Automatically Predicting Possible Loci of Variable Number of Tandem Repeats

Chia-Hung Chang[1], Han-Yu Chuang[1], Yi-Hung Chiang[1], Chien-Shun Chiou[2], Hsueh-I Lu[3], and Cheng-Yan Kao[1]
1. Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan
2. Center for Disease Control, Taiwan
3. Institute of Information Science, Academia Sinica, Taipei, 115, Taiwan
Correspondence to: Han-Yu Chuang
Email: {r91056, r90002}@csie.ntu.edu.tw
Phone: +886-2-23635336 #401

Abstract

Variable number of tandem repeats (VNTR) stands for a tandem repeat which has variation in length and the number of repeated segments between individuals. VNTRs are useful as molecule markers in many applications, such as DNA fingerprinting, genetic disease analysis and molecular typing of prokaryotes. It costs a large amount of money and time to identify these special loci by using traditional biological experiments. Here we develop a novel tool, VNTR analyzer, to identify and analyze VNTR in genome sequence by comparing genomic sequences of different strains. In this study, we demonstrate its ability to detect VNTRs by analyzing 3 bacterial species: ***Staphylococcus aureus***, ***Xylella fastidiosa*** and ***Salmonella enterica***. *The results showed that our program could find VNTRs accurately in a short time. Moreover, the program provides biologists a colorful visualization tool for further analysis.*

1. Introduction

Short-sequence DNA repeats (SSR) occur in large quantities in many eukaryotic and prokaryotic genomes [19]. Some of them are called variable number of tandem repeats (VNTR) [23] or minisatellites [1] as their numbers of repeated segments vary between individuals. VNTRs are inherently unstable units that undergo frequent variation in the number of copies through slipped strand misalignment during DNA synthesis [12]. These phenomena imply that VNTR regions seem to be evolutionary hotspots. The variation is seen when a population contains fragments of different lengths in the same genomic region [5,13]. It is recently shown as a useful technique in DNA fingerprinting for forensic investigation and paternity analysis to distinguish individuals by combination of alleles at VNTR loci [23]. Moreover, polymorphic VNTR regions are also be used for the molecular typing of several bacterial species, including *Yersinia pestis* [7], *Francisella tularensis* [2], *Salmonella enterica* [22], *Mycobacterium tuberculosis* [8], *Xylella fastidiosa* [10], *Haemophilus influenzae* [20], *Staphylococcus aureus* [4], *and Bacillus anthracis* [14]. Furthermore, several inherited human diseases, such as Huntington's disease [11] and Myotonic dystrophy [21], are now associated with the intragenic expansion of repeat DNA sequences [6, 17].

Nowadays, identifying VNTRs in many laboratories costs large amount of money and time. Biologists perform tandem repeat finding programs, such as REPuter [16], TRF [9] and so on[15,24], to search tandem repeats[3]. This approach follows that tandem repeats are directly used as VNTR candidates. For each VNTR candidate, they use PCR-based experiment to examine whether it is a true VNTR or not. Many VNTR candidates are then dropped from further analyses due to lack of variation in length. Using such a strategy, biologists need to design and synthesize primers for each possible tandem repeat. Moreover, the parameter settings deeply affect the number of tandem repeats predicted by many tandem repeat finding programs. The looser the parameter settings are, the more the tandem repeats are predicted.

Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04)
0-7695-2173-8/04 $ 20.00 © 2004 IEEE

Consequently, biologists get more candidates to verify which may be impossible to be real VNTRs. It costs a large amount of money and wastes unnecessary time.

In order to identify VNTR loci more efficiently, we develop a new tool called VNTR analyzer, a Variable Number of Tandem Repeat identification program. The program identifies VNTR candidates by comparing complete sequencing genomes of different strains of the same species. VNTR analyzer locates loci of each tandem repeat in test strains and counts copies of its repeated segments in each locus. Afterwards, the program outputs highly potential VNTR candidates and related VNTR information in tables. Moreover, it gives user a colorful chart displaying the optimal alignment between the maximal-copy locus and the minimal-copy one of each VNTR candidate. Such visualization provides further analysis for users to confirm the fitness of the outputted VNTR candidates.

For this paper, we applied the VNTR analyzer on 3 bacteria genomes of which some VNTR information have been verified by biological experiments: *Staphylococcus aureus*, *Xylella fastidiosa*, and *Salmonella enterica.* Our experiments showed that most of VNTR loci list in the literature could be successfully recognized by our VNTR analyzer**.** Few of them are missing in our prediction due to no changes of copy numbers in the present genomes. Besides, our results are not affected dramatically by the parameter settings.

