Malays. Appl. Biol. (2016) 45(2): 75–85

THE PROFILING OF SHORT TANDEM REPEATS (STRs) FROM NEXT-GENERATION SEQUENCING (NGS) DATA BY ESTABLISHING A WHOLE GENOME STRs PIPELINE FOR FORENSIC BIOINFORMATICS

NUR NABILAH, A.[1*] and VENKATARAMANAN, S.[1]

[1]*Department of Diagnostic and Allied Health Science, Faculty Health and Life Sciences, Management and Science University, 40100, Shah Alam, Selangor, Malaysia*
**Email: nurnabilahalias@yahoo.com*

Accepted 26 October 2016, Published online 21 December 2016

ABSTRACT

Repeat region with length of one to six base pairs (bp), found in DNA sequences is known as short tandem repeats (STRs). Currently, high-throughput next-generation sequencers have facilitated the effective polymorphic STR markers' identification. In this study, a new whole genome STRs pipeline has been established in order to call and profile STRs from next-generation sequencing (NGS) data. Firstly, genome sequences of *Helicobater pylori* strain, CPY1124 and PeCan4 as reference genome were retrieved from European Nucleotide Archive (ENA) database which then the quality of sequences were checked using FastQC. The assembly of genome sequences was done by VELVET *de novo* assembler. Unordered contigs from VELVET's output was realigned using multiple genome alignment (MAUVE) to obtain ordered contigs sequence. Lastly, STRs calling and profiling by Tandem Repeat Finder was done with the parameters of (2: match, 7: mismatch and 7: indels). These parameters are for Smith-Waterman style local alignment using wrap-around dynamic programming. As a result, this new pipeline enables to identify polymorphic and unique STRs which are GTTTG and AAACCC from CPY1124. This pipeline has been compared with other available STRs profiling pipeline like pSTR Finder and Tandem Repeat Database (TRDB) for validation purpose. The similar output producing by both tools thus indicates the reliability of this new pipeline for future usage.

Key words: Short tandem repeats (STRs), next-generation sequencing data (NGS), *Helicobacter pylori, H. pylori*, whole genome and STRs pipeline

INTRODUCTION

Short Tandem Repeats (STRs)

Short tandem repeats (STRs), also known as microsatellites or simple sequence repeats (SSRs) is a molecular marker consisting of one to six bases repeated for 5 to 50 times (Hartwell *et al*., 2008; Walker & Rapley, 2005; Weising *et al*., 2005). As for example (ACCC)n where n implies number of ACCC repeats. They are mostly found in non-coding region of the chromosomes. Furthermore, they can be efficiently analyzed by polymerase chain reactions (PCR), besides being co-dominant (Jayabalan, 2006), highly reproducible and multi-allelic, and capable of being automated. In fact, they become important in genes identification (Zhao *et al*., 2013) and genetic analysis of population.

STRs have several variants with many mutations. According to Ballantyne *et al* (2010), the spontaneous mutation rate of STRs is 3.78×10^{-4} to 7.44×10^{-2} in the human Y-chromosome, which is greater than the rate of copy number variation (CNV), 1.7×10^{-6} to 1.2×10^{4} (Lupski, 2007). This greater mutation rate of STRs becomes a major influence towards their variation. Therefore, detecting various STRs by processing billions of short, raw reads is essential for personal genomes analysis.

Almost 17% of human genes containing STRs in their open reading frames (ORFs) (Gemayal *et al.,* 2010). Different people will have different number of these repeats and they are unique found in everyone. Due to higher number of STRs in human genome and their highly polymorphic characteristic, this repeat region will act as effective profiling marker (Butler *et al*., 2007) and human

* To whom correspondence should be addressed.

identification marker for evolutionary and forensic analysis.

In the 1990s, STR analysis became popular in forensics field which it started to be used for individual genetic fingerprinting. Tetra- or penta-nucleotide repeats are being used in forensic analysis as they give a high degree of error-free data while being robust enough to survive degradation in non-ideal conditions. Unlike shorter repeat sequences like mono-, di- and tri-nucleotides, they tend to suffer from PCR stutter and preferential amplification as well as the fact that several genetic diseases are associated with tri-nucleotide repeats such as Huntington's disease.

Next-generation Sequencing (NGS)

In recent years, there has been emergence of a new technology, called the next-generation sequencing (NGS) technique which increases the sequencing's capability far beyond the traditional Sanger method. Up to present, the entire human genome can be sequenced within several hours and days by using NGS machine (Illumina, 2012). Its benefits such as high-throughput, lower in cost, faster in analysing time and highly-scalable indirectly leads to many new projects in sequencing genomes of thousands of individual humans, animal and plant. Additionally, this technology could significantly faster the analysis of mixed DNA samples and complicated paternity cases.

Some STR-containing reads in whole genome that cannot be mapped completely to reference genome due to higher number of mismatch or indels related to STRs have been solved by the advanced technology of high-throughput next-generation sequencers like Illumina HiSeq 4000. It is able to facilitate the identification of polymorphic STR markers and can output many short reads that may be meaningful for uncovering STR-related disease.

Apart from that, although STRs is important in evolution and disease discovery, it is very challenging to perform accurate genotyping of STRs from NGS data (Treangen & Salzberg, 2011). In fact, there is still less exploration in analysing short read of STRs. Therefore, the identification of polymorphic STRs from NGS data may be helpful in detecting significant genetic markers while searching for conserved STRs can be functional elements in gene regulation.

MATERIALS AND METHODS

Retrieving Whole Genome Sequences

To demonstrate whether the proposed pipeline is capable in detecting polymorphic and unique STR from NGS data, *Helicobacter pylori* strains, CPY1124 and PeCan4 in fastq format were downloaded from the European Nucleotide Archive (ENA) database as benchmark datasets. The accession numbers of CPY1124 and PeCan4 strains were SRR400671 and ERR351243 respectively. This PeCan4 can also be retrieved under the accession number of CP002074 and NC_014555 in DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL) or GenBank databases. In this analysis, PeCan4 was chosen as the reference genome which will be used in alignment step by multiple genome alignment (MAUVE).

