Bioinformatics Tools and Moringa Genomics

12

Saranya Nallusamy and Jayakanthan Mannu

Abstract

Bioinformatics is an interdisciplinary field of biology, computer science, and mathematics. The advancement of high throughput genomics and proteomics technologies has produced large volume of genomics and proteomics data, which can be accessible from open databases. Exploring this big data could resolve many of the biological complexity. In this chapter, we discussed the role of bioinformatics in analyzing the genomics and proteomics data of Moringa.

12.1 Introduction

Moringa oleifera is a miracle tree in the ecosystem which is also called as drumstick tree/ben oil tree/horse radish tree or simply moringa. It serves as food (super food) with enriched nutritional value, livestock hood, agricultural applications, and possesses enormous medicinal properties (Matic et al. 2018). As an indigenous plant with Indian origin, it has been used as traditional medicine over centuries (Fahey 2005). All the plant parts like flower, leaf, root, seed, and stem are edible, and its phytochemicals and crude extract were proven to have antioxidant, anticancer, anti-inflammatory, antidiabetic, and antimicrobial activities against more than 300 human diseases (Anwar et al. 2007; Goyal et al. 2007).

With the advancement of high throughput omics technology, genomics and transcriptomics data of *Moringa oleifera* are made available in the public repositories. Bioinformatics studies of gene prediction, annotation, and pathway analysis provides insight on the biology of genes and their interaction mechanism that are required to understand the key biological and metabolic functions in moringa. Orthology analysis has shown its close evolutionary phylogenetic relation with *Carica papaya*, *Theobroma cocoa, Arabidopsis thaliana,* and *Vitis vinifera* (Tian et al. 2015; Pasha et al. 2020) compared to other species in viridiplantae kingdom.

Comparative genomics and transcriptomics studies provide massive knowledge on the similarities and differences between different plant species in the ecosystem. Moreover, biological databases and software are the key players that pave the scientists to explore new aspects in the way of metabolic and genetic engineering of moringa to produce value-added products in huge volumes. Apart from the huge medicinal value, it is also interesting to know the application of moringa gum in calico printing.

S. Nallusamy (✉) · J. Mannu
Department of Plant Molecular Biology and Bioinformatics, Centre for Plant Molecular Biology & Biotechnology, Tamil Nadu Agricultural University, Coimbatore, India
e-mail: saranya.n@tnau.ac.in

N. M. Boopathi et al. (eds.), *The Moringa Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-80956-0_12

Moringa oleifera being a reservoir of medicinal and nutritional properties was not completely characterized at the molecular and physiology level. There exists a lot of space to explore the mechanism of function and interaction that prevails within the *Moringa oleifera* genome.

12.2 Bioinformatics tools and databases

Bioinformatics is the interdisciplinary science in which informatics is applied in any biological data by means of computational tools and databases that renders scientific community the ease of data storage, access, interpretation, and analysis. High-throughput sequencing of any plant genome produces huge volume of data. Genomic databases like NCBI–Genome (Sayers et al. 2019), plant genomic database (http://www.plantgdb.org), Gramene (http://www.gramene.org/), Phytozome (Goodstein et al. 2012), and Ensemble plants (https://plants.ensembl.org/index.html) provides the genome sequence and associated information of plants.

These databases also contain tools integrated with their server, so that anyone can browse for the information regarding genes, chromosomes, markers, restriction sites, function, pathway, etc., without any restriction to access. Availability of genomic resources is the key to acquire molecular level genetic knowledge of an organism. Some of the mostly used tools and database information are provided in Table 12.1.

12.3 Pairwise Sequence Alignment

Most powerful tool that is being employed in bioinformatics is the Basic Local Alignment Search Tool (BLAST) (Johnson et al. 2008), which is a biological sequence comparison program based on the similarity. This tool compares the user-given biological sequences (nucleotide or protein sequences) to the sequences already stored in the databases and calculates the statistical significance of matches.