The rest of this paper is organized as follows. Section 2 introduces the proposed approach. Section 3 presents the experiment results of predicted VNTRs in the 3 bacterial genomes. Section 4 contains conclusions and the future research directions.

2. SYSTEM AND METHOD

Identifying and analyzing VNTR by our program is a 4-step processes based on comparing the differences in sequences among several complete sequenced strains and visualized those differences. Figure 1 shows the work flowchart of the VNTR analyzer. Firstly, it looks for all possible tandem repeats in the first test strain. Next, for each possible tandem repeat, the program marks 2 primers to flank it**.** Using these primers, the program locates the corresponding loci of each possible tandem repeat in other tested strains. In the third, it calculates the number of copies in corresponding loci of each strain. After comparing the numbers of copies among all test strains, the program identifies those tandem repeats with various copies as VNTR candidates. Moreover, the VNTR

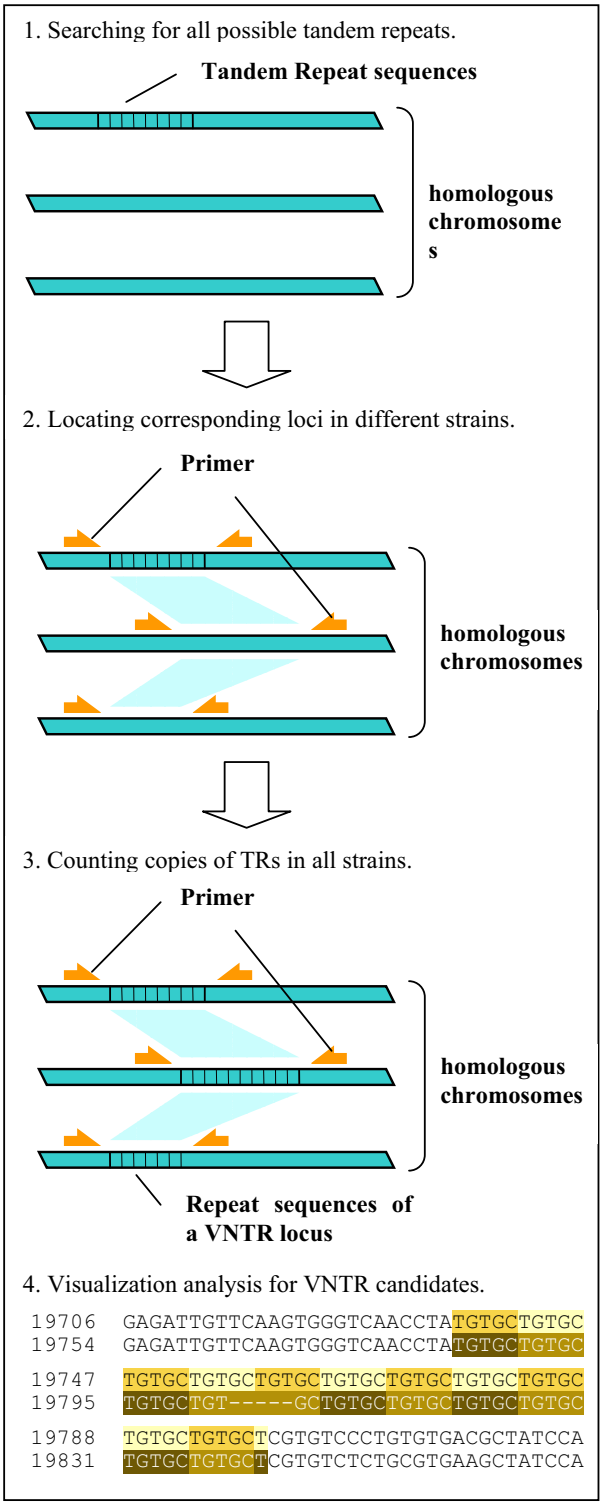

Figure 1. Flowchart of the VNTR analyzer.

analyzer outputs their sequences and maximal variation of copies into a table. Finally, for further analysis, we design an easy-to-use visualization tool, a colorful chart showing the optimal alignment of two loci with maximal and minimal copies for each VNTR candidate.

Step 1. Searching for all possible tandem repeats

To find out all possible tandem repeats in the first test strain, we use a commonly used program called TRF[2]. For details of its usage and parameter settings, please see "http://tandem.bu.edu/trf/trf.unix.help.html". The reported tandem repeat patterns and loci from TRF will be used in Step 2.