FastQC Quality Control Checking

H. pylori strain, CPY1124 undergoes FastQC checking (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) version 0.9.1 in Linux server to check if there were any problems or biases related to this sequence. A few criteria have been evaluated, for instances per base sequence quality, per base GC content, duplicate sequences, over represented *k-mers* and many more. However, in this study, per base sequence quality and duplicate sequences will be focused more as the crucial criteria to be fulfilled for the next assembly step.

Assembly of Short Reads Sequence

Whole genome sequence of *H. pylori* strains of CPY1124 that consists of two files, referring to reverse and forward sequences files, SRR400671_1.fastq.gz and SRR400671_2.fastq.gz respectively were assembled by using Velvet *de novo* assembler. Velvet assembler was able to read sequences and remove overlapping sequence reads in order to produce high quality of contigs sequences. A few parameters such as expected coverage, *k-mer* length, minimum coverage, cut-off value and length of minimum contigs were determined to build the *de Bruijn* graph. The length of *k-mer*, expected coverage value and cut-off value were set as 31, 21 and 2.81 respectively. Then, 200bp was set for minimum contigs length as it was the minimum length of genome sequence to be submitted into GenBank. At the end of this assembly step, contigs file, named as contigs.fa was expected to be produced which contains unordered contigs sequences (assembled transcript).

Reordering Unordered Contigs

The contigs.fa file that contains unordered contigs sequences of CPY1124 was used as the input in Multiple Genome Alignment (MAUVE). This unordered contigs sequence was compared with the reference genome sequence, PeCan4 to reorder the contigs sequences, thus producing an ordered sequence of it. The reordering process may take half an hour, from four to seven iterations or even more than that depending on type of operating system.

In this study, MAUVE is run under 64-bit Window OS. As for its output, it produced a set of ordered and oriented contigs in Fasta format, located in the last of the iterated alignment. Then, ordered contigs can be viewed by aligning with progressive Mauve. By using Mauve, these two complete genome alignments can be analyzed on their similarity degree and conserved regions between them (Darling *et al.*, 2004).

Short Tandem Repeats (STRs) Calling

Tandem Repeat Finder was used to call all the possible STRs in ordered sequence of CPY1124 and the alignment parameters were set to (2, 7, 7), referring to match, mismatch and indels respectively. Minimum size of patterns was set to 10bp and the minimum alignment score to report repeats is 50. Two files were returned after the execution of this tandem repeat finder. The first file was summary table, describing the location and statistical properties of the tandem repeats found and another file was alignment file, containing alignment of each repeat with its consensus sequence. Those files were connected, thus by clicking any indices in the table will open second browser window that contains detailed explanation of the selected alignment.

There were several information to be focused in summary table such as repeat's indices in sequence, period size, copy number, size of consensus pattern, percent matches between adjacent copies, percent indels between adjacent copies, alignment score, percent composition for each nucleotides (A, C, G, T) and entropy measure based on percent composition.

RESULTS

Quality evaluation for the genome sequence of *H. pylori* strain, CPY1124 was done using FastQC to determine whether it was of good quality or not. Both of the reverse and forward sequences of CPY1124 showed good quality criteria based on higher quality scores they obtained as almost all the bases were on green background of Box Whisker plot (Figure 1 and 2). However, there were still duplications resulted in CPY1124 strain (Figure 3 and 4).

As the final result for assembly step by Velvet *de novo* assembler, unordered contigs sequence of CPY1124 has been produced, therefore it was important for it to be realigned by MAUVE (Edwards *et al.,* 2013). Based on Figure 5, the conserved and unique regions of *H. pylori* strains CPY1124 can be observed when this strain was compared with PeCan4 strain. There were identical conserved regions being observed between both strains which indicated by the same colour block between them. Besides, the presence of less number of white gaps also proved that they have high similarities in term of homology characteristic.

