Original Article

GenoSSRFinder: a tool for rapid, precise, and targeted simple sequence repeat detection in genomic studies

GenoSSRFinder: uma ferramenta para detecção de repetição de sequência simples rápida, precisa e direcionada em estudos genômicos

M. Alhawatema[a]*

[a]Tafila Technical University, Faculty of Science, Department of Applied Biological Science, Tafila, Jordan

Abstract

The GenoSSRFinder is a new tool enables the research of Simple Sequence Repeats (SSRs) in DNA sequences and genomes much simpler and more precise in short time. The analysis is carried out by targeting a certain SSR in genome and gene sequences. This utility is quick, accurate, and does its function quite well. It quickly goes across the sequence, revealing all of the locations at which the selected SSR may be found. This tool will tell researchers where selected SSR begins and where it stops, how long it is, how often it repeats, and how long each repetition is. GenoSSRFinder gets the findings quickly, and they will be simple to comprehend. Therefore, when studying SSRs, researchers will have more time to use to thorough work as a result of this time savings. In addition, it provides a valuable information since it is highly precise. GenoSSRFinder is simple to use and produces high-quality findings. It is also accelerating SSRs gene research, which is a direct result of the new approach we use to analyse SSRs. Three case studies in this study demonstrated the usefulness of this program by immediately studying a particular SSR that was associated with genetic illness, biodiversity and criminal science in living organisms. This demonstration explains that GenoSSRFinder might be utilized in a wide variety of fields, such as the research of genetic illnesses, the biodiversity and genetic studies, or even in criminal investigations.

Keywords: GenoSSRFinder, SSRs, simple sequence repeats.

Resumo

O GenoSSRFinder é uma nova ferramenta que possibilita a pesquisa de Simple Sequence Repeats (SSRs) em sequências de DNA e genomas muito mais simples e precisos em curto espaço de tempo. A análise é realizada visando um determinado SSR no genoma e nas sequências de genes. Este utilitário é rápido, preciso e faz sua função muito bem. Ele percorre rapidamente a sequência, revelando todos os locais em que o SSR selecionado pode ser encontrado. Essa ferramenta informará aos pesquisadores onde o SSR selecionado começa e onde termina, quanto tempo dura, com que frequência se repete e quanto tempo dura cada repetição. O GenoSSRFinder obtém as descobertas rapidamente e elas serão simples de compreender. Portanto, ao estudar SSRs, os pesquisadores terão mais tempo para usar em trabalhos minuciosos como resultado dessa economia de tempo. Além disso, fornece uma informação valiosa, pois é altamente precisa. O GenoSSRFinder é simples de usar e produz achados de alta qualidade. Também está acelerando a pesquisa de genes SSRs, que é um resultado direto da nova abordagem que usamos para analisar SSRs. Três estudos de caso neste estudo demonstraram a utilidade deste programa ao estudar imediatamente um SSR específico que foi associado a doenças genéticas, biodiversidade e ciência criminal em organismos vivos. Esta demonstração explica que o GenoSSRFinder pode ser utilizado em uma ampla variedade de campos, como a pesquisa de doenças genéticas, a biodiversidade e estudos genéticos, ou mesmo em investigações criminais.

Palavras-chave: GenoSSRFinder, SSRs, repetições de sequência simples.

1. Introduction

Studying and understanding Simple Sequence Repeats (SSRs) requires analysing enormous amounts of genomic data (Ellegren, 2004). This is one of the many challenges that come with the task of searching genomes data for SSRs, which is both interesting and difficult (Ellegren, 2004). These simple sequence repeats, which are also known as microsatellites, are hiding in different regions of the complex genome (Moxon et al., 2006; Kashi and King 2006). They are repeating sequences and may be anywhere in genome sequence from one to six base pairs in length, and they are found all across the genomes of a wide range of living organisms (Moxon et al., 2006). As a result, they are progressively getting attention as they are involved in many functions in living organism (Moxon et al., 2006). The SSRs perform a wide variety of tasks

