Beni-Suef University Journal of Basic and Applied Sciences

RESEARCH

Open Access

RFLP-kenzy: a new bioinformatics tool for in silico detection of key restriction enzyme in RFLP technique

Nora Laref[1*], Khadidja Belkheir[1], Mohamed Belazreg[1] and Abdelhadi Hireche[2]

Abstract

Background Today, several bioinformatics tools are available for analyzing restriction fragment length data. RFLP-kenzy is a new bioinformatic tool for identifying restriction key enzyme that cut at least 1 sequence and a maximum of n-1 sequence.

Results This bioinformatic tool helps researchers to select appropriate enzymes that yield different RFLP patterns, especially from overly identical sequences with single nucleotide mutation or other small variations. By using RFLP-kenzy, multiple DNA sequences could be analyzed simultaneously and the key enzymes list is provided. The present paper also demonstrates the ability of RFLP-kenzy to identify the key enzymes through the analysis of 16S rRNA sequences and the complete genome of various genera of microorganisms.

Conclusion From the results, several key enzymes were provided indicating the importance of this new tool in the selection of appropriate restriction enzymes.

Keywords Script, key enzymes, RFLP, Closely related sequences, Mutations

1 Background

RFLP is a method widely used for DNA analysis, established by Grodzicker and collaborators in 1974 [1]. It is based on the digestion of genomic DNA into fragments of different sizes, using known restriction enzymes [2]. This analytic technique is widely used as a tool in molecular biology. It allows the analysis of nucleotide base substitutions, insertions, deletions, duplications, or inversions within the whole genome which are noted by different RFLP patterns after the removal of a restriction site or the creation of a new one [3].