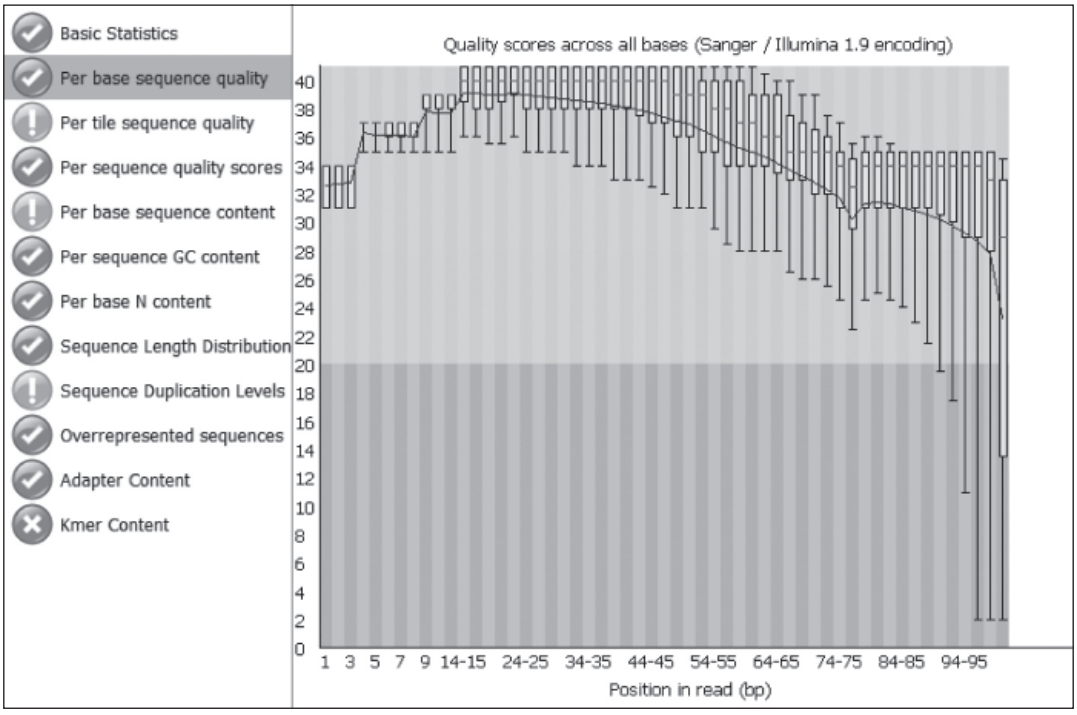

Fig. 1. Per Base Sequence Quality of SRR400671_1.fastq.gz

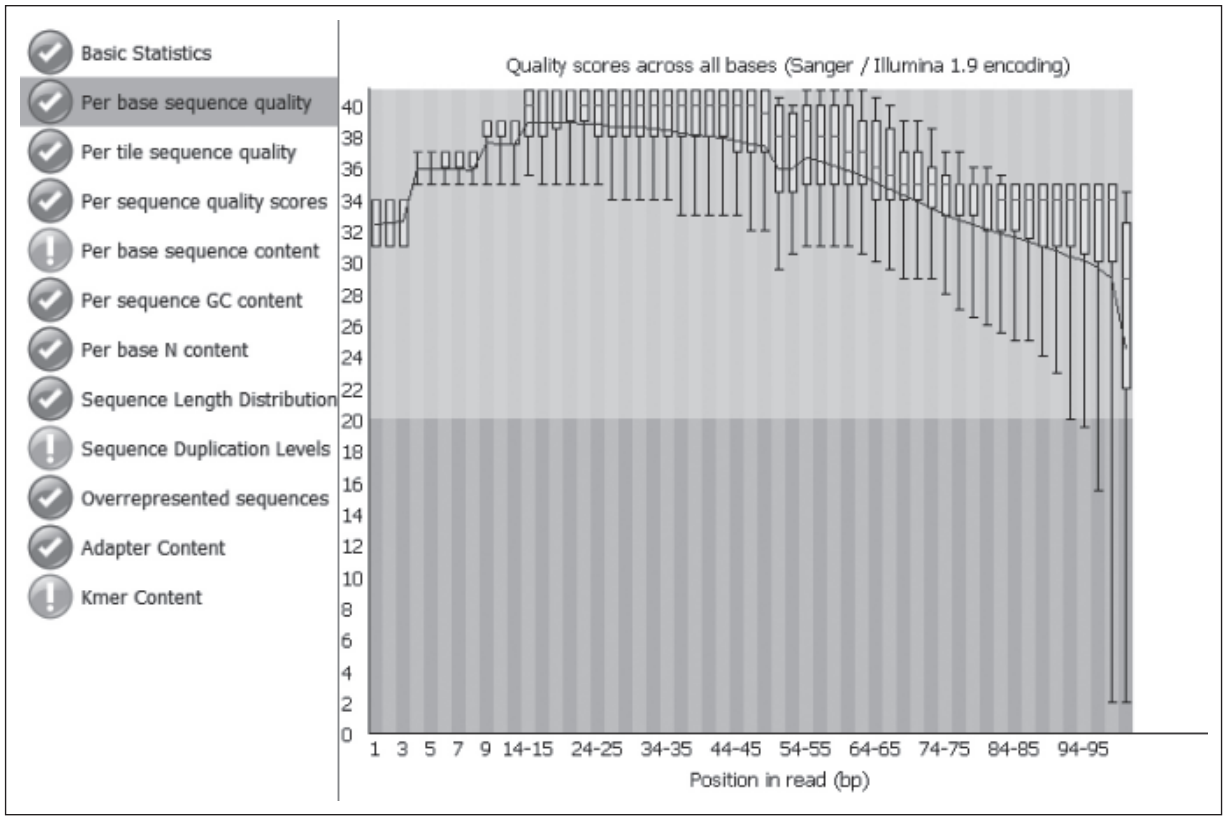

Fig. 2. Per Base Sequence Quality of SRR400671_2.fastq.gz

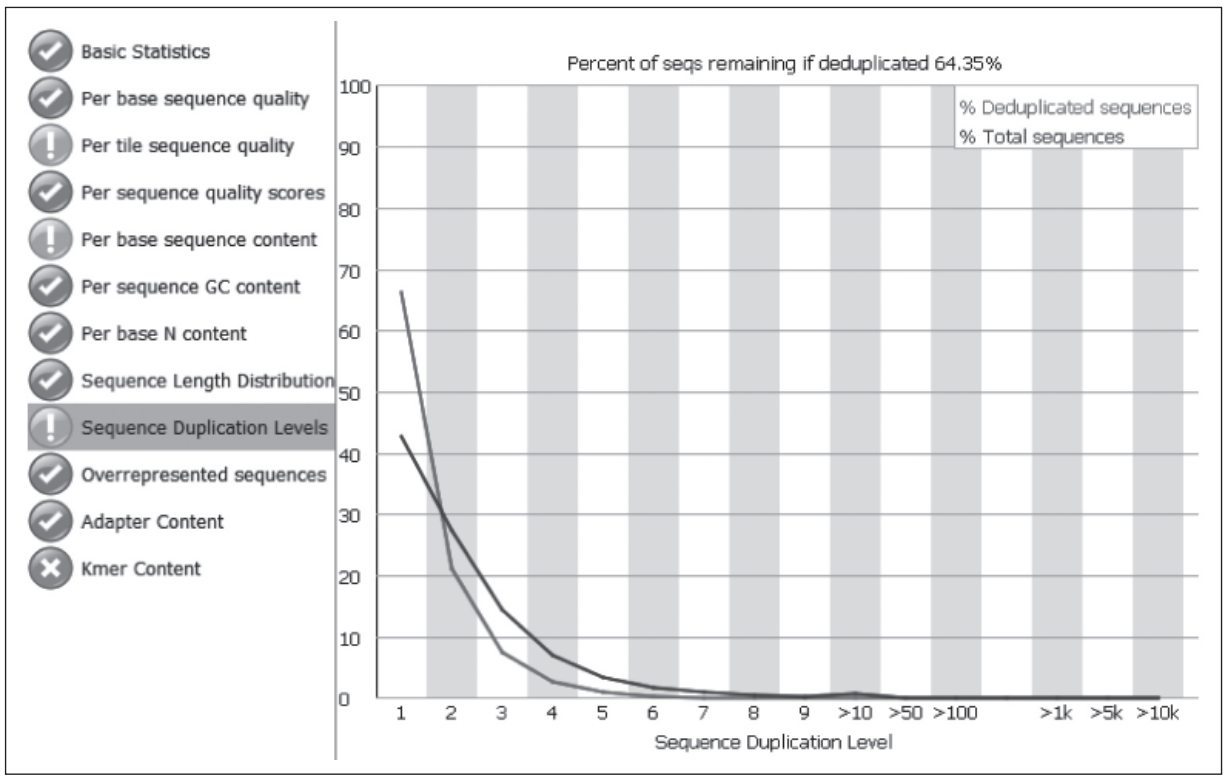

Fig. 3. Sequence Duplication Level of SRR400671_1.fastq.gz

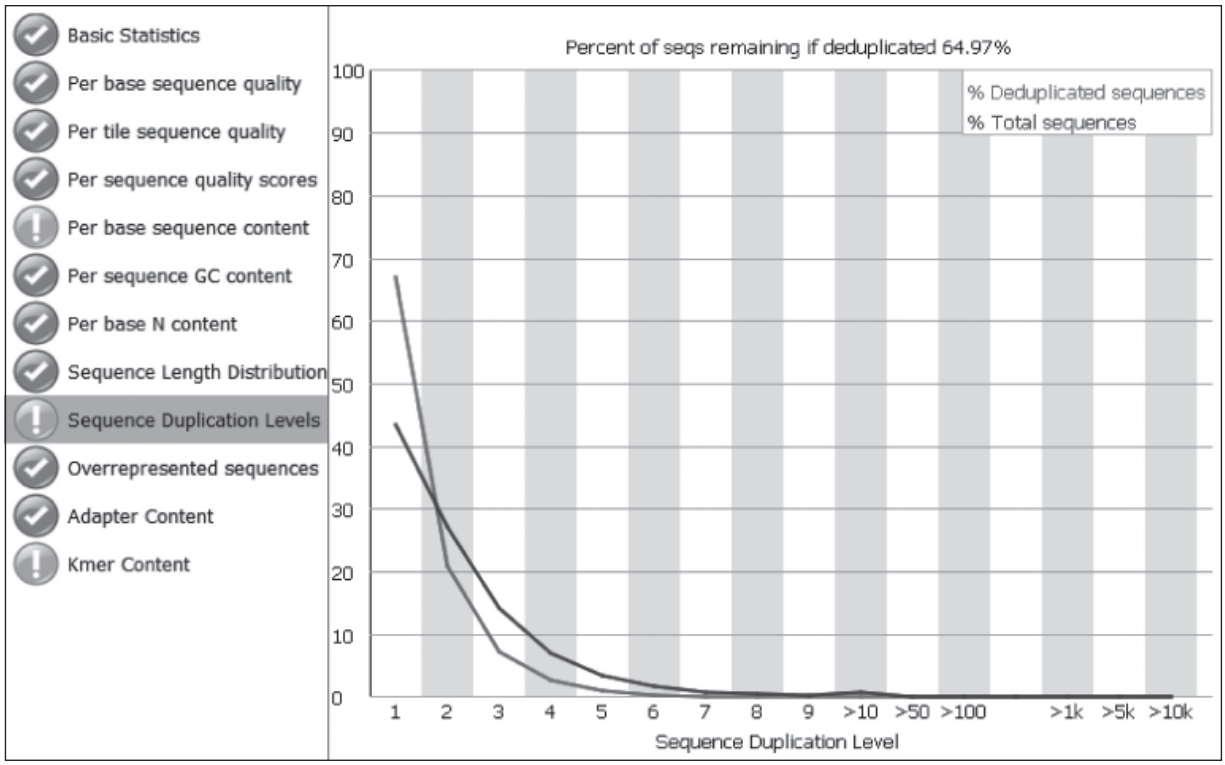

Fig. 4. Sequence Duplication Level of SRR400671_2.fastq.gz

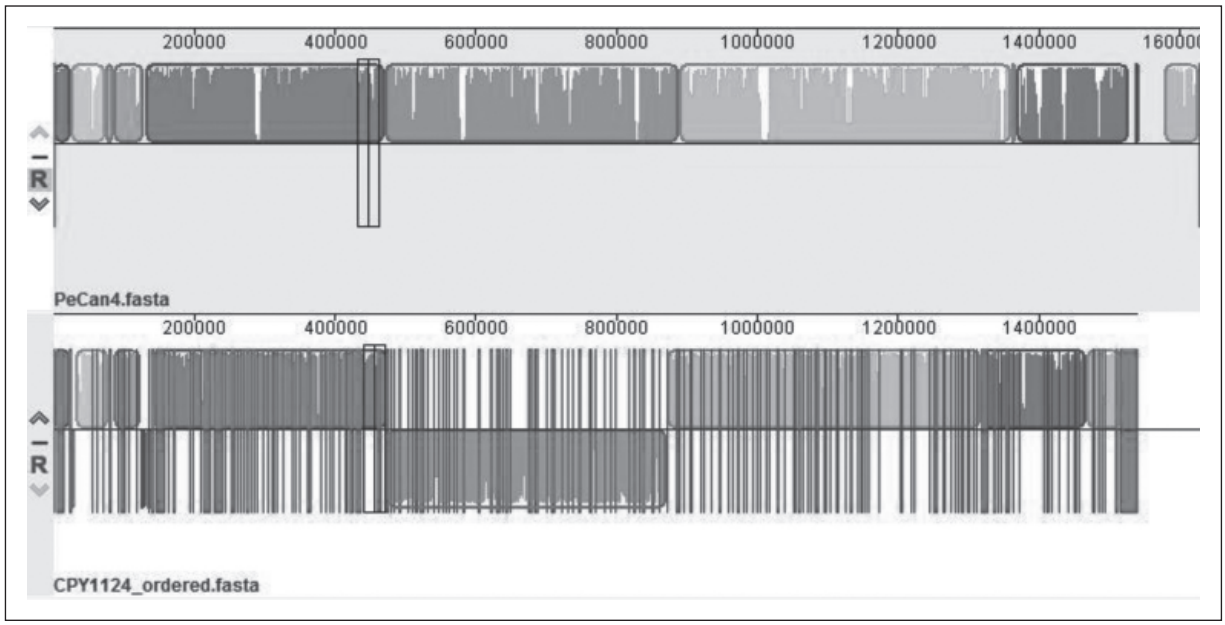

Fig. 5. Ordered Contigs Sequence of CPY1124.