throughout the biology and evolution of these living organisms (Moxon et al., 2006), including playing an important role in the genetic diversity of living organism and they are used as genetic markers. In addition to this, they are implicated in the manifestation of some diseases and play a vital part in the process of gene transcription (Pearson et al., 2005; Gymrek et al., 2016). The study of SSRs has the potential to provide important insights into a variety of biological processes (Kashi and King, 2006). Dinucleotide repeats like (AT)n, which are present in yeast, are linked to DNA bending and Z-DNA synthesis, whereas key mononucleotide repeats like (A)n and (C)n in humans play a crucial role in the regulation of transcription (Li et al., 2002). The physical structure and function of DNA are therefore profoundly affected by these dinucleotide repeats (Li et al., 2002). Another well-known class of SSRs is trinucleotide repeats. (CAG)n, (CGG)n, and (CTG)n repeats have been linked to many human genetic diseases (Pearson et al., 2005). Myotonic dystrophy, Fragile X syndrome, and Huntington's disease are only a few examples (Pearson et al., 2005). The need for studying SSRs to the understanding many of functions that they play has increased a demand for efficient tools that can be used to analyse them quickly and precisely. GenoSSRFinder is a python-based software was developed to satisfy this need. It is a simple software that was developed particularly for the purpose of enhancing SSR research. The main goal of GenoSSRFinder is to make this essential part of genomic research easier to understand, which may be accomplished by giving a targeted and efficient tool for SSR studies. Therefore, it is designed to rapidly and accurately identify particular SSRs that are present within genomic data.

GenoSSRFinder is a new tool (GenoSSRFinder download for PC) (GenoSSRFinder, 2023) in genomic research that aims to improve the speed and accuracy of the search for SSRs within genomic data. This program is an add forward in the field of bioinformatics concerning SSRs studies when taking into account the essential part that SSRs play in gaining an understanding of the genetic diversity of organisms (Kashi and King, 2006; Buschiazzo and Gemmell, 2006). Therefore, this a new tool will be of extremely beneficial in this field. GenoSSRFinder has the ability of finding and searching a particular SSR in genomic data, in addition to giving information about it. Therefore, it is a significant tool for academics since it provides an approach to the analysis of SSRs that is both simple and quick to utilize.

The hypothesis of this study is that the introduction of GenoSSRFinder, and an examination of its potential involvement in the field of genomics . This study aims to show its potential uses through three cases as studies to examine the function that SSRs play in genetic diversity, genetic illnesses, and even in criminal investigations.

2. Methodology

2.1. Introduction to GenoSSRFinder: an innovative approach

The development of GenoSSRFinder requires the use of a specific approach to develop a software capable of precisely targeting a particular SSR within big genomic data and DNA sequences quickly. To accomplish this aim, Python, a flexible and easy programming language was our tool of choice (Van Rossum and Drake 2009; Harris et al., 2020) as it has ready libraries such as Biopython which has ready codes for DNA sequences (Cock et al., 2009).

2.2. Python key libraries utilized

Python is a powerhouse of a language with a variety of libraries that were used for developing GenoSSRFinder (Harris et al., 2020). The programming codes uses some important known key libraries like Pandas (McKinney, 2010), NumPy (Harris et al., 2020), and Biopython (Cock et al., 2009). High-performing data analysis was made available by the open-source data analysis Pandas (McKinney, 2010). Mathematical operations were calculated using NumPy library which made them easier to perform on arrays and matrices containing numeric data (Harris et al., 2020) . Genomic data files were often decoded using Biopython, a python programming language with a suite of tools for biological computing and molecular biology (Cock et al., 2009).

2.3. File reading and processing in GenoSSRFinder

The first step of GenoSSRFinder was to open and read the Fasta file that is the most common genomic file format used. We have included a method that can read and process Fasta files to guarantee proper accepting this file format. This project was greatly aided by the SeqIO module in Biopython (Cock et al., 2009).

2.4. Developing a user-friendly interface for SSR Identification

This was done by creating input fields where users could provide the file path of the genomic data by downloading the Fasta file of genome sequence or gene sequence (Figure 1). The detail of the repeat unit of interest is easily filled by user. The user select a certain SSR to search for by filling repeat unit, and then minimum and maximum repetitions (Figure 1).

2.5. SSR search and recording: a detailed process

Upon completion of the input stage, the software starts searching for the specified SSR selected by users. This operation is undertaken by a function and examines the genome sequence, recording every occurrence of the SSR based on input information provided by user. Each occurrence of selected SSR is followed by the recording of important details including the SSR start and SSR end within the genomic sequence or DNA sequences, the length of the repeat sequence, and the frequency of repetition.