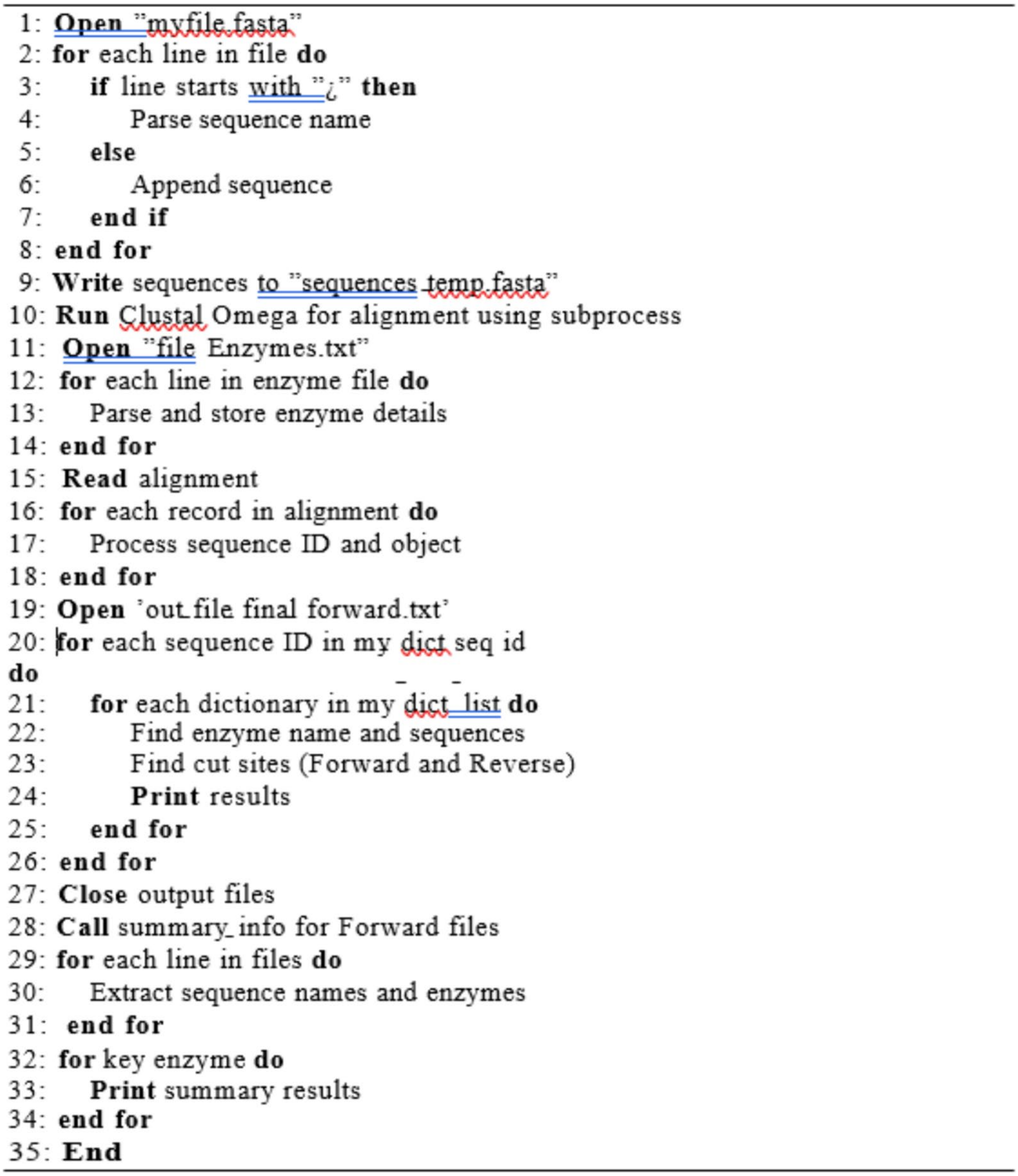
Digestions of sequences could be simulated in silico to find appropriate enzymes for experimental analysis [4–6]. Programs available on the web such as TAP-TRFLP (http://rdp.cme.msu.edu), torast (http://www.torast.de), and MiCA (http://mica.ibest.uidaho.edu/MICA (microbial community analysis)), NEB cutter 2.0, TReFID, or Cis SERS [7] are usually used to perform the in silico digestion of genes, but all these tools could not indicate the appropriate enzyme which allows a rapid resolution with accurate results, especially to differentiate sequences with high similarities. The current study focuses on the construction of a new script called RFLP-kenzy that helps biologists to find rapidly the key enzymes that can be used in an experimental study for reliable detection of mutation in DNA sequences. This tool could be used to identify new species, closely species separation, or human disease detection.

*Correspondence:
Nora Laref
nora.laref@univ-relizane.dz
[1] Department of Biology, Faculty of Sciences and Technology, University Ahmed Zabana of Relizane, Relizane, Algeria
[2] College of Information Technology, United Arab Emirates University, Abu Dhabi, United Arab Emirates

2 Materials and methods

2.1 Script construction

The key enzyme tool algorithm was created using Python 3, programming language version 3.11, and the Visual Studio IDE programming tool. For script editing, the file "enzyme_list_file.txt" was generated from the Bio. Restriction library. In general, this script processes eukaryotic, bacterial, or viral FASTA file sequences, which are first aligned with the Clustal Omega package 1.2.2-win64 (an external tool that must be installed in the given path c:\clustal-omega) and saved in Clustal file format (Fig. 1). The second part focuses on checking restriction sites in the sequences. In this step, all restriction sites are determined and stored for easy access in a list of dictionaries, recorded separately as "forward_site.txt," according to the enzymes available in the "enzyme_list_file.txt" (Fig. 1). The aligned sequences are then read using the AlignIO module from the Biopython library. Each sequence is searched for restriction sites corresponding to the enzymes listed in the enzyme file. The positions of the cut sites are identified. Results, including the sequence, enzyme, and the cut site information, are provided. Finally, this newest script selects, stores, and displays the key enzymes in an "out_file_final_forward.



```
Open ”myfile.fasta”
for each line in file do
    if line starts with ”¿” then
        Parse sequence name
    else
        Append sequence
    end if
end for
Write sequences to ”sequences temp.fasta”
Run Clustal Omega for alignment using subprocess
Open ”file Enzymes.txt”
for each line in enzyme file do
    Parse and store enzyme details
end for
Read alignment
for each record in alignment do
    Process sequence ID and object
end for
Open ’out_file final forward.txt’
for each sequence ID in my dict_seq_id do
    for each dictionary in my dict_list do
        Find enzyme name and sequences
        Find cut sites (Forward and Reverse)
        Print results
    end for
end for
Close output files
Call summary_info for Forward files
for each line in files do
    Extract sequence names and enzymes
end for
for key enzyme do
    Print summary results
end for
End
```