Tandem Repeat Finder was used to call the possible STRs in ordered sequence of CPY1124 strain and the results obtained can be observed in Table 1, showing the sequence location and its statistical properties of various tandem repeats found in CPY1124 strain. STRs analysis on those tandem repeats was manually done and as the final result, two unique and polymorphic STRs have been obtained which were AAACCC and GTTTG. AAACCC STR was found repeating so many times on two different locations in CPY1124 strain which are on (770 – 817) sequence index and (86 -- 116)

Table 1. Summary of Sequence Location and Statistical Properties of Tandem Repeats Found in CPY1124

Sequence Index	Sequence Description	Number of Repeats
5	NODE_363_length_704_cov_180.048294	1
10	NODE_46_length_29817_cov_163.648529	1
30	NODE_179_length_2109_cov_256.657196	1
59	NODE_150_length_1787_cov_171.451599	2
139	NODE_23_length_19640_cov_146.881821	2
151	NODE_137_length_26171_cov_132.700974	1
214	NODE_288_length_11839_cov_144.783600	1
265	NODE_671_length_595_cov_173.875626	1
273	NODE_81_length_9254_cov_232.632706	1
274	NODE_302_length_848_cov_256.259430	1
279	NODE_578_length_1523_cov_138.957977	1
291	NODE_590_length_243_cov_8.637860	1
303	NODE_618_length_629_cov_211.882355	1
307	NODE_128_length_34091_cov_143.769470	1
314	NODE_122_length_1908_cov_380.372131	1
337	NODE_155_length_493_cov_60.551723	1
340	NODE_197_length_972_cov_255.145065	1

sequence index (Table 2 and 3). Meanwhile for GTTTG with five repeat units, has been found on (1194 – 1237) sequence index (Table 4).

DISCUSSION

As to be highlighted, the core of this new pipeline consists of five main steps as mentioned above. FastQC was responsible as the first evaluator tool to check the quality of CPY1124 sequence, whether it was of good quality or not. As CPY1124 was a NGS data, it might contain higher number of PCR duplication due to library construction during sequencing process (Illumina, 2012), thus led to higher sequence duplication level in it. Velvet *de novo* assembler was performed to assemble reverse and forward sequences of CPY1124 in order to construct *de Bruijn* graph. Velvet was chosen as the assembler tool in this research due to its effectiveness in time and efficiency in space (Zerbino, 2009). Velvet effectively manipulates *de Bruijn* graph by simplifying and compressing the data into several nodes without any loss of the graph data. Besides, it also corrects the overlapping regions in sequences by using repeat solver algorithm and eliminates any detected errors in them (Chaisson *et al.*, 2009). The assembly process resulted unordered contigs sequence of CPY1124. Then, it was realigned by using multiple genome alignment (MAUVE).

MAUVE has a viewing system which enables it to view the reordering structure of genome sequence (Darling *et al.*, 2004). The first sequence was used as a reference orientation to locally collinear blocks (LCBs), which was also known as homologous region of sequence. In this analysis, LCBs were shared by CPY1124 and PeCan4 strains. Additionally, MAUVE was used to determine the colour block of homology sequence with a unique colour, indirectly enables clear visualization of homology sequence between targeted *H. pylori* strains, CPY1124 and reference genome strain, PeCan4 (Figure 5). There were identical conserved regions being observed between both strains which indicated by the same colour block between them. In fact, the presence of less number of white gaps also proved that they have high similarities in term of homology characteristic (Rissman *et al.*, 2009).

Tandem Repeat Finder, an online package was used to call all the possible STRs in ordered sequence of CPY1124 strain as it can identify and search for tandem repeats which were more likely being masked in larger homologous region. Its detection ability was specified by identity percentage and frequency of indels. This finder at the end will align repeat copies against a consensus sequence, displaying patterns of common mutations.

Table 2. Summary Table of GTTTG STR on (1194 – 1237) Sequence Indices

Sequence: NODE_150_length_1787_cov_171.451599
Parameters: 2 7 7 80 10 50 10
Length: 1817

Tables: 1

This is table 1 of 1 (2 repeats found)

Click on indices to view alignment

Table Explanation

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
1194--1244	9	5.7	9	97	0	84	0	3	45	50	1.20
1194--1237	5	9.6	5	79	18	51	0	2	43	54	1.12

Tables: 1

Table 3. Summary Table of AAACCC STR on (770 – 817) Sequence Indices

Sequence: NODE_578_length_1523_cov_138.957977
Parameters: 2 7 7 80 10 50 10
Length: 1553

Tables: 1

This is table 1 of 1 (1 repeats found)

Click on indices to view alignment

Table Explanation

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
770--817	6	8.0	6	85	0	60	45	45	4	4	1.41

Tables: 1

Table 4. Summary Table of AAACCC STR on (86 – 116) Sequence Indices

Sequence: NODE_618_length_629_cov_211.882355
Parameters: 2 7 7 80 10 50 10
Length: 659

Tables: 1

This is table 1 of 1 (1 repeats found)

Click on indices to view alignment

Table Explanation

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
86--116	6	5.2	6	92	0	53	54	45	0	0	0.99

Tables: 1

These patterns unmasked the overall history of duplications that produced the tandem repeat, thus providing a potentially useful tool in phylogenetic research.

As a result, this pipeline enables to identify polymorphic and unique STRs which were GTTTG and AAACCC. The probabilistic model was used in Tandem Repeat Finder in order to interpret the results. The probability of success, which was also known as matching probability (PM), represents the average of percent identity between the copies whereas indels probability (PI) specifies the average of percent indels between the copies (Benson, 1999). This probability model fits with the parameter used in this analysis which is (2: match, 7: mismatch, 7: indels). Therefore, it showed that the STRs results obtained were reliable. In fact, a verification step has been done in order to validate the results. The similar results produced by pSTR (Table 5) and TRDB (Figure 6, 7 and 8) thus proving the reliability of this new pipeline.

Several improvements can be applied on this suggested pipeline, for example by performing the laboratory work to select the polymorphic STRs for targeted species as they can be highly reproducible in laboratory (Jones *et al*., 1997). This protocol will involve initial STR-primers screening, multiplex-PCR amplification, fragment analysis and genotyping. In fact, this method has been applied to discover simple sequence repeats (SSRs) of various tropical trees such as *Shorea leprosula* (Lee *et al*., 2004), *S. platyclados* (Ng *et al*., 2009; 2013) and *Gonystylus bancanus* (Ng *et al*., 2009). According to Weising *et al* (2005), designing flanking primers for isolation of STR loci must meet certain essential criteria so that the primers can work effectively by amplifying the templates specifically instead of multiple sites amplification. The lengths of primers designed are usually 20 to 24 bp to work well. The GC content of a pair of primers should be around 45 to 55% which corresponds to melting temperature (T_m) around 55º to 60º. Primers with higher Tm will misprime at lower temperature whereas lower T_m may lead to malfunction of primers at high annealing temperature. The repetition of nucleotide sequences inside of a primer sequence should be avoided since the occurrence of repeats may lead to mis prime occurrence. These laboratory procedures can be done as additional verification step although it will take time to really isolate the polymorphic STRs for certain species.

Besides, instead of using Tandem Repeat Finder, STRs can be identified and located by using Microsatellite Identification Tool (MISA) in Linux environment or QDD. QDD, an open-source for SSR search tool package, is the latest tool that focuses on STR marker discovery through NGS read analysis. It provides a pipeline from raw NGS reads to STR identification and corresponding primer design (Chen *et al*., 2014).

CONCLUSION

Since few decades ago, advances in technology of sequencing and computational analysis tools have led to study and analysis of genomic variations. In this study, *H. pylori* strain, CPY1124 is used to demonstrate whether this proposed pipeline is capable in detecting polymorphic and unique STR from NGS data or not. By undergoing these five core steps in suggested pipeline that are being mentioned above, there are two polymorphic and unique STRs can be called from the whole genome of *H. Pylori* strain, CPY1124 which are GTTTG and AAACCC on particular location index of that genome. Verification process for this pipeline has been done by comparing the result obtained with other available STRs pipeline, for example pSTR Finder and Tandem Repeat Database. Both pipelines also have produced GTTTG and AAACCC as their STRs calling outputs. This proven that the suggested pipeline is able to do STRs calling and also profiling from the NGS data. Also, this study supports the efficiency of NGS data in providing faster and reliable resources for STR development.

Table 5. Results of pSTR Finder

Source Id	Source Repeat Unit	Source Repeat Number	Source STR Index	Source STR Sequence
NODE_618_length_629_cov_ 211.882355	6	5.2	86	AAACCCAAACCCAAA CCCAAACCAAAACCCA
NODE_578_length_1523_cov_ 138.957977	6	8	770	AAACCCAAACCCAAACCCAAGCCC AAACCTAAGCCTAAACCCAAACCC
NODE_150_length_1787_cov_ 171.451599	5	9.6	1194	GTTTGGTTGGTTTGGTTGG TTTGGTTGGTTTGGTT

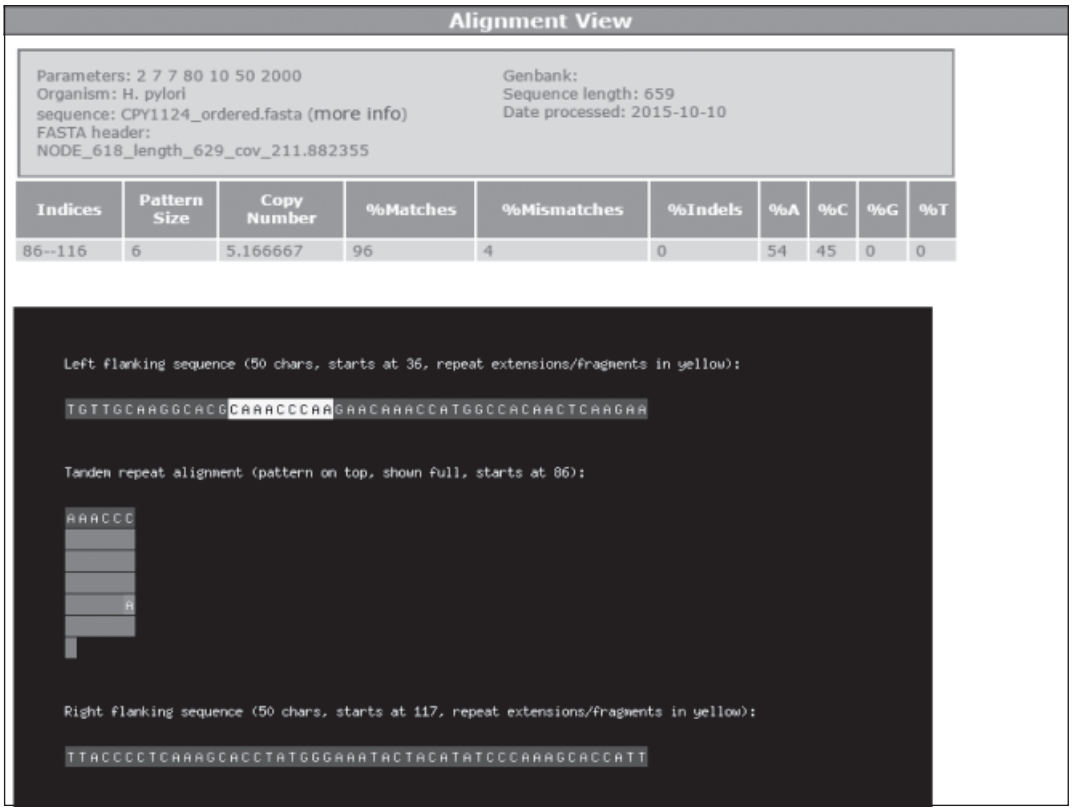

Fig. 6. TRDB Result: AAACCC on (86 – 116) Sequence Index.