2.6. Presentation of results: fast and user-friendly

We take the Accuracy in our approach to deal with as an important factor. Therefore, we integrated error-checking procedure to deal with any error when inputting data by users. Once the search is done, the results of GenoSSRFinder are presented for users on a box on screen that is generated by a Pandas DataFrame (McKinney, 2010). The results are then shown to users in seconds, and they are shown in easy

format, so that, researchers are able then to understand and analyse the data. This, reducing the time needed to study SSRs and making genomic research faster. To do this, we used the effective library NumPy (Harris et al., 2020) in the programming codes to make the process faster and takes less time. As a result, GenoSSRFinder is both accurate and precise (McKinney, 2010). It uses also simple interface and easy to use by researchers (Figure 1).

The development process also involved testing process to confirm the software's robustness by running this software on genomic data may times. It has an approach that has an advantage over the traditional programs for SSRs analysis that take hours to days for analysing genomic data against particular SSR. By concentrating on particular SSRs, it allows for a focused and precise examination of genomic data. The technique behind GenoSSRFinder underscores its potential as a tool in genomic research to search for a certain SSR in genomic data.

3. Results

The results generated using GenoSSRFinder are of value and remarkable (Figure 2-5). Once initiated, GenoSSRFinder rapidly scan across the entire genome sequence. The software then generates an in-depth account, outlining all the occurrences of the specified SSR. This account comprises crucial details, such as the SSR's
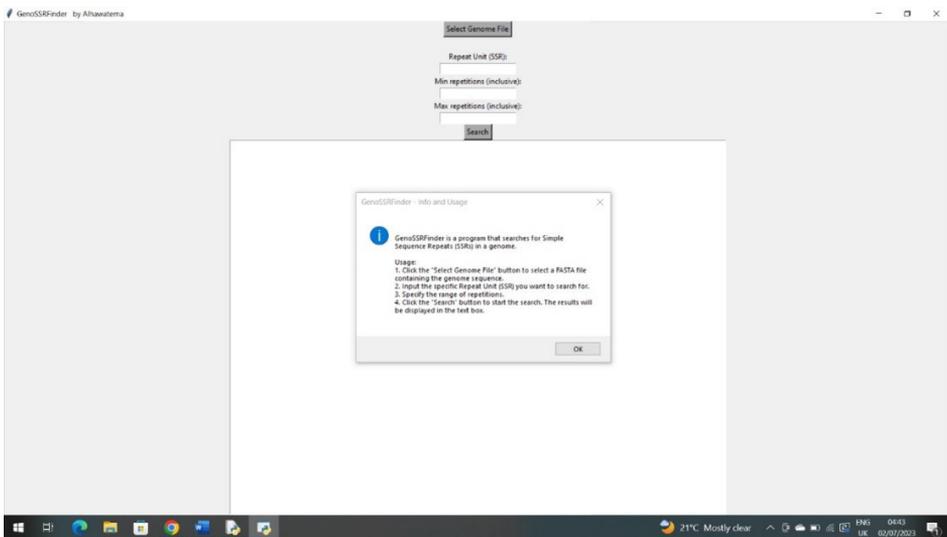
Figure 1. Displays the GenoSSRFinder user interface for searching a specific SSR.
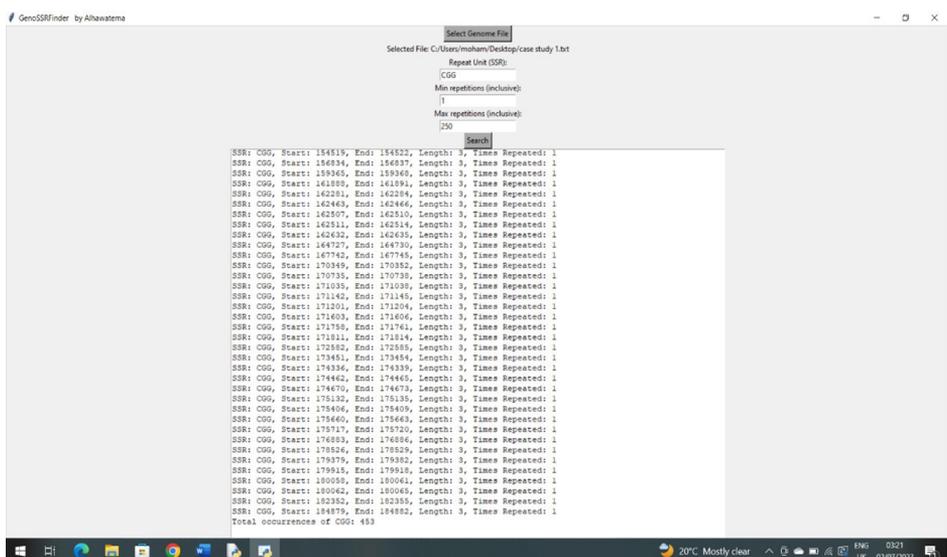
Figure 2. Application of GenoSSRFinder unveils significant trinucleotide SSR (CGG)n expansion in the FMR1 gene linked to Fragile X Syndrome.
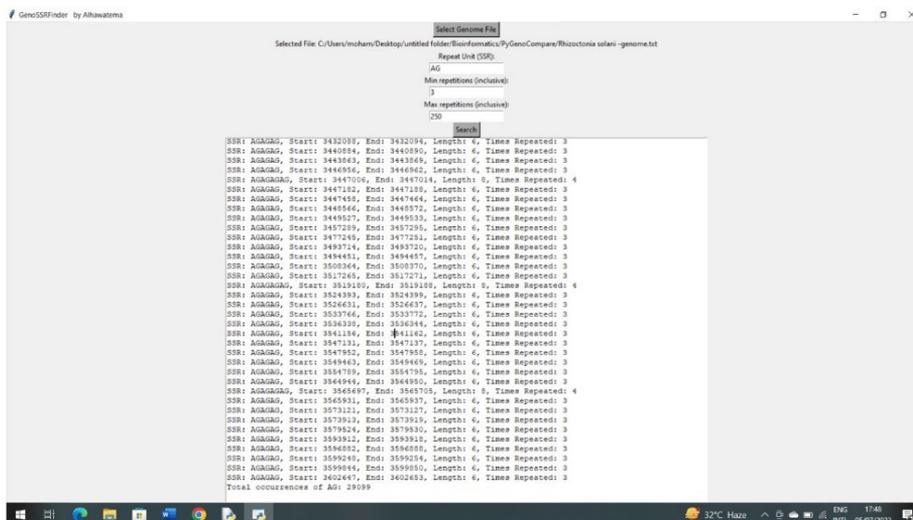
Figure 3. Comparative analysis of GenoSSRFinder's detection of SSR(AG)n repeats in *Rhizoctonia Solani* unveils notable genetic diversity.
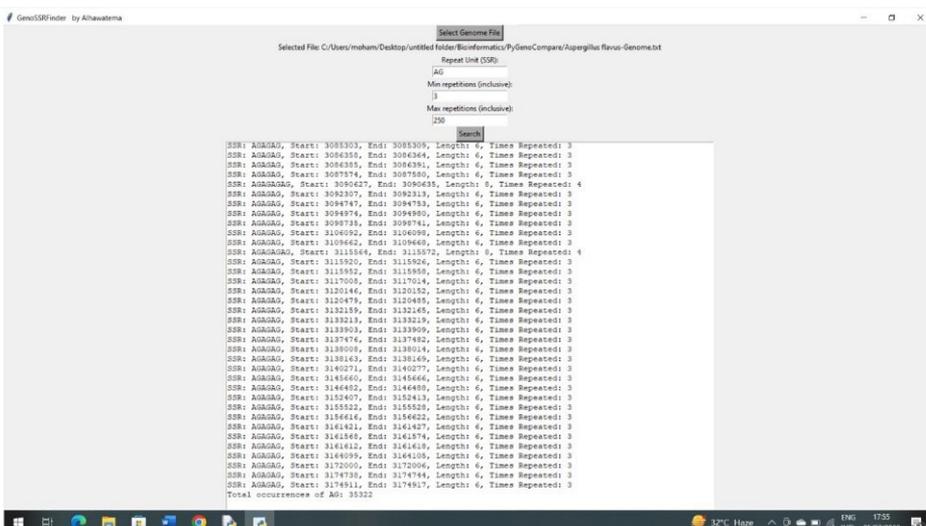
Figure 4. Comparative analysis of GenoSSRFinder's detection of SSR(AG)n repeats in *Aspergillus Flavus* unveils notable genetic diversity.

start and end, length, and the frequency of its repetition. The results are presented on screen.

The user-friendly nature of the results simplifies interpretation, making the research process more streamlined. The findings generated through GenoSSRFinder have consistently shown a high degree of accuracy, a testament to the software's efficiency. The dependable precision of the tool assures researchers that they are dealing with reliable data.