Match is found by pairwise aligning the user-given sequences to the database sequences. Information pertaining to function and evolutionary relationship could be retrieved by using this tool. BLOSUM (Blocks Substitution Matrix), PAM (Point Accepted Mutation) matrices are used for scoring the similarity between the sequences.

12.4 Multiple Sequence Alignment and Phylogenetic Tree Construction

To compare more than two sequences, unlike the pairwise similarity program, tools like CLUSTALW (Higgins and Sharp 1988), MAFFT (Katoh et al. 2019) are used. Multiple sequence alignment forms the basis for understanding the conserved and variable regions across the gene families. Other informations like motif, domain, signature, fingerprint, etc., are obtained by means of performing multiple sequence alignment.

Moreover, phylogenetic tree construction programs highly rely on the multiple sequence alignment programs to infer evolutionary relationship between the set of sequences. Programs like PHYLIP (Felsenstein 1993) and MEGA (Kumar et al. 2018) are mostly used for the phylogenetic tree construction.

Evolutionary information could be obtained in the form of gene tree or species tree based on the biological data available for study. Phylogenetic tree in simple can be compared with that of real tree which has branches, leaves, and root. Sequences with similar characteristics are grouped together in separate branches (internal node) and with the edge as the terminal node.

The phylogenetic tree can be rooted or unrooted depending on the ancestral sequence information availability. If ancestor or the primary sequences are present, there exists a root from which other sequences evolve by means of duplication or deletion or insertion or mutation. Methods such as clustering, Maximum likelihood, Maximum Parsimony, genetic algorithm, and Bayesian simulation are employed to construct the phylogenetic tree.

Table 12.1 List of prioritized bioinformatics databases and tools for *Moringa oleifera* genome and transcriptome analysis

S. no	Type	Databases/tools	Description	Weblink/References
1	Sequence databases and tools	**NCBI resources** Gene, Protein, Genome, GEO datasets, Bioproject, SRA experiment	NCBI provides public access to gene, protein, and genome and transcriptome data	https://www.ncbi.nlm.nih.gov/ (Sayers et al. 2019)
		Uniprot	Resource for the protein sequence and annotation data	https://www.uniprot.org/ (Bateman et al. 2020)
		Sequence Manipulation Suite	Tool serves the purpose of generating, formatting, and analyzing short DNA and protein sequences	https://www.bioinformatics.org/sms2/ (Stothard 2000)
		PFAM	Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models	https://pfam.xfam.org/ (Finn et al. 2014)
		Motif Scan	Tool to find all the motif in a sequence	https://myhits.sib.swiss/cgi-bin/motif_scan (Hau et al. 2007)
		InterPro	InterPro performs functional analysis of proteins by classifying them into families and predicting domains and important sites. InterPro makes use of protein signature information from several different databases	http://www.ebi.ac.uk/interpro/ (Finn et al. 2017)
2	Genomics/transcriptomics tools	Cufflink	Transcriptome data analysis package Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples	https://github.com/cole-trapnell-lab/cufflinks (Trapnell et al. 2012)
		SOAPdenovo2	Genome short read assembly program	https://github.com/aquaskyline/SOAPdenovo2 (Luo et al. 2015)
		OMICSBOX/Blast2GO	Genome Transcriptome and metagenomics data analysis package	Standalone commercial software (Götz et al. 2008; Workbench)
		CLC genomics workbench		
		HMMER	Hidden Markov Model can be employed to find the sequence similarity based on the profile of the aligned sequences. Build HMM profile can then be used to search large sequence databases to find related sequences, even those distantly related	http://hmmer.org/ (Eddy 2011)

(continued)

Table 12.1 (continued)