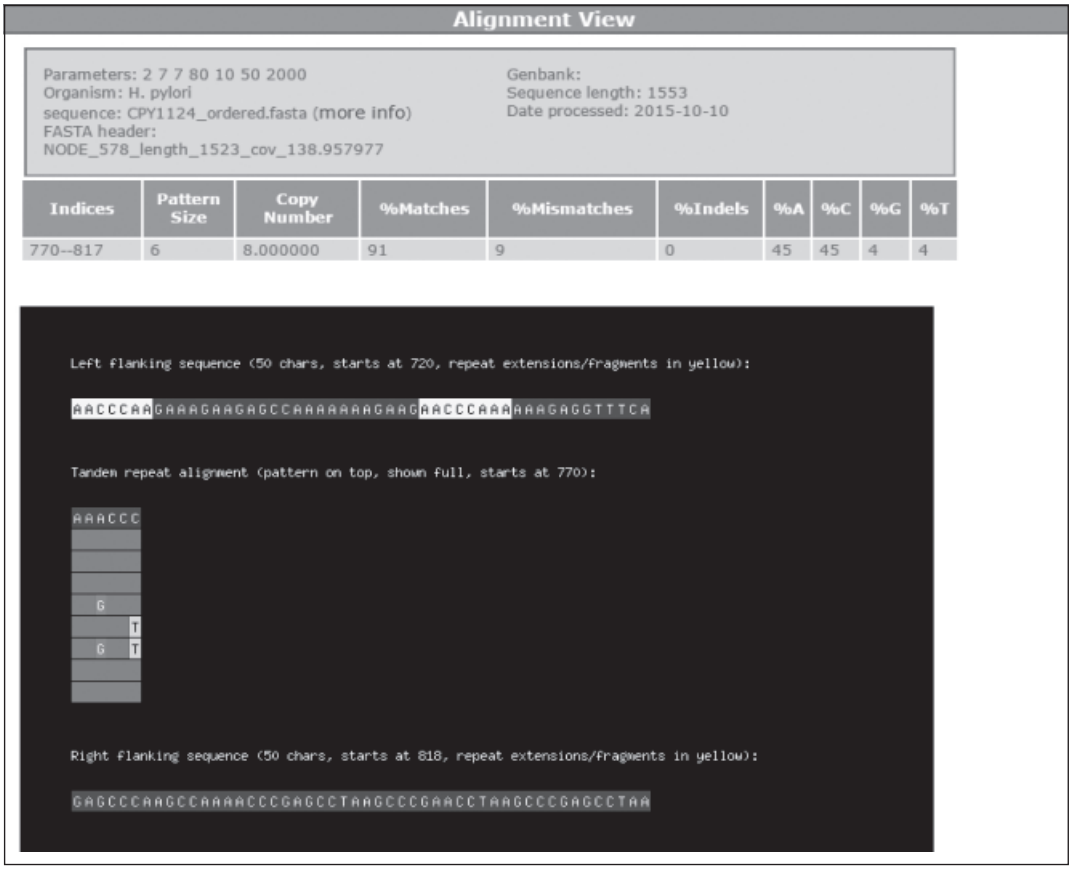

Fig. 7. TRDB Result: AAACCC on (770 – 817) Sequence Index.