Moreover, GenoSSRFinder considerably reduces the time typically needed for SSR analysis. The software has demonstrated its ability to deliver results more promptly than traditional SSR analysis tools. This is attributable to its focus on particular SSRs, which facilitates a more specific search. The speed at which the software operates is one

of its most defining features. By saving precious time, GenoSSRFinder enables researchers to engage in more detailed investigations.

These efficient and precise outcomes can significantly accelerate the pace of genomic research. The outputs created by the software are not only fast and accurate but also comprehensive. Consequently, GenoSSRFinder offers a complete view of the specified SSR within the genome, ensuring that no potential SSR instance goes unnoticed.

Thus, GenoSSRFinder grants researchers a more profound comprehension of the genome. Coupled with the software's user-easy design, these findings position GenoSSRFinder as the optimal solution for SSR analysis. In essence, the outcomes derived from GenoSSRFinder underpin its potential to control SSR analysis. The software
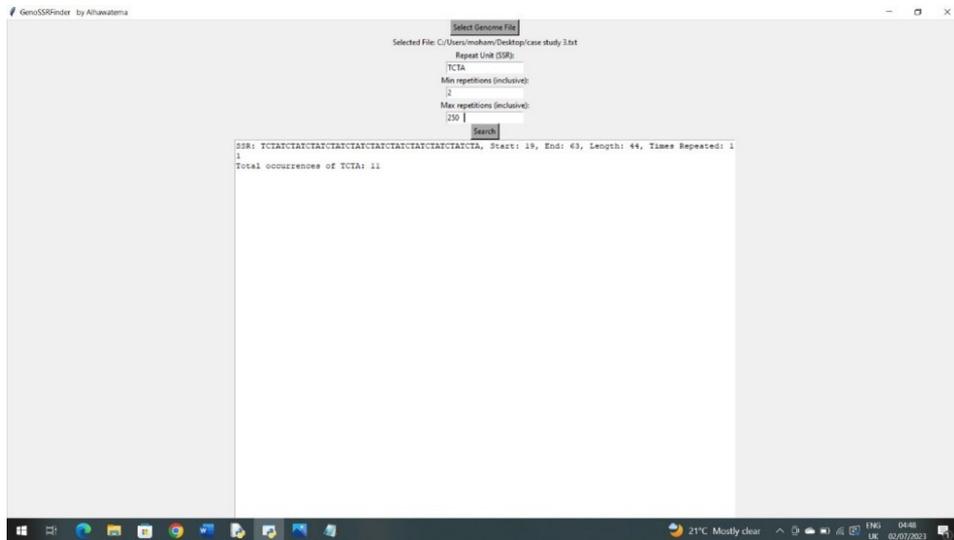
Figure 5. GenoSSRFinder identifies SSR Pattern (TCTA)n in the D18S51 locus, forensic genetic analysis.

has proven to be a beneficial resource for researchers for analysing certain SSR within genomic data looking for genetic illnesses, the biodiversity and genetic studies, or even in criminal investigations.

3.1. Case Study 1: GenoSSRFinder as an instrument for genetic disorder investigation

To deepen our grasp of genetic diseases, it's crucial to identify and explore specific SSRs associated with genetic disorders (Andrew et al., 1993; Brook et al., 1992). GenoSSRFinder has the potential to be expanded so that it may search for SSRs related with a variety of additional genetic conditions. For instance, Huntington's disease has been connected to the Huntingtin gene, which contains the trinucleotide repeat (CAG)n. It has been determined that the FMR1 gene, which includes the trinucleotide repeat (CGG)n, is connected with the fragile X syndrome (Verkerk et al., 1991). In addition, there is a link between myotonic dystrophy and the DMPK gene, which has the trinucleotide repeat (CTG)n mutation (Andrew et al., 1993; Brook et al., 1992).

Therefore, SSRs were studying an extremely rare genetic condition tied to a unique repeat, for example, expansion of trinucleotide SSRs are linked with several neurological diseases, such as fragile X syndrome particularly when the SSR (CGG)n has an expansion to be more than 200 times in fragile X syndrome compared with 40 times in normal (Jin and Warren, 2000). GenoSSRFinder was employed to identify this SSR, significantly accelerating the research by identifying the SSR's location and repetition frequency (Figure 2). We used GenoSSRFinder to search for the SSR(CGG) in (GenBank #(L29074.1, Homo sapiens fragile X mental retardation syndrome protein (FMR1)), the sequence was retrieved from the GenBank database (NCBI, 2023). Based on results from Figure 2, GenoSSRFinder revealed that the total occurrence of SSR(CGG)n =453 in FMR1 gene which is connected to fragile X syndrome (Jin and Warren, 2000).