Step 2. Locating corresponding loci in different strains

For each repeat pattern and its locus found in Step 1, our program subsequently locates the corresponding loci of each tandem repeat in other test strains. First, we choose a 26bp-long sequence segment as a primer located on 50bp upstream of a tandem repeat in the first test strain. Then we find the most similar segments to the primer, synthesized from the first test strains, in other test strains by approximately matching with the primer. Meanwhile, we design another primer located on 50bp downstream of the tandem repeat, and find its most likely occurrences in other test strains on the same way. Flanked by the upstream and downstream primers, the corresponding loci of a tandem repeat are located in each test strain. This process is repeated to find out all corresponding loci for each tandem repeat.

In the above instructions, we regard the region between the upstream and downstream primers as a corresponding locus of a tandem repeat, and call it a possible locus in our program. To avoid getting the wrong result when the corresponding locus does not exist, we limit the difference in the length of possible corresponding loci among the first and other strains to be at most 8 times. If it is over 8 times, the region between the corresponding primers should not be a corresponding locus of a tandem repeat. Therefore, we extend the search region for 16bps on both tails, and then start the whole process from making primers again. Furthermore, the program will search the sequences in the reverse direction alternatively since the loci of some sequences may be reverse complemented. After checking by the above searching procedure for over 50 times, we would consider the tandem repeat does not exist in the strain, and drop the tandem repeat from VNTR candidates.

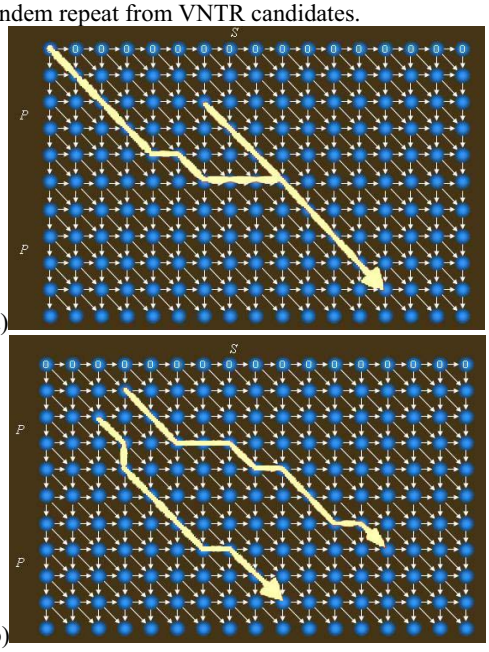

Figure 2. Two different paths with the same maximal score when counting copy number of a VNTR locus. Contrasting with (b), the two paths in (a) joined into the one at some point.

Step 3. Computing the variation of copies of tandem repeats in all strains

We now have the loci of tandem repeats in different strains. We use Wraparound dynamic programming (WDP)[18,25] to locally align the repeated segments in the corresponding loci of each tandem repeat. Then we count the number of copy of each locus, the different length rate (i.e. DLR) (Eq. 1) and the copy difference (i.e. CD) (Eq. 2) defined as follows. The one with copy difference more than some threshold are regarded as VNTR candidates.

Let a be the set of all possible loci of a tandem repeat; a_i is one locus in test strain i, and a_j is another one, where i =1,2,..,number of test strains; $|a_i|$ is the length of one locus in test strain i, where i =1,..,number of test strains.

Different Length Rate= $\frac{|a_i|-|a_j|}{|a_1|}$, where $|a_i| \geq |a_k|, |a_j| \leq |a_k|$, $\forall k$ = 1,2,..,number of test strains. …….(1)

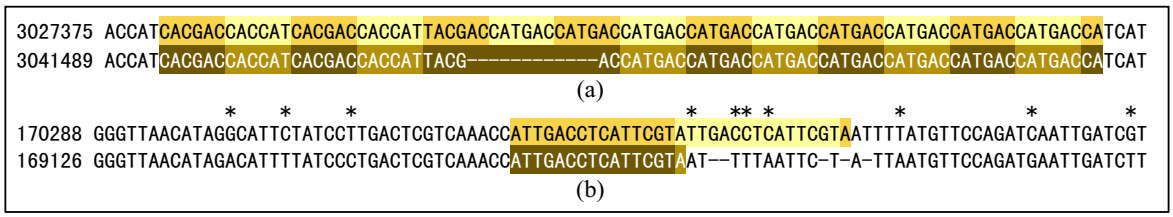

Figure 4. The visualization analysis of two possible loci of VNTRs. (a) Ty2 and CT18 VNTR loci in *yohM* gene of *Salmonella enterica*, Pattern=CACGAC, and (b) 9a5c and Temecula1 of *Xylella fastidiosa*, Pattern=ATTGACCTCATTCGT. Symbol "*" denotes mismatch between sequences.