S. no	Type	Databases/tools	Description	Weblink/References
3	SSR Marker database and tools	Gramene SSR database	Plant SSR marker information	https://archive.gramene.org/markers/ (Tello-Ruiz et al. 2018)
		Krait	A robust and flexible tool for fast investigation of microsatellites in DNA sequences	https://github.com/lmdu/krait (Du et al. 2018)
		MISA	The MISA microsatellite finder (Thiel et al. 2003) is a tool for finding microsatellites in nucleotide sequences	MISA (Beier et al. 2017)
4	Phylogenetic analysis	PHYLIP	Phylogenetic tree construction software	https://evolution.genetics.washington.edu/phylip.html (Felsenstein 1993)
		MEGA (Molecular Evolutionary Genetics Analysis)	Enables the study of phylogenetic and evolutionary relationship in biological sequences. Includes statistical analysis packages	https://www.megasoftware.net/ (Kumar et al. 2018)
5	Sequence alignment	BLAST	Pairwise sequence alignment	https://blast.ncbi.nlm.nih.gov/ (Boratyn et al. 2013)
		CLUSTALW	Multiple sequence alignment	https://www.ebi.ac.uk/Tools/msa/clustalo/ (Higgins and Sharp 1988)
		MAFFT		https://mafft.cbrc.jp/alignment/server/ (Katoh et al. 2019)
6	Orthology analysis	Orthofinder	Software for analysis of phylogeny of orthogroups, gene duplication events in protein sequences	https://github.com/davidemms/OrthoFinder (Emms and Kelly 2019)
		Benchmarking Universal Single-Copy Orthologs (BUSCO)	Assessment of completeness of genome/transcritpome	http://busco.ezlab.org/ (Seppey et al. 2019)
		OrthoDB	Database of orthologs	(www.orthodb.org) (Kriventseva et al. 2019)
		OrthoMCL	Tool used for constructing orthologous groups across multiple eukaryotic taxa, using a Markov Cluster algorithm to group (putative) orthologs and paralogs	https://orthomcl.org/orthomcl/ (Li et al. 2003)

(continued)

Table 12.1 (continued)

S. no	Type	Databases/tools	Description	Weblink/References
7	Pathway database	KEGG pathway	Pathway maps representing the information of molecular interaction, reaction, and relation networks	https://www.genome.jp/kegg/pathway.html (Kanehisa et al. 2017)
		Reactome		https://reactome.org/ (Fabregat et al. 2018)
8	Protein–protein Interaction database and tools	StringDB	Database of known and predicted protein–protein interactions	https://string-db.org/ (Szklarczyk et al. 2020)
		Cytoscape	Software for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles	https://cytoscape.org/
9	Non-coding RNA databases and tools	Coding Potential Calculator (CPC2)	Assessment of the coding and non-coding ability of RNA transcripts	http://cpc2.cbi.pku.edu.cn/ (Kang et al. 2017)
		Coding Potential Assessment Tool (CPAT)		http://lilab.research.bcm.edu/cpat/ (Wang et al. 2013)
		RFAM	The Rfam database stores the information of RNA sequence families such as the structural RNAs, non-coding RNA genes, and cis-regulatory elements	http://rfam.xfam.org/ (Kalvari et al. 2020)
		PmiREN (Plant miRNA ENcyclopedia)	Functional information of plant miRNA	http://www.pmiren.com/
		Noncode	Collection of Plant non-coding sequences	http://www.noncode.org/ (Xiyuan et al. 2017)
		CANTATAdb		http://cantata.amu.edu.pl/ (Szcześniak et al. 2016)
		psRNATarget	Prediction of target mRNA sequence for miRNA	http://plantgrn.noble.org/psRNATarget/ (Dai and Zhao 2011)

12.5 Genomics Tools

Genome/Transcriptome sequencing results in the generation of raw reads of large volume. These reads exist as an input data for the plethora of bioinformatics tools to carry out the steps such as alignment, annotation, and analysis. Quality checking and trimming of reads are the primary step before start of the alignment/assembly of raw reads.

FastQC (FastQC 2015; Andrews 2010) is the tool used for the quality checking of the raw reads based on the parameters such as GC content, duplication level, length distribution, etc. Trimming of raw reads is mainly performed to remove adapter sequences that may occur by chance in the sequencing procedure for which software like Trimmomatic (Bolger et al. 2014) is mostly used.