3.2. Case Study 2: GenoSSRFinder, as a tool on genetic diversity

Simple sequence repeats, or SSRs, has a considerable influence on genetic variety, which is essential for the continued existence of a species. Because it is able to locate and examine these SSRs in a wide variety of living organism(Moxon et al., 2006). GenoSSRFinder has become an indispensable tool for the study of biodiversity and conservation biology. Research on genetic diversity often makes use of the SSR gene, which has the repeat pattern (AG)n and is found in a great variety of fungal species (Nybom, 2004). The capability of GenoSSRFinder to quickly discover and study individual SSRs has the potential to improve our knowledge of genetic diversity. We used GenoSSRFinder to search for the SSR pattern (AG)n in *Rhizoctonia solani* chromosome 1 (GenBank #(NC_057370.1) and *Aspergillus flavus* chromosome 1 (GenBank #(NC_054691.1). The sequences were retrieved from the GenBank database (NBCI, 2023). Based on results from Figures 3 and 4. GenoSSRFinder results revealed that the total occurrence of SSR(AG)n= 29099 in *Rhizoctonia solani* compared with *Aspergillus flavus* that has SSR(AG)n= 35333

3.3. Case Study 3: GenoSSRFinder as a tool in forensic science practices

DNA profiling, an integral part of forensic science utilized for identification purposes, requires accurate detection of SSRs to yield successful results. (Butler, 2006).

Identification in forensics often on loci sequences like vWA and D8S1179, both possessing the repeat pattern (TCTA)n, alongside D18S51, characterized by the repeat pattern (AGAA)n (Butler, 2006). The quick detection capabilities and detailed SSR information provided by GenoSSRFinder could simplify the procedures involved in forensic inquiries, possibly enabling faster case resolutions We used GenoSSRFinder to search for the SSR pattern (TCTA)n in the D18S51 locus chromosome (GenBank

#(MH105190.1). The sequence was retrieved from the GenBank database (NCBI, 2023) (Figure 5).

The results of GenoSSRFinder for searching (TCTA)n in the human D18S51 locus indicated that D18S51 locus has a total occurrence = 11 (Figure 5). This result is of important in forensic science for certain SSR numbers comparing between individuals examined (Butler, 2006).

4. Discussion

It is observed from obtained results of analysis of case study 1, 2, and 3 by GenoSSRFinder that the methodology is associated with the tool have profound implications as GenoSSRFinder has an efficient and precise approach to SSR searching. Therefore, GenoSSRFinder holds the potential to speed genomic research of SSRs. This simple program is of great importance in such as genomics to study SSRs concerning biodiversity, genetic diseases and even a forensic science . This program is designed to improve the SSRs studies as many scientists interested in studying the function and importance of such SSR in living organisms will start soon using it.

GenoSSRFinder is a tool used to study genetic patterns of SSRs, and it is useful and accurate and time saving in SSRs research compared with other tools available. It uses Python programming and the best of its libraries in making to search certain important genetic markers SSRs in short time and is quicker than most other tools available. (Harris et al., 2020; McKinney, 2010; Cock et al., 2009), which is of importance in many fields related to gene studies including genetic diseases, the biodiversity and genetic studies, and even in criminal investigations (Verkerk et al. 1991; Jin and Warren, 2000).

In the first study case (Figure 2), we used GenoSSRFinder to look at genetic diseases. We focused on special genetic markers related to diseases like fragile X syndrome (Verkerk et al. 1991). This way, GenoSSRFinder was able to get a lot of important data. The tool was great at finding SSR(CGG)n in a sequence retrieved from GenBank and counting this certain SSR, showing us how useful it can be for studying genetic diseases(Verkerk et al. 1991; Jin and Warren, 2000; Nybom, 2004; Butler, 2006)

In addition to genetic disease, GenoSSRFinder may be used in many fields such as genetic diversity (Nybom, 2004) . We used GenoSSRFinder in the second study case to study genetic diversity by searching a certain repeat (AG) which is important for genetic diversity in different types of fungi (Figures 3 and 4). The retrieved sequences of *Rhizoctonia solani* and *Aspergillus flavus* were searched by GenoSSRFinder for certain repeat (AG). GenoSSRFinder was really good at finding and studying the SSR(AG) we were interested in; this makes our work faster and simpler. This shows that GenoSSRFinder can be used for studies about biodiversity (Nybom, 2004).