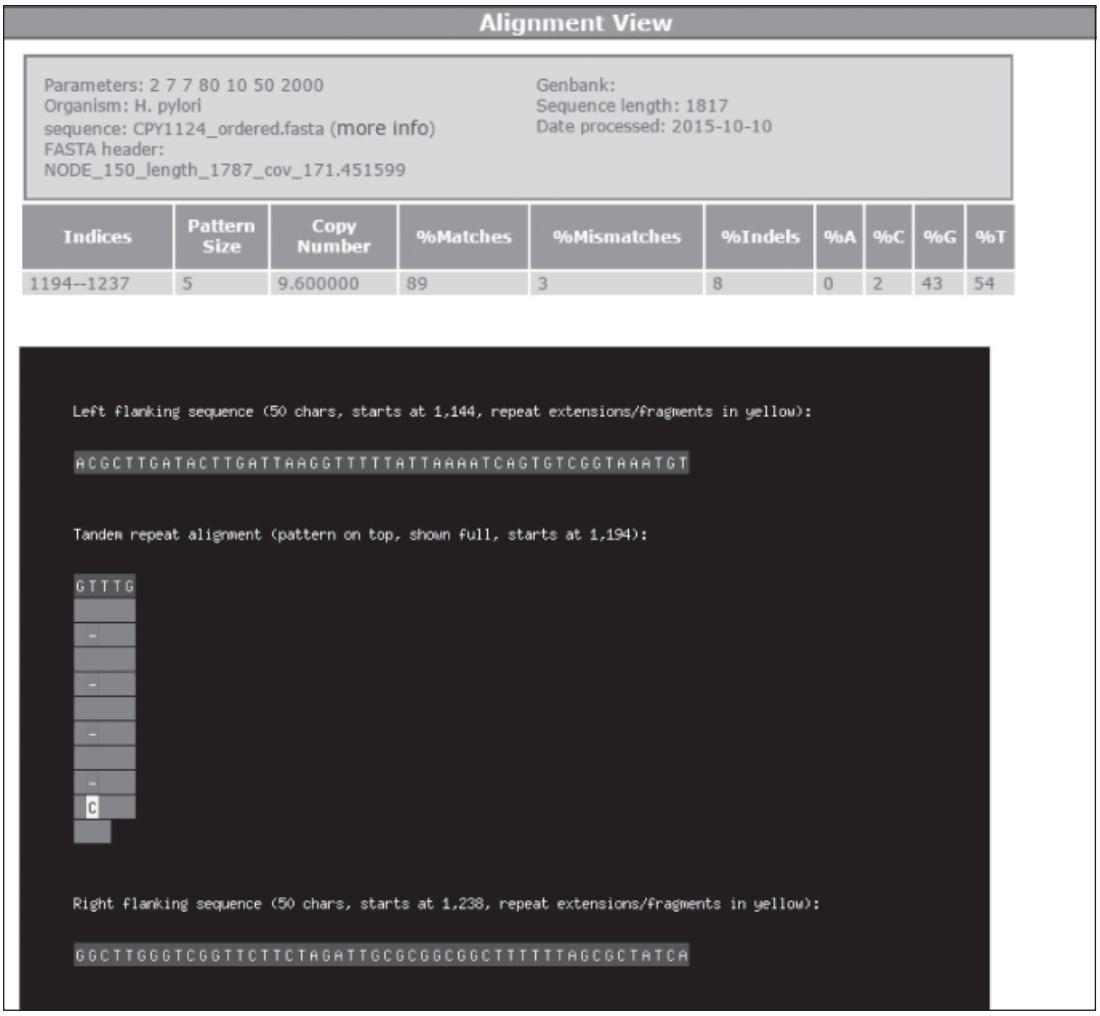

Fig. 8. TRDB Result: GTTTG on (1194 – 1237) Sequence Index.

ACKNOWLEDGEMENT

I would like to thank Management and Science University for giving opportunity, guidance and support in doing this research project and thus completing the manuscript successfully.

REFERENCES

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A. & Kayser, M. 2010. Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and rorensic implications. *American Journal of Human Genetics*, **87(3)**: 341–353.

Benson, G. 1999. Tandem repeats: a program to analyze DNA sequences. *Nucleic Acids Research*, **27(2)**: 573–580.

Butler, J.M., Coble, M.D. & Vallone, P.M. 2007. STRs vs. SNPs: Thoughts on the future of forensic DNA testing. *Forensic Science, Medicine, and Pathology*, **3(3)**: 200–205.

Chaisson, M.J., Brinza, D. & Pevzner, P.A. 2009. *De novo* fragment assembly with short mate-paired reads: Does the read length matter? Genome Research 19: 336–346 *Clinical. Microbiology and Infection.* **15**: 971–976.

Chen, C., Sio, C., Lu, Y., Chang, H., Hu, C. & Pai, T. 2014. Identification of conserved and polymorphic STRs for personal genomes, *BMC Genomics,* **15(10)**: 1–16.

Darling, A.C.E., Mau, B., Blattner, F.R. & Perna, N.T. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, **14**: 1394–1403.

Edwards, R.A., Olson, R., Disz,T., Pusch, G.D., Vonstein, V., Stevens, R. & Overbeek, R. 2013. Real time metagenomics: using k-mers to annotate metagenomes. *Bioinformatics*, **28**: 3316–3317.

Gemayel, R., Vinces, M.D., Legendre, M. & Verstrepen, K.J. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics,* **44**: 445–477.

Hartwell, L.H., Hood, L., Goldberg, M.L., Reynolds, A.E., Lee, M. & Veres, S.R.C. 2008. *Genetics from genes to genomics*, 3rd ed. Boston: McGraw Hill.

Illumina. 2012. *An Introduction to Next-Generation Sequencing Technology*, 1–12pp.

Jayabalan, N. 2006. *Plant biotechnology*. New Delhi: A.P.H. Publishing Corporation, 73pp.

Jones, C.J., Edwards, K.J., Castaglione, S., Winfield, M.O., Sala, F., Wiel, C. Van De & Karp, A. 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories, *Molecular Breeding*, **3(8):** 381–390.

Lee, S.L., Tani, N., Ng, K.K.S. & Tsumura, Y. 2004a. Isolation and characterisation of 20 microsatellite loci for an important tropical trees *Shorea leprosida* (Dipterocarpaceae) and their applicability to *S. parvifolia*. *Molecular Ecology Notes,* **4**: 222-225.

Lupski, J.R. 2007. Genomic rearrangements and sporadic disease. *Nature Genetics*, **39(7)**: 43–47.

Ng, K.K.S, Lee, S.L. & Koh, C.L. 2004. Spatial structure and genetic diversity of selected tropical tree species with contrasting breeding systems and different ploidy levels. *Molecular Ecology*, **13**: 657-669.

Ng, K.K.S., Lee, S.L., Ng, C.H., Tnah, L.H., Lee, C.T. & Tani, N. 2009. Microsatellite markers of *Gonystylus bancanus* (Thymelacaccae) for population genetic studies and DNA fingerprinting. *Conservation Genetic Resources* **1**: 153-157.

Rissman, A.I., Mau, B., Biehl, B.S., Darling, A.E., Glasner, J.D. & Perna, N.T. 2009. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, **25(16)**: 2071-2073.

Treangen, T.J. & Salzberg, S.L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews. *Genetics*, **13(1)**: 36–46.

Walker, J.M. & Rapley, R. 2005. Microsatellite analysis. Medical bio-methods handbook. New Jersey, USA: Humana Press Inc., **33**: 463–469pp.

Weising, K., Nybom, H., Wolff, K. & Kahl, G. 2005. *DNA fingerprinting in plants: Principles, methods, and applications*, 2nd ed. Boca Raton, Florida, USA: CRC Press, Taylor & Francis Group.

Zerbino, D.R. 2009. Genome assembly and comparison using de Bruijn graphs. *Molecular Biology*, **149**: 23-30.

Zhao, Y., Williams, R., Prakash, C.S. & He, G. 2013. Identification and characterization of gene-based SSR markers in date palm (*Phoenix dactylifera* L.), *BMC Plant Biology*, **12**: 1-8.