Next step is the assembly step which can either be reference-based and de novo-based. If reference genome is available, then the former methodology is employed, otherwise the latter is used. A number of assembly algorithms like Trinity Grabherr (Grabherr et al. 2011), STAR (Dobin et al. 2013), and SOAPdenovo (Luo et al. 2015) software are used for alignment/assembly process.

After assembly, gene prediction programs like AUGUSTUS (Stanke and Morgenstern 2005), FGENESH (Salamov and Solovyev 2000), GENSCAN (Burge and Karlin 1997) are used for predicting the coding part of the genome. Annotation is performed by similarity search against sequences of closely related species using BLAST, MAKER (Cantarel et al. 2008), BlastKOALA, and GhostKOALA tools (Kanehisa et al. 2016).

In this procedure, gene ontology terms such as Molecular Function, Cellular component, and Biological process are mapped based on their similarity with the already annotated sequences. This analysis provides the knowledge on its function, pathway, and their interaction mechanism. Other downstream analysis like orthology, non-coding region annotation is performed in order to understand gene duplication events, divergence that has occurred during the evolutionary period of time, and to decode the regulatory role of non-coding part of the genome.

System biology and synthetic biology studies have joined hand with genomic studies (Jamil et al. 2020) to engineer the biological system at the metabolic level, genetic level, or protein level mainly to increase the production of metabolites of pharmaceutical and medicinal value.

Comparative genomics studies involve the comparison of whole genomes of two or more species to understand the intron–exon organization, gene structure, and similarity/dissimilarity/conservation observed among the distribution of gene families. BLAST tool could be used for comparison of genomes. Genome level alignment and comparison also provides information on evolutionary events like speciation, duplication, and horizontal gene transfer events.

12.6 Moringa Genomics

First draft genome sequence of *Moringa oleifera Lam.* was first published in the year 2015 (Tian et al. 2015) from Yunnan Agricultural University, China. 19,465 annotated protein-coding genes were predicted from the 457× coverage DNA sequencing data for the *Moringa oleifera* sample. Based on the 17-mer frequency distribution, the estimated genome size was 315 Mb. Clustering analysis to understand the distribution of gene families among other plant species such as *Vitis vinifera*, *Cajanus cajan*, *Carica papaya,* and *Malus domestica* showed that these five different plant species possess similar numbers of gene families, with a core set of 10,215 shared genes. 198 single-copy gene families were found among the 12, 298 gene families reported in *Moringa oleifera*.

Second draft genome, as well as transcriptome data, was published by Chang et al. (2019) using *Moringa oleifera* sample grown at the World AgroForestry Center campus in Kenya. Genome size was observed to be 216.76 Mb. Transcription factors such as bHLH, NAC, ERF, MYB-related, C2H2, MYB, WRKY, bZIP, FAR1,

C3H, B3, G2-like, Trihelix, LBD, GRAS, M-type MADS, HDZIP, MIKC MADS, HSF, and GATA were found in abundance in comparison with other transcription factors.

Recently, RNA-sequencing enabled the analysis of gene expression samples from five different tissues (leaf, root, stem, seed, and flower) of *Moringa oleifera* plant (Bhagya variety) at the University of Agricultural Sciences, GKVK, Bangalore, India (Pasha et al. 2020). Pathway analysis was performed to understand the biosynthesis of secondary metabolites such as quercetin, kaempferol, benzylamine, and ursolic/oleanolic acid synthesized by *Moringa oleifera* genes which have profound medicinal values.

Being a drought-tolerant plant, stress-related transcription factors and enzymes related to production of metabolites of medicinal value were mostly expressed in the *Moringa oleifera* transcriptome leaf analysis. For example, pathway analysis showed the involvement of seven enzymes such as 4-Coumarate-CoA ligase (4CL), Chalcone synthase (CHS), Chalcone flavone isomerase (CHI), Flavonone 3-hydroxylase (F3H), Flavonol synthase (FLS), Tricin synthