which was 84 bp. So the GCA repeats in II-2 were 84 bp and 153 bp long. Moreover, the mean of the two alleles was 118 bp, which was also consistent with our estimation (123 bp with 115~140 bp confidence interval).

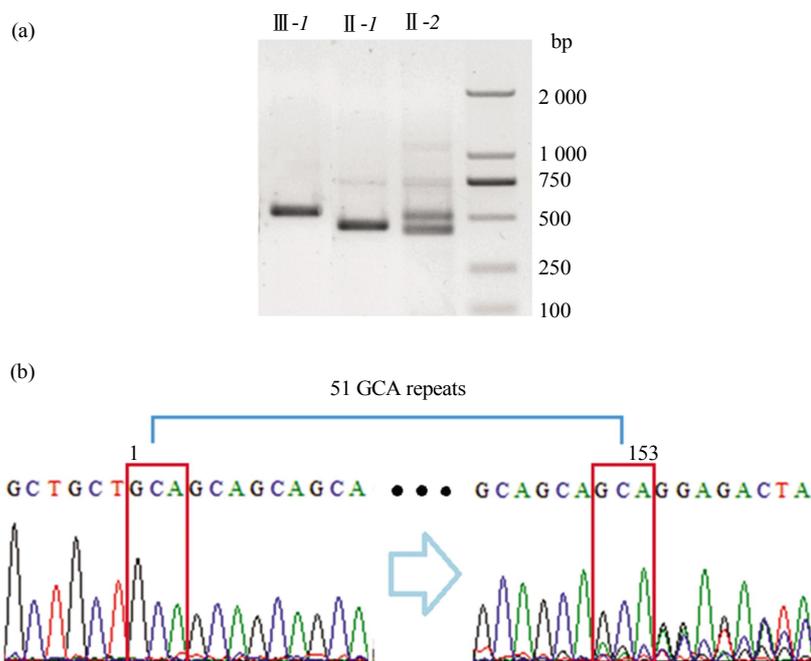


Fig. 5 Experimental validation of long STR variation

(a) Agarose gel electrophoresis results of GCA repeats PCR products of II-1, II-2 and III-1. (b) Sanger sequencing results of III-1. As it is a long sequence, so the repeat region is partially shown in the figure.

3.3.4 Re-diagnosis of the case

We found that the length variation of the exonic GCA repeats of the AR gene was the most reasonable disease-causing variant in the family. GCA repeat length variations in AR have been reported to cause

SBMA, which is an adult-onset, slowly progressive motor neuropathy^[8]. Furthermore, the syndrome of SBMA has been found to vary among different cases, sometimes being misdiagnosed as amyotrophic lateral sclerosis or other MNDs^[23-24]. It might explain why the

patients we investigated had not been diagnosed before. Finally, using our novel method of long STR variation analysis, we diagnosed the patients as having SBMA. It showed that genetic testing was quite necessary for the diagnosis of neurological disorders. Besides, we also found the long GCA repeat variation in individual IV -2, who was a teenager without any symptoms. However, because of our diagnosis of SBMA, he has received early intervention.

4 Discussion

Length variations of STRs are quite important in human genetic diseases. However, STRs in personal genomes remain largely uncharacterized. To overcome the limitation of analyzing STRs using NGS data, we originally used features of coverage depth to solve the problem of STR analysis in personal genomes. We proposed a novel method for identifying long STR variations. Through a study of familiar MND, we validated our method and showed its application value in human genetic disease research. In the near future, third generation sequencing may totally solve problems of identifying long STR variations because its average read length is more than 1 kb. As the error rate of third generation sequencing is large now because of immature technologies, so it is still not widely used. However, our method can use a large number of existing genome or exome NGS data to characterize STR variations and discover disease-related long STR variations.

There are some factors influence the effectiveness of our method. Firstly, the mean depth of read coverage is obvious one important factor. As our method is based on depth profile, so low coverage depth would affect both detection of depth shift and regression modeling of depth profile. Fortunately, recent NGS-based techniques, especially for exome sequencing, mean depth of read coverage is always more than 100×, which is sufficient for our methods. Secondly, insert size of paired-end sequencing (related to library fragment size) is another key factor, which limits the detectable length of long STR variations. Most STRs in the reference genome are shorter than 100 bp, but their variations in the personal genome are unpredictable. If the length of STR variation is much longer than insert size of read pairs, lots of read pairs would be totally repeated sequences, which could not be mapped to reference genome. So the depth profile is

incomplete and can not be used to identify long STR variations. Hence, the length of STR variations can be identified is limited by the mean insert size of paired-end sequencing. Most of NGS libraries are fragmented to 200~300 bp fragments, which is proper for most of long STR variations. Finally, there is a limitation of our methods, which is heterozygous. As different alleles might have different lengths of STR, so the length estimated by our method is a kind of mean value of both alleles, which can not be separated. However, we have shown that the mean value estimated by our method was consistent with real mean value of two alleles in II -2 of the MND family.

Our method is effective and accurate in identifying long STR variations. We are currently developing an easy-to-use R package for other researchers. Moreover, we believe that our method will inspire the development of other NGS data processing tools. The application of our method in familial MND also suggested that we should consider the limitations of general NGS data analysis to arrive at accurate conclusions in human genetic disease research. There are still many types of variations apart from STR can not be well characterized by NGS, new technology, such as third-generation sequencing, might give us more information about personal genomes in the precision medicine era.