Let b be the maximal number of copies in all test strains of a tandem repeat, and C is the minimal number of copies.

Copy Difference = $b - c$ …….(2)

Finally we output VNTR information into a table, including locations of VNTR candidates in the first strain, lengths of the regions between the upstream and downstream primers in each test strains, different length rate, copies of each strain, and copy difference. Besides, we also output gene information of each VNTR locus. The gene information comes from NCBI GenBank.

Step 4. Visualization tool for VNTR candidate

When we use local alignment method to align repeated segments and a possible locus for counting copy numbers, there are two or more paths with the same maximal score in the alignment graph occasionally (Figure 2a or Figure 2b). Therefore, one sequence would have two or more different copy numbers in this way. When we recognize VNTRs only by comparing the copy numbers of different strains, the confidence of the truthfulness of VNTRs would be affected by the uncertainty of copy numbers. To strengthen our prediction power, we develop a novel visualization tool to observe the true sequence changes.

We use the following example to show how our visualization works. Here, there are 2 strains of the same VNTR locus. One is with 7 copies (Figure3a), and the other is with 6 copies (Figure 3b). After aligning these sequences, we can see a segment of gaps in their alignment result (Figure 3c). To easily observe such repeat changes between 2 stains, we color the information of repeated units with two different colors taking places by turns (Figure 3d upper panel). Also there exists another optimal alignment that obtains the same maximal score as we mentioned above (Figure 3d lower panel). Figure 3d is the visualization of VNTR information in this example.

In general cases, we choose 2 loci corresponding to the same VNTR to do the end-space-free alignment. One is the sequence of the maximal copies with proper extension in both tails, and the other is of the minimal copies with the extension. Figure 4a shows the visualization of a real VNTR identified in *yohM* gene of *Salmonella enterica*, and Figure 4b is the result of a false VNTR in alignment of segments in 9a5c and Temecula1 of *Xylella fastidiosa*. Although the copies and length of loci in Figure 4b are totally different, it is not a VNTR since there are no continuous gaps of length approximate as a pattern in our visualization analysis. By using this visualization tool, not only users can decide whether the candidate is a true VNTR in a short time, but biologists can get hints to do further research through looking at the sequence changes.

3. RESULT AND DISCUSSION

In this study, we used our program to analyze 3 species, *Staphylococcus aureus, Xylella fastidiosa, and Salmonella enterica.* Their VNTR loci had been proved and used in present DNA fingerprinting. In the experiments, the parameter values used in TRF are p_M = 75, p_I = 20, Minscore = 50, and MaxPeriod = 2000. The experiment results demonstrate that our program can find VNTRs accurately in a short time. From Table 1, we can see that the average running time of our program on these 3 species is some ten minutes (on the machine with *Intel P4* 2.8G). As the genome size of the interested specie grows, the running time is getting longer expectably. Besides, the number of primers tried to properly flank a possible locus of VNTR is another important factor to affect the running time. Although the genome size of *Xylella fastidiosa* is a little bit smaller than that of *Staphylococcus aureus*, the number of primers we tried in *Xylella fastidiosa* is almost twice as big as the one in *Staphylococcus aureus*. Thus, the running time for finding VNTRs in *Xylella fastidiosa* is longer. Moreover, Table 1 summarizes the numbers of VNTRs candidates in these 3 species at each analyzing stage and their consistency with the ones in the literature. We discuss the details of the predicted VNTRs in the following.