Fig. 1 Algorithm representation of the process of RFLP-kenzy

txt," with special considerations for enzymes that cut only one sequence or all sequences minus one.

2.2 Microorganisms used for the script assay

To evaluate the performance and reliability of this virtual bioinformatic tool, validation studies were carried out using partial or complete genomes of diverse microorganisms collected from GenBank https://www.ncbi.nlm.nih.gov/. Case studies included the 16S ribosomal RNA gene, partial sequences of 16 bacteria, namely *Escherichia marmotae* strain HT073016, *Escherichia ruysiae* strain OPT1704, *Escherichia vulneris* strain NBRC 102420, *Escherichia coli* strain NBRC 102203, *Escherichia fergusonii* ATCC 35469, *Escherichia albertii* strain LMG 20976TNBRC 107761, *Leclercia adecarboxylata* strain CIP 82.92, *Escherichia blattae* DSM 4481(NBRC 105725 = CIP_104942), *Escherichia hermannii* strain CIP 103176, *Salmonella typhimurium*, variant C37, variant C79, variant C11, variant C85, variant C170, variant C89 and variant C9, the complete genome of *SARS*-related *Coronavirus*, strain BtKY72, isolate SC/L75.18/2021, isolate GD/L18.18/2020, isolate GD/L17.18/2020, isolate GD/L19.18/2020, isolate BatCoV_B20-50, isolate Rs56, isolate Rs7952, isolate Ra7909, isolate Rs7907, isolate Rs7905, isolate Rs7896 and mitochondrion, complete genomes of *Trichodermalixii* MUT3171, *Trichoderma atroviride* strain ATCC 26799, *Trichoderma gamsii* strain KUC1747, *Trichoderma virens* strain Gv29-8, *Trichoderma koningiopsis* strain POS7, *Trichoderma simmonsii* strain GH-Sj1, and *Trichoderma harzianum* strain CBS 226.95 (Table 1).

2.3 Analytical validation of RFLP-kenzy

For analytical validation of the script, sequences from 1263 to 30,146 bp (Table 1) were analyzed by all the library restriction enzymes. After analysis, the results are provided as a list of key enzymes.

3 Results

3.1 RFLP-kenzy and output results description

RFLP-kenzy is a new bioinformatic tool developed to facilitate the selection of appropriate restriction enzymes that allow rapid distinction between sequences. By using RFLP-kenzy, key enzymes that cut only one sequence or all sequences minus one are shown on an individual tab as _file_final_forward.txt. Once the analysis is complete, the results including the analyzed sequence, enzymes, and the cut site information are also provided in an output file as previously described.

3.2 Analytical validation of RFLP-kenzy

The initial validation of the RFLP-kenzy tool revealed different restriction key enzymes able to differentiate 16 bacterial genomes collected from NCBI. When considering the case study of *Escherichia* species, several restriction enzymes including PflPt14I, Bco11035III, Mch946II, Cdi13746V, Cko11077IV, Cal14237I, BseRI, MunI, MfeI, Van9116I, TpyTP2I, PspPRI, Eco9699II, Bve1B23I, BtgZI, and RdeGBII were selected as key enzymes because each one recognizes only one restriction site and cut just one sequence as indicated in the supplementary file (S1). For example, unique RFLP patterns for *Escherichia marmotae* HT073016 (Gen Bank code KJ787692.1) could be obtained using the enzyme PflPt14I (A|G)GCCCAC) which cuts at position: [262], or the enzyme Bco11035III (GAAGC(C|T) which cuts at position: [76], or the enzyme Mch946II (A|G|C)CGATCT which cuts at position: [276], or either the enzyme Cdi13746V (A|G)GAAAG(A|G) which cuts at position: [71]. The same thing is also applicable to the enzyme Cko11077IV (TGACAG) which only cuts the sequence of the strain *Escherichia vulneris* NBRC 102420 (Gen Bank code AB681776.1) at position [641] or to the enzymes Cal14237I (GGTTAG), BseRI (GAGGAG), and MunI /MfeI (CAATTG) that only recognize the respective specific sites at positions [1130], [841], and [473] in the sequence of *Leclercia adecarboxylata* CIP 82.92 (Gen bank code JN175338.1).