References

- [1] Gymrek M, Golan D, Rosset S, *et al.* lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res*, 2012, **22** (6): 1154-1162
- [2] Merkel A, Gemmell N. Detecting short tandem repeats from genome data: opening the software black box. *Brief Bioinform*, 2008, **9**(5): 355-366
- [3] Doi K, Monjo T, Hoang P H, *et al.* Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics*, 2014, **30**(6): 815-822
- [4] Madsen B E, Villesen P, Wiuf C. Short tandem repeats in human exons: a target for disease mutations. *BMC Genomics*, 2008, **9**: 410
- [5] Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res*, 2008, **18** (7): 1011-1019
- [6] Walker F O. Huntington's disease. *Lancet*, 2007, **369** (9557): 218-228
- [7] Tiryaki E, Horak H A. ALS and other motor neuron diseases. *Continuum (Minneapolis Minn)*, 2014, **20** (5 Peripheral Nervous System Disorders): 1185-1207
- [8] Sulek A, Hoffman-Zacharska D, Krysa W, *et al.* CAG repeat

- polymorphism in the androgen receptor (AR) gene of SBMA patients and a control group. *J Appl Genet*, 2005, **46**(2): 237–239
- [9] Baragona R, Battaglia F. Outliers detection in multivariate time series by independent component analysis. *Neural Comput*, 2007, **19**(7): 1962–1984
- [10] Karolchik D, Hinrichs A S, Furey T S, *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 2004, **32**(Database issue): D493–496
- [11] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, 1999, **27**(2): 573–580.
- [12] Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 1996, **5** (3): 299–314
- [13] Kent W J, Sugnet C W, Furey T S, *et al.* The human genome browser at UCSC. *Genome Res*, 2002, **12**(6): 996–1006
- [14] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009, **25**(14): 1754–1760
- [15] Mckenna A, Hanna M, Banks E, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 2010, **20**(9): 1297–1303
- [16] Cingolani P, Platts A, Wang Le L, *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, 2012, **6**(2): 80–92
- [17] Liu X, Wu C, Li C, *et al.* dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat*, 2016, **37**(3): 235–241
- [18] Sherry S T, Ward M H, Kholodov M, *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 2001, **29** (1): 308–311
- [19] Smith K R, Bromhead C J, Hildebrand M S, *et al.* Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol*, 2011, **12**(9): R85
- [20] Abecasis G R, Cherny S S, Cookson W O, *et al.* Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 2002, **30**(1): 97–101
- [21] Sung Y J, Gu C C, Tiwari H K, *et al.* Genotype imputation for African Americans using data from HapMap phase II versus 1000 genomes projects. *Genet Epidemiol*, 2012, **36**(5): 508–516
- [22] Thorvaldsdottir H, Robinson J T, Mesirov J P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 2013, **14** (2): 178–192
- [23] Harutunian G M, Beydoun S R, Rison R A. Kennedy disease misdiagnosed as polymyositis: a case report. *BMC Res Notes*, 2013, **6**: 389
- [24] Parboosingh J S, Figlewicz D A, Krizus A, *et al.* Spinobulbar muscular atrophy can mimic ALS: the importance of genetic testing in male patients with atypical ALS. *Neurology*, 1997, **49** (2): 568–572

一种基于二代测序辨识短串联重复序列长度变异的新方法及其在人类遗传疾病研究中的应用*

严章明¹⁾ 王瑶²⁾ 刘珂³⁾ 向书念¹⁾ 孙之荣^{1)**}

¹⁾ 清华大学生命科学学院生物信息学教育部重点实验室, 北京 100084; ²⁾ 南昌大学第一附属医院, 南昌 330006;

³⁾ *Department of Statistics, University of California, Berkeley, CA 94720, USA*

摘要 二代测序技术的涌现推动了基因组学研究, 特别是在疾病相关的遗传变异研究中发挥了重要作用. 虽然大多数遗传变异类型都可以借助于各种二代测序分析工具进行检测, 但是仍然存在局限性, 比如短串联重复序列的长度变异. 许多遗传疾病是由短串联重复序列的长度扩张导致的, 尤其是亨廷顿病等多种神经系统疾病. 然而, 现在几乎没有工具能够利用二代测序检测长度大于测序读长的短串联重复序列变异. 为了突破这一限制, 我们开发了一个全新的方法, 该方法基于双末端二代测序辨识短串联重复序列长度变异, 并可估计其扩张长度, 将其应用于一项基于全外显子组测序的运动神经元疾病临床研究中, 成功地鉴定出致病的短串联重复序列长度扩张. 该方法首次原创性地利用测序读长覆盖深度特征来解决短串联重复序列变异检测问题, 在人类遗传疾病研究中具有广泛的应用价值, 并且对于其他二代测序分析方法的开发具有启发性意义.

关键词 二代测序, 短串联重复序列, 长度变异, 运动神经元疾病

学科分类号 R857.3, Q31

DOI: 10.16476/j.pibb.2016.0144

* 国家自然科学基金(31171274)和国家重点基础研究发展计划(2012CB725203)资助项目.

** 通讯联系人.

Tel: 010-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

收稿日期: 2016-04-22, 接受日期: 2016-05-13