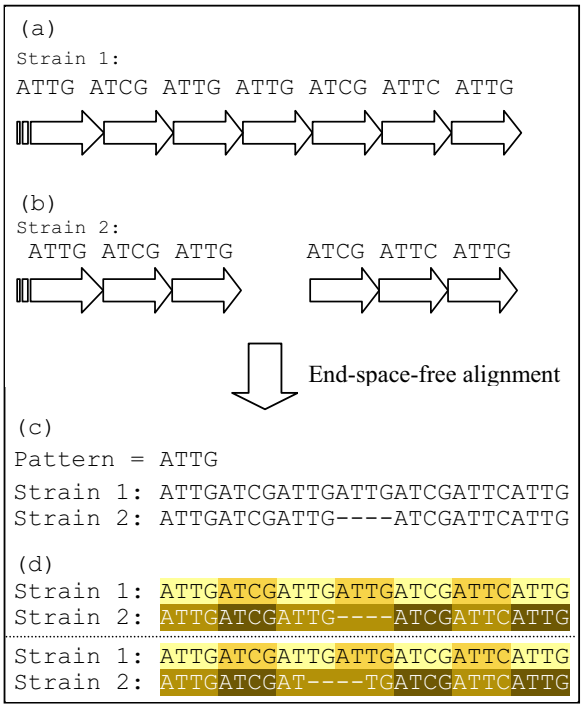

Figure 3. Alignment of VNTR loci with different copies and color repeat units. (a) VNTR locus with 7 copies; (b) VNTR locus with 6 copies; (c) Alignment of two loci; (d) Two alignments with the same maximal scores, and shows the visualization tool of our program.

Staphylococcus aureus

There are 3 completely sequencing strains of *Staphylococcus aureus*, *Staphylococcus aureus* N315, Mu50, and MW2. The sequences were downloaded from the web site of the National Center for Biotechnology Information (NCBI). We used N315 as the first test strain to search for tandem repeats, and then compared copy numbers among possible loci of other two strains. The parameter values using in TRF are Match = 2, Mismatch = 7, Delta = 7. WDP and end-space-free alignment used the same parameter settings as TRF did.

The duplicated and multi-reported tandem repeats in the result of TRF were excluded by our program. Thus, there are 76 possible tandem repeats in *Staphylococcus aureus*. Our program discovered 41 VNTR candidates as their copy numbers and length vary in these 3 strains. This variation of copies is over 0.5 when we check the result table in copies comparison between strains. 33 of them remain after further visualization analysis. Therefore, we discovered more than one missing repeat. So as we know, these are very powerful potential VNTR candidates. Of the 33 VNTR candidates we predicted, sdrC, SdrD, sdrE, clfA, clfB, ssp, coa, and spa have been verified as real VNTRs in the literature [4].

In multiple-locus VNTR analysis for the molecular typing of clinical isolates with several bacterial species, it is important to clearly differentiate among unrelated isolates. But at the same time, one should also demonstrate the relationship of all organisms isolated from individuals infected through the same source. Therefore, the more proper VNTR loci we use, the more discriminatory power we get. For getting more VNTR loci, we loosed the parameters used in TRF, WDP, and colorful alignment to be Match = 2, Mismatch=5, Delta=5. In Table 1, the number of tandem repeats in such parameter setting increased to 212, and the number of VNTR candidates after Step 3 was 50. Also the number of VNTR candidates remaining after visualization analysis increased to 41. Similarly, by setting the parameters values as Match=2, Mismatch=4, Delta=4, the result rose to 659, 72 and 52. Furthermore, we can see that the number of tandem repeats predicted by TRF changed more dramatically as the parameters changed than the number of VNTRs recognized by our programs. It is quite reasonable but also emphasizes the risk of directly using tandem repeats as VNTR candidates without further analysis. Table 2 gives the loci information of VNTR candidates in *Staphylococcus aureus.*

Xylella fastidiosa

The complete genome sequences of 2 *Xylella fastidiosa* strains 9a5c and Temecula1 were also obtained from NCBI web site. We use 9a5c as the first test strain in our program. The parameter values using in TRF, WDP and colorful alignment are also Match=2, Mismatch=7, Delta=7. The reported number of possible tandem repeats and VNTR candidates are 112 and 68, respectively. 60 VNTR candidates remains after visualization analysis. Most of them are short in length of 3bp to 33bp. SSR20, SSR28, SSR30, and SSR 40 in the literature[10] were identified in our program. Although SSR26 and SSR36 were not regarded as VNTRs in previous studies, SSR26 shows enormous change in number of copies and length in our analysis. SSR26 has 37 copies in 9a5c, but only 1.62 copies in Temecula. In SSR36 locus, the 9a5c has only 10.5 copies, but Temecula has 31.37 copies. So we suggest that these 2 loci may worth further investigation.

Table 1. Running time and VNTR candidates in 3 species. *S. aureus* uses 3 test parameter settings of Match, Mismatch and Delta, (2, 7, 7), (2, 5, 5) and (2, 4, 4). Factors affecting the running time are marked with symbol "*". "#" indicated the number of VNTR candidates without duplicated and multi-report in 3 analysis stages. The row "Occurring in Known VNTRs" is the number of predicted VNTR loci shown in the literature. For example, "7 of 8" means that there are 8 known VNTRs in the literature, and 7 of them are accurately predicted by our program.

Specie	*S. aureus* (2, 7, 7)	*S. aureus* (2, 5, 5)	*S. aureus* (2, 4, 4)	*Xylella fastidiosa*	*Salmonella enterica*
***Genome Size**	$3*10^6$			$2.6*10^6$	$4.9*10^6$
***Strain used**	3			2	3
***#TR**	76	212	659	112	52
***Primer tried**	962	2269	6648	1663	1792
Running Time	11m	23m	67m	17m	32m
#VNTRs after comparing	41	50	72	68	23
#VNTRs after visualization	33	41	52	60	18
Occurring in Known VNTRs	7 of 8			4 of 5	6 of 7

Salmonella enterica

Salmonella enterica ser. Typhi strains TY2 and CT18 and *Salmonella enterica* ser. Typhimurium strain LT2 SGSC1412 were completely sequenced and the sequences can be downloaded from NCBI website. LT2 was used as the first test strain, and the same parameter setting as above experiments was taken. TRF reported 52 tandem repeats, and 23 of them remained as VNTR candidates after comparing copies among different strains. After visualization analysis, we predicted 18 potential VNTR candidates. 8 VNTR candidate loci will be cloned by previous study from STTR1 to STTR8. Our prediction of STTR1, 3, 4, 5, 7, and 8 loci as VNTRs are as well as previous studies [22]. STTR2 locus has no variation in length and copies both in our result and the literature [22]. The one we miss is STTR6 [22]. It is difficult to search for its similar loci between TY2 and CT18 since the strains we used may not be infected with phage.

4. CONCLUSION

We have demonstrated that our program is an efficient and automatic approach for finding VNTRs accurately. With only 2 or 3 completely sequencing strains we conducted in the experiments, the program found the known VNTR loci and predicted new ones that worth further study. VNTR analyzer not only predicts VNTRs, but also provides researchers a colorful visualization tool to investigate the possible influence of sequence changes.

In the future, we proceed to find enough and proper VNTR loci of some bacteria which are not familiar with in present molecular typing technology and epidemiology. With the support from Center for Disease Control Taiwan, we are working on *Streptococcus pyogenes* and *Shigella flexneri*. Although *Streptococcus pyogenes* have 42 tandem repeats with Match=2 Mismatch=7 Delta=7, there is only one of them with more than 10 copies. The more the copies a tandem repeat has, the larger the possibility it will be a VNTR. The phenomena of tandem repeats in *Streptococcus pyogenes* implies that we cannot get VNTR candidates just by comparing the type of repeats, homogeneity, and number of copies. So, we adjusted the parameter to Match=2 Mismatch=4 Delta=4, and got 267 possible tandem repeats. But it caused another problem -- the level of homogeneity is getting lower as the tandem repeats is getting more. To reduce the cost and the required time in real biological experiments, the capacity of our program will be shown. We will synthesize real primers to verify our prediction of VNTRs in these 2 bacterial genomes. Furthermore, we will develop online tools and databases of VNTRs to help biologists for querying VNTRs of their interests with friendly graphic interfaces.

Table 2. Report of highly potential VNTR candidates in *Staphylococcus aureus.* The columns from left to right respectively are "location of tandem repeat locus in first strain"," Possible locus in each strains and difference length rate", "copy number in each strains and copy difference", "pattern size of tandem repeats" and "Gene information". Note that if we use different parameters in WDP and TRF ex. Match = 2 Mismatch =4 Delta =4 to TRF and Match = 2 Mismatch =3 Delta =3 to WDP , then copy number × pattern size ≠ length of primer loci. By adjusting parameters to (2, 4, 4), we got many additional VNTR candidates located in ORFs, and 2 of them are named as (mapW) & fmtB(mrp) marked with symbol "*".

Indices	Possible locus				Copy number				Pattern	Gene name
Start--End	1	2	3	DLR	1	2	3	CD	Size	
43150--43507	357	277	317	0.22	8.94	6.94	7.94	2	40	
122940--123176	236	188	236	0.2	9.87	8.16	9.87	1.7	24	spa
266167--266619	452	371	452	0.17	5.59	4.48	5.59	1.11	81	coa
311526--311693	167	224	167	0.34	3.05	1.41	3.05	1.63	55	
343171--345674	2503	3998	2503	0.59	4.9	7.32	4.38	2.94	511	-
444579--444658	79	151	0	0.91	2.22	1.22	0	1	36	set9
485155--485543	388	388	187	0.51	1.93	1.93	0.93	1	201	SA0423
526783--526833	50	29	92	1.26	2.42	1.42	4.42	2.99	21	
607366--607893	527	533	527	0.01	30.33	27.27	30.33	3.05	18	sdrC
611812--612416	604	490	604	0.18	33.27	27.33	33.27	5.94	18	sdrD
680338--680544	206	472	206	1.29	3.1	7.1	3.1	4	66	
777130--777265	135	0	79	0.41	2.41	0	1.41	1	56	
818071--818207	136	80	136	0.41	2.47	1.47	2.47	1	55	
823565--823872	307	127	307	0.58	5.25	1.71	5.25	3.54	59	
850242--851179	937	811	775	0.17	52.11	49.22	43.11	9	18	clfA
874296--874420	124	68	124	0.45	2.25	1.25	2.25	1	56	
888859--889113	254	40	254	0.84	5.95	0.95	5.95	5	43	
1021696--1021796	100	55	100	0.45	5.61	2.11	5.61	3.5	18	sspA
1056336--1056721	385	70	385	0.81	6.12	1.12	6.12	5	63	
1117862--1118187	325	128	325	0.6	2.51	0.91	2.51	1.6	128	
1142714--1142835	121	188	121	0.55	2.17	3.17	2.17	1	56	-
1215617--1215826	209	145	209	0.3	3.28	1.18	3.28	2.09	64	
1439902--1440593	691	460	691	0.33	2.99	2.09	2.99	0.9	231	ebhA
1653005--1653197	192	77	192	0.59	3.33	1.26	3.33	2.07	57	
2076113--2076165	52	64	52	0.23	3.77	4.44	1.55	2.88	27	-
2145378--2145682	304	105	304	0.65	3.04	1.1	3.04	1.94	100	
2281793--2281861	68	20	68	0.7	2.76	0.76	2.76	2	26	-
2440666--2440834	168	60	168	0.64	2.77	0.77	2.77	2	61	
2560410--2562528	2118	2110	1351	0.36	5.51	0.1	3.51	5.41	383	-
2572160--2572395	235	151	235	0.35	5.61	3.61	5.61	2	42	fnb
2634837--2635124	287	98	287	0.65	2.26	1.26	2.26	1	127	
2718517--2719170	653	719	653	0.1	36.33	41.33	36.33	5	18	clfB
2813887--2813924	37	19	37	0.48	2.11	1.11	2.11	1	18	
*2003357--2004612	1255	1570	1255	0.25	3.89	4.89	3.89	1	321	truncated(mapW)
*2225145--2225284	139	160	139	0.15	7.61	6.61	7.61	1	18	fmtB(mrp)

Reference

[1] A.J. Jeffreys, V. Wilson, and S.L. Thein, **"**Hypervariable 'minisatellite' regions in human DNA", *Nature*, 314, 1985, pp. 67-73.

[2] A. Johansson, I. Goransson, P. Larsson, and A. Sjostedt, **"**Extensive allelic variation among *Francisella tularensis* strains in a short-sequence tandem repeat region", *J. Clin. Microbiol*., 39, 2001, pp. 3140-3146.

[3] A.M. Klevytska, L.B. Price, J.M. Schupp, P.L. Worsham, J. Wong, and P. Keim, "Identification and Characterization of Variable-Number Tandem Repeats in the *Yersinia pestis* Genome", *J. Clin. Microbiol.*, 39, 2001, pp. 3179-3185.

[4] A. Sabat, J. Krzyszton-Russjan, W. Strzalka, R. Filipek, K. Kosowska, W. Hryniewicz, J. Travis, and J. Potempa, "New method for typing *Staphylococcus aureus* strains: multiple-locus variable-number tandem repeat analysis of polymorphism and genetic relationships of clinical isolates", *J. Clin. Microbiol*., 41(4), 2003, pp. 1801-4.

[5] B.S. Weir, "Population genetics in the forensic DNA debate", *Proc Natl Acad Sci U S A*., 89(24), 1992, pp. 11654-9. Review.

[6] C.T. Caskey, A. Pizzuti, YH Fu, Jr R.G. Fenwick, and D.L. Nelson, "Triplet repeat mutations in human disease", *Science*, 256, 1992, pp. 784-789.

[7] D.M. Adair, P.L. Worsham, K.K. Hill, A.M. Klevytska, P.J. Jackson, A.M. Friedlander, and P. Keim, "Diversity in a variable-number tandem repeat from *Yersinia pestis*", *J. Clin. Microbiol.*, 38, 2000, pp. 1516-1519.