For this same case study, other key enzymes which cut more than one sequence and could be used to separate these species are also provided by the RFLP-Kenzy tool (S1). For example, some key enzymes can cut all sequences at specific sites except for one sequence. In such case, species with this particular sequence could be quickly separated from the rest of the studied *Escherichia* group as is the case for *Escherichia blattae* DSM 4481 and for *Escherichia albertii* LMG 20976T that do not contain any restriction sites for the following enzymes Sma325I, Hpy300XI, CstMI, SalI, Pac19842II or XmaJI, AvrII, AspA2I, and BlnI, respectively (S1). The results also showed that key enzymes such as Hca13221V, HbaII, Sbo46I, or Hca13221V could be used to differentiate both of *Escherichia coli* NBRC 102203, *Escherichia hermannii* strain CIP 103176, *Escherichia fergusonii* ATCC 35469, and *Leclercia adecarboxylata* CIP 82.92, respectively (S1).

In the initial validation of RFLP-kenzy tool, the restriction patterns of some *Salmonella typhimurium* variants were further analyzed. Results reveal here also several key enzymes able to cut just one sequence such as Rtr1953I, Spe19205IV, Cko11077IV, MaqI, NgoAVII, and Bsp3004IV (S2). By using the key enzymes mentioned above, *Salmonella typhimurium* variant C85, variant C170, and variant C9 could rapidly discriminate each one from the other and from the rest of the *Salmonella* species. On the other hand, *Salmonella typhimurium* variant

Table 1 Species used for the analytical validation of RFLP-kenzy

Species	Gen Bank	Sequences length
Escherichia marmotae HT073016	KJ787692.1	1504 bp
Escherichia ruysiae OPT1704	LR745848.1	1538 bp
Escherichia vulneris NBRC 102420	AB681776.1	1465 bp
Leclercia adecarboxylata CIP 82.92	JN175338.1	1527 bp
Escherichia albertii LMG 20976T	AJ508775.1	1494 bp
Escherichia blattae DSM 4481	JN175333.1	1525 bp
Escherichia coli NBRC 102203	AB681728.1	1467 bp
Escherichia fergusonii ATCC 35469	AF530475.1	1473 bp
Escherichia hermannii strain CIP 103176	JN175345.1	1478 bp
Salmonella typhimurium variant C37	EF057784.1	1323 bp
Salmonella typhimurium variant C79	EF057787.1	1263 bp
Salmonella typhimurium variant C11	EF057786.1	1281 bp
Salmonella typhimurium variant C85	EF057785.1	1302 bp
Salmonella typhimurium variant C170	EF057783.1	1326 bp
Salmonella typhimurium variant C89	EF057782.1	1341 bp
Salmonella typhimurium variant C9	EF057781.1	1347 bp
SARS-related coronavirus isolate SC/L75.18/2021	OQ297704.1	29,274 bp
SARS-related coronavirus isolate GD/L18.18/2020	OQ297703.1	29,745 bp
SARS-related *coronavirus* isolate GD/L17.18/2020	OQ297702.1	29,720 bp
SARS-related *coronavirus* isolate GD/L19.18/2020	OQ297701.1	29,879 bp
SARS-related *coronavirus* isolate BatCoV_B20-50	ON378802.1	29,612 bp
SARS-related *coronavirus* isolate Rs56	MW681002.1	30,146 bp
SARS-related *coronavirus* isolate Rs7952	OL674081.1	29,525 bp
SARS-related *coronavirus* isolate Ra7909	OL674077.1	29,803 bp
SARS-related *coronavirus* isolate Rs7907	OL674076.1	29780bp
SARS-related *coronavirus* isolate Rs7905	OL674075.1	29,676 bp
SARS-related *coronavirus* isolate Rs7896	OL674074.1	29,385 bp
Trichoderma lixii strain MUT3171 mitochondrion	NC_052832.1	29,791 bp
Trichoderma atroviride strain ATCC 26799 mitochondrion	MN125601.1	32,758 bp
Trichoderma gamsii strain KUC1747 mitochondrion, complete genome	NC_030218.1	29,303 bp
Trichoderma virens strain Gv29-8 mitochondrion, complete genome	CP071114.1	27,943 bp
Trichoderma koningiopsis strain POS7 mitochondrion, complete genome	MT816499.1	27,560 bp
Trichoderma simmonsii strain GH-Sj1 mitochondrion, complete genome	NC_063562.1	28,668 bp
Trichoderma harzianum CBS 226.95 mitochondrion, complete genome	MN564945.1	27,632 bp