On application that is interesting we used GenoSSRFinder for forensic science, helping us with DNA profiling. GenoSSRFinder was used for searching for the SSR pattern (TCTA)n of the D18S51 locus chromosome which is an important in forensic science investigation and giving us a valuable information (Butler, 2006). The results of

GenoSSRFinder (Figure 5) indicated that it may be of importance in the forensic science for certain SSR studies between individuals examined (Butler, 2006). This could make it as a tool to be used by forensic science researchers as it makes analysing certain SSRs easier and quicker. The easy design of GenoSSRFinder also enhances the idea that many people interested in studying SSRs will soon use it as the user interface and the results of GenoSSRFinder are easy to understand. Therefore, many researchers are more likely to use it. Thus, GenoSSRFinder not only improves research of SSRs but also will help solving many problems concerning SSRs analysis. By allowing researchers to target a specific SSR in seconds that plays important function within genome. Therefore, GenoSSRFinder is more than simply a program; it accelerates research for comparison of targeted regions of different organisms' genomes. For instance, by it we can compare certain SSRs between different genomes in short time. GenoSSRFinder is aimed to become an integral part of genomics research for targeting particular SSRs that play important functions in living organisms because of its ease of use with high quality output in fast time comparing with traditional SSR analysis tools.

5. Conclusion

GenoSSRFinder has emerged as a new method tool for SSR examination. It places emphasis on a researcher-oriented approach, which is characterized by effectiveness and precision, distinguishing it from other similar platforms. Concentrating on particular SSRs, GenoSSRFinder enables a more focused examination, a precision that is immensely advantageous in the many fields of genomic studies.

By lessening the duration time needed by researchers for analysis, GenoSSRFinder facilitates more frequent and research efforts. Moreover, the software's high degree of precision guarantees that the investigations are grounded in trustworthy data. Consequently, GenoSSRFinder augments both the volume and quality of genomic investigations. It is clear that GenoSSRFinder is really good at studying these special genetic markers. Its design focuses on being fast and accurate, and it's great at giving us detailed and correct information in all kinds of research regarding SSRs. The hypothesis regarding the introduction of GenoSSRFinder, and it is potential uses in the field of genomics was supported by it is effective uses in the three cases studies that were performed in this study showing it is potential uses in genetic diseases, biodiversity, or forensic science.

References

ANDREW, S.E., GOLDBERG, Y.P., KREMER, B., TELENIUS, H., THEILMANN, J., ADAM, S., STARR, E., SQUITIERI, F., LIN, B., KALCHMAN, M.A., GRAHAM, R.K. and HAYDEN, M.R., 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature Genetics*, vol. 4, no. 4, pp. 398-403. http://dx.doi.org/10.1038/ng0893-398. PMid:8401589.

BROOK, J.D., MCCURRACH, M.E., HARLEY, H.G., BUCKLER, A.J., CHURCH, D., ABURATANI, H., HUNTER, K., STANTON, V.P., THIRION, J.P., HUDSON, T., SOHN, R., ZEMELMAN, B., SNELL, R.G., RUNDLE, S.A., CROW, S., DAVIES, J., SHELBOURNE, P., BUXTON, J., JONES, C., JUVONEN, V., JOHNSON, K., HARPER, P.S., SHAW, D.J. and HOUSMAN, D.E., 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*, vol. 69, no. 2, pp. 385. http://dx.doi.org/10.1016/0092-8674(92)90154-5. PMid:1568252.

BUSCHIAZZO, E. and GEMMELL, N.J., 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays*, vol. 28, no. 10, pp. 1040-1050. http://dx.doi.org/10.1002/bies.20470. PMid:16998838.

BUTLER, J.M., 2006. Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences*, vol. 51, no. 2, pp. 253-265. http://dx.doi.org/10.1111/j.1556-4029.2006.00046.x. PMid:16566758.

COCK, P.J., ANTAO, T., CHANG, J.T., CHAPMAN, B.A., COX, C.J., DALKE, A., FRIEDBERG, I., HAMELRYCK, T., KAUFF, F., WILCZYNSKI, B. and DE HOON, M.J., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, vol. 25, no. 11, pp. 1422-1423. http://dx.doi.org/10.1093/bioinformatics/btp163. PMid:19304878.