[8] E. Mazars, S. Lesjean, A.L. Banuls, M. Gilbert, V. Vincent, B. Gicquel, M. Tibayrenc, C. Locht, and P. Supply, "Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome", *Mol. Microbiol*., 36, 2000, pp. 762-771.

[9] G. Benson, "Tandem repeats finder: a program to analyzed DNA sequences", *Nucleic Acids Res.*, 27, 1999, pp. 573–580.

[10] H.D. Coletta-Filho, M.A. Takita, A.A. de Souza, C.I. Aguilar-Vildoso, and M.A. Machado, "Differentiation of strains of *Xylella fastidiosa* by a variable number of tandem repeat analysis", *Appl. Environ. Microbiol.*, 67, 2001, pp. 4091-4095.

[11] Huntington's Disease Collaborative Research Group, "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes", *Cell*, 72, 1993, pp. 971-983.

[12] M. Strand, T.A. Prolla, R.M. Liskay, and T.D. Petes, "Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair", *Nature*, 365(6443), 1993, pp. 274-6.

[13] P.E. Smouse, and C. Chevillon, "Analytical aspects of population-specific DNA fingerprinting for individuals", *J. Hered*., 89(2), 1998, pp. 143-50.

[14] P. Keim, L.B. Price, A.M. Klevytska, K.L. Smith, J.M. Schupp, R. Okinaka, P.J. Jackson, and M.E. Hugh-Jones, "Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*", *J. Bacteriol*., 182, 2000, pp. 2928-2936.

[15] R. Kolpakov, G. Bana, and G. Kucherov, "mreps: efficient and flexible detection of tandem repeats in DNA", *Nucleic Acids Res.*, 31, 2003, pp. 3672-3678.

[16] S. Kurtz, and C. Schleiermacher, "REPuter: fast computation of maximal repeats in complete genomes", *Bioinformatics*, 15, 1999, pp. 426–427.

[17] S. Mohmood, A. Sherwani, F. Khan, R.H. Khan, and A. Azfer, "DNA trinucleotide repeat expansion in neuropsychiatric patients", *Med Sci Monit*., 9(9), 2003, RA237-45. Review.

[18] V. Fischetti, G. Landau, J. Schmidt, and P. Sellers, In A. Apostolico, M. Crochemore, Z. Galil, and U Manber (eds), "Identifying periodic occurrences of a template with applications to a protein structure", *Proceedings of the 3rd Annual Symposium on Combinatorial Pattern Matching*, *Lecture Notes in Computer Science*, 644, Springer-Verlag, Berlin, 1992, pp. 111–120.

[19] van A. Belkum, S. Scherer, van L. Alphen, and H. Verbrugh, "Short sequence DNA repeats in prokaryotic genomes", *Microbiol. Mol. Biol*., 62, 1998, pp. 275-293.

[20] van A. Belkum, W.J. Melchers, C. Ijsseldijk, L. Nohlmans, H. Verbrugh, and J.F. Meis, "Outbreak of amoxicillin-resistant *Haemophilus influenzae* type b: variable number of tandem repeats as novel molecular markers", *J. Clin. Microbiol.*, 35, 1997, pp. 1517-1520.

[21] Y.-H. Fu, A. Pizzuti, R.G. Fenwick, J. King, S. Rajnarayan, P.W. Dunne, J. Dubel, G.A. Nasser, T. Ashizawa, P. de Jong, B. Wieringa, R. Korneluk, M.B. Perryman, H.F. Epstein, and C.T. Caskey, "An unstable triplet repeat in a gene related to myotonic muscular dystrophy", *Science*, 255, 1992, pp. 1256-1258.

[22] Y. Liu, M.A. Lee, E.E. Ooi, Y. Mavis, A.L. Tan, and H.H. Quek, "Molecular typing of *Salmonella enterica* serovar typhi isolates from various countries in Asia by a multiplex PCR assay on variable-number tandem repeats", *J. Clin. Microbiol.*, 41(9), 2003, pp. 4388-94.

[23] Y. Nakamura, M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin, and et al., "Variable number of tandem repeat (VNTR) markers for human gene mapping", *Science*, 235(4796), 1987, pp. 1616-22.

[24] Y. Wexler, Z. Yakhini, Y. Kashi, and D. Geiger, "Finding Approximate Tandem Repeats in Genomic Sequences", *Research in Computational Molecular Biology* 2004.

[25] W. Miller, and E. Myers, "Approximate matching of regular expression", *Bull. Math. Biol*, 51, 1989, pp. 5-37.