C170 or variant C11 could also be rapidly selected based on the Sba460II, BtgZI, or Pcr308II restriction patterns, respectively, because these two species do not contain any restriction sites for these enzymes (S2).

The validation of RFLP-kenzy tool included also the analysis of 11 complete genome of the *SARS*-related *Coronavirus*. As shown in the supplementary file (S3), restriction endonucleases able to cut from one to 10 sequences are selected as key enzymes. For example, PspOMII CGCCCA(A|G) and Sst E37I CGAAGA C which cut only the isolate Rs7907 and the isolate SC/L75.18/2021 at positions [271] and [18442], respectively, allow the rapid distinction of these two isolates from the rest of the studied *Coronavirus* (S3). Also, BatCoV_B20-50 isolate could be easily differentiated from the other *Coronavirus* (S3) based on Sse8387I CCTGCAGG, SdaI CCTGCAGG, PliMI CGCCGAC, or SbfI CCTGCAGG RFLP patterns because these endonucleases cut the BatCoV_B20-50 genome sequence at the unique positions of [12334], [12334], [28637], and [12334], respectively (S3). In addition, other key enzymes able to cut from two to 10 sequences are also provided in the supplementary file(S3) proving the usability of this new bioinformatic tool in selecting key enzymes able to differentiate these closely related *Coronavirus* isolates.

In the last example, the analysis of the restriction profile of 7 *Trichoderma mitochondrion* complete genome was also conducted for the validation of RFLP-kenzy tool. Results showed that different key enzymes are designed in the supplementary file (S4). Among these endonucleases, numerous key enzymes cut only one sequence as is the case for XmaIII, BstZI, BseX3I, RspPBTS2III, SstE37I, Sth20745III, EclXI, RdeGBI, Eco52I, EagI, GdiII, and UbaF13I, which allow to distinguish easily the *Trichoderma atroviride* ATCC 26799 strain from the other *Trichoderma* species. The following enzymes such as Van91I, Eco31I, Pae10662III, AccB7I, Bso31I, FspAI, SacII, PflMI, KspI, Sfr303I, BsaI, Cfr42I, BspTNI, and SgrBI or enzymes as HdeNY26I, RpaB5I, AquII cleave only the *Trichoderma gamsii* strain KUC1747or the *Trichoderma virens* strain Gv29-8 allowing specifically the differentiation of these two strains, respectively (S4).In addition, other enzymes such as TaqII, Nbr128II, BseYI, PspFI, Bpu10I, and GsaI or as SpoDI which cut, respectively *Trichoderma koningiopsis* strain *POS7* or *Trichoderma simmonsii* strain *GH-Sj1* provide in silico restriction patterns to separate with efficiency both these strains (S4). As mentioned earlier, RFLP-kenzy tool reveals other restriction enzymes capable to cut more than one sequence, allowing thereby the separation of the uncut sequences (S4).

3.3 Comparison with other tools

To highlight the advantages of the proposed tool, the performance of RFLP-kenzy was compared with other existing virtual programs including Pdraw 32, REDiges, NEBcutter, and C*is*SERS [7, 8] tools. Unlike the previous tools, RFLP-kenzy allows the user input multiple genes or complete genome at the same time for comparatives studies. Also, by using RFLP-kenzy, the entire list of restriction enzymes from the rebase database is used for the analysis, which in this way offers the possibility of evaluating all restriction enzymes. Furthermore, RFLP-kenzy tool is not a web server-based program an advantage for high-throughput analysis that requires high internet connection and server availability which is a limit in the case of web server tools. Another advantage of RFLP-kenzy is that details of the cut site's positions, the analyzed sequence, and the restriction enzymes are also provided in addition to the key enzymes lists. With the RFLP-kenzy tool, any restriction enzyme could be added to the enzyme_list_file.txt allowing assessment of tagged endonucleases. In addition, this script runs locally on the IDE editor and could run on a notebook like Colab or Jupyter. Also, RFLP-kenzy does not require the installation of the Java Virtual Machine (JVM) as is the case for many other tools.

The script of RFLP-kenzy is available for free in the Supplementary file (S5).

4 Discussion

RFLP analysis is a method that examines DNA sequence variations by comparing the patterns of DNA fragments generated through restriction enzyme cleavage. However, one of the main challenges of RFLP techniques is to identify the appropriate restriction enzymes with specific recognition sites in particular to differentiate between the closely similar sequences with minor variations [9, 10]. This is the reason why several in silico tools have been developed in the past for the virtual analysis of the RFLP patterns, but each method has its own advantages and limitations as described above [8]. Based on our previous study in which we proved that too closely related species from lactobacilli group, with more than 99% of 16S rRNA gene sequences similarities, could be separated only on basis of their 16S rRNA RFLP patterns by using key enzymes which cut 1 sequence in the minimum and n-1 sequence in the maximum [11]. In the present work we propose, a free virtual RFLP tool that allows the selection of all key enzymes that cut at least 1 sequence and at most n-1 sequences. This tool provides the possibility of analyzing simultaneously the RFLP of multiples sequences using all restriction endonucleases available in the Bio.Restriction library. After the analysis, all restriction key enzymes are then listed in a list text.

Moreover, the results of the RFLP-kenzy validation demonstrated the effectiveness of this newly tool in selecting the appropriate restriction enzymes allowing the rapid distinction between closely related isolates like *Salmonella Typhi* variants that are usually reported as difficulty discerned by using restriction enzymes [12]. In such cases, this tool could be very useful in short-term epidemiological surveillance of typhoid fever for example. Also, results showed that several restriction enzymes are selected as the key ones by using RFLP-kenzy and the whole genome of SARS *Coronavirus* strains. These enzymes could rapidly distinguish between new variants in this ribovirus group that accumulate mutations without any correction systems [13, 14]. With the rapid emergence of *Coronavirus* variants, the development of simple and accurate tools is very important in particular to detect and track mutations [15]. Furthermore, the complete mitochondrion genome analysis of *Trichoderma* species by RFLP-kenzy tool also provides many key enzymes that could be very useful for rapid identification and selection of species in this fungal group with biocontrol traits for preventing diseases in plant [16].

5 Conclusion

In this work, we provide a free, simple, and new virtual restriction tool called RFLP-kenzy. Through the in silico analysis of the restriction patterns of four examples, we demonstrated that RFLP- kenzy is a useful and easy tool for identifying key enzymes that allow fast separation of partial or complete genomes of eukaryotic, prokaryotic, or viral organisms. By using this virtual new tool, restriction enzymes digestion patterns are simulated, and the appropriate enzymes are designed without traditional laboratory experiments, helping in this way researchers to save time and resources. For further work, we would like to increase the capacity of the RFLP-kenzy tool to analyze large sized sequences of more than 20 mega bp.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43088-024-00531-8.