ELLEGREN, H., 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews. Genetics*, vol. 5, no. 6, pp. 435-445. http://dx.doi.org/10.1038/nrg1348. PMid:15153996.

GENOSSRFINDER [online], 2023 [viewed 6 May 2023]. Available from: https://drive.google.com/file/d/1Eg2gohLz6VIzx6kApX2Ng8hv4WopZa3i/view

GYMREK, M., WILLEMS, T., GUILMATRE, A., ZENG, H., MARKUS, B., GEORGIEV, S., DALY, M.J., PRICE, A.L., PRITCHARD, J.K., SHARP, A.J. and ERLICH, Y., 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*, vol. 48, no. 1, pp. 22-29. http://dx.doi.org/10.1038/ng.3461. PMid:26642241.

HARRIS, C.R., MILLMAN, K.J., VAN DER WALT, S.J., GOMMERS, R., VIRTANEN, P., COURNAPEAU, D., WIESER, E., TAYLOR, J., BERG, S., SMITH, N.J., KERN, R., PICUS, M., HOYER, S., VAN KERKWIJK, M.H., BRETT, M., HALDANE, A., DEL RÍO, J.F., WIEBE, M., PETERSON, P., GÉRARD-MARCHANT, P., SHEPPARD, K., REDDY, T., WECKESSER, W., ABBASI, H., GOHLKE, C. and OLIPHANT, T.E., 2020. Array programming with NumPy. *Nature*, vol. 585, no. 7825, pp. 357-362. http://dx.doi.org/10.1038/s41586-020-2649-2. PMid:32939066.

JIN, P. and WARREN, S.T., 2000. Understanding the molecular basis of fragile X syndrome. *Human Molecular Genetics*, vol. 9, no. 6, pp. 901-908. http://dx.doi.org/10.1093/hmg/9.6.901. PMid:10767313.

KASHI, Y. and KING, D.G., 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, vol. 22, no. 5, pp. 253-259. http://dx.doi.org/10.1016/j.tig.2006.03.005. PMid:16567018.

LI, Y.C., KOROL, A.B., FAHIMA, T., BEILES, A. and NEVO, E., 2002. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, vol. 11, no. 12, pp. 2453-2465. http://dx.doi.org/10.1046/j.1365-294X.2002.01643.x. PMid:12453231.

MCKINNEY, W., 2010. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*. Austin, USA: SciPy Organizers, vol. 445, pp. 51-56. http://dx.doi.org/10.25080/Majora-92bf1922-00a.

MOXON, R., BAYLISS, C. and HOOD, D., 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annual Review of Genetics*, vol. 40, no. 1, pp. 307-333. http://dx.doi.org/10.1146/annurev.genet.40.110405.090442. PMid:17094739.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION - NCBI, 2023 [viewed 6 May 2023]. *GenBank* [online]. NCBI. Available from: www.ncbi.nlm.nih.gov/genbank/

NYBOM, H., 2004. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology*, vol. 13, no. 5, pp. 1143-1155. http://dx.doi.org/10.1111/j.1365-294X.2004.02141.x. PMid:15078452.

PEARSON, C.E., EDAMURA, K.N. and CLEARY, J.D., 2005. Repeat instability: mechanisms of dynamic mutations. *Nature Reviews. Genetics*, vol. 6, no. 10, pp. 729-742. http://dx.doi.org/10.1038/nrg1689. PMid:16205713.

VAN ROSSUM, G. and DRAKE, F.L., 2009. *Python 3 Reference Manual*. Scotts Valley: CreateSpace.

VERKERK, A.J., PIERETTI, M., SUTCLIFFE, J.S., FU, Y.H., KUHL, D.P., PIZZUTI, A., REINER, O., RICHARDS, S., VICTORIA, M.F., ZHANG, F.P., EUSSEN, B.E., VAN OMMEN, G.-J.B., BLONDEN, L.A.J., RIGGINS, G.J., CHASTAIN, J.L., KUNST, C.B., GALJAARD, H., THOMAS CASKEY, C., NELSON, D.L., OOSTRA, B.A. and WARREN, S.T., 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*, vol. 65, no. 5, pp. 905-914. http://dx.doi.org/10.1016/0092-8674(91)90397-H. PMid:1710175.