Additional file 1
Additional file 2
Additional file 3
Additional file 4
Additional file 5
Additional file 6
Additional file 7
Additional file 8
Additional file 9
Additional file 10

Acknowledgements
Not applicable.

Author contributions
LAREF and BELKHEIR were involved in the idea conception. BELAZREG and HIRECHE helped in RFLP-kenzy designing. LAREF and BELAZREG contributed to sequences analysis. BELKHEIR and LAREF helped in results analysis. BELKHEIR and LAREF contributed to draft and manuscript writing. All authors approved the final version of the manuscript.

Funding
No funding for this manuscript.

Availability of data and material
All the data are reported in the manuscript.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interest.

Received: 29 March 2024 Accepted: 15 July 2024
Published online: 02 September 2024

References

1. Grodzicker T, Anderson C, Sharp PA, Sambrook J (1974) Conditional lethal mutants of adenovirus 2-simian virus 40 hybrids I. Host range mutants of Ad2+ND1. J Virol 13:1237–1244
2. DI Felice F, Micheli G, Camilloni G (2019) Restriction enzymes and their use in molecular biology: an overview. J Biosci 44:38
3. Yang W, Kang X, Yang Q, Lin Y, Fang M (2013) Review on the development of genotyping methods for assessing farm animal diversity. J Anim Sci Biotechnol 4:2
4. Chen X, Luo C, Ma X, Chen M (2009) VIRS: a visual tool for identifying restriction sites in multiple DNA sequences. Biotechnol Prog 25:1525–1527
5. Cheng YH, Liaw JJ, Kuo C (2018) REHUNT: a reliable and open source package for restriction enzyme hunting. BMC Bioinform 19:178
6. Szubert J, Reiff C, Thorburn A, Singh BK (2007) REMA: a computer-based mapping tool for analysis of restriction sites in multiple DNA sequences. J Microbiol Methods 69:411–413
7. Sharpe RM, Koepke T, Harper A, Grimes J, Galli M, Satoh-Cruz M (2016) CisSERS: customizable in silico sequence evaluation for restriction sites. PLoS ONE 11:e0152404
8. Vincze T, Posfai J, Roberts RJ (2003) NEBcutter: a program to cleave DNA with restriction enzymes. Nucleic Acids Res 31:3688–3691
9. Chen L, Teasdale MT, Kaczmarczyk MM, Freund GG, Miller MJ (2012) Development of a lactobacillus specific T-RFLP method to determine lactobacilli diversity in complex samples. J Microbiol Methods 91:262–268
10. Huang CH, Lee FL, Liou JS (2010) Rapid discrimination and classification of the Lactobacillus plantarum group based on a partial dnaK sequence and DNA fingerprinting techniques. Antonie Van Leeuwenhoek 97:289–296
11. Laref N, Belkheir K (2022) Application of 16S rRNA virtual RFLP for the discrimination of some closely taxonomic-related lactobacilli species. J Genet Eng Biotechnol 20:167
12. Tien YY, Ushijima H, Mizuguchi M, Liang SY, Chiou CS (2012) Use of multilocus variable-number tandem repeat analysis in molecular subtyping of Salmonella enterica serovar Typhi isolates. J Med Microbiol 61:223–232
13. Markov PV, Mahan G, Beer M (2023) The evolution of SARS-CoV-2. Nat Rev Microbiol 21:361–379
14. Belkheir K, Laref N (2024) The inhibitory effect of vitamins on Omicron virus via targeting the ACE2 receptor. In silico analysis. J King Saud Univ Sci 36:103082
15. Ratcliff J et al (2022) Highly sensitive lineage discrimination of SARS-CoV-2 variants through allele-specific probe PCR. J Clin Microbiol 60:1–14
16. Guzmán-Guzmán P et al (2023) Trichoderma species: our best fungal allies in the biocontrol of plant diseases-A review. Plants 12:432

Publisher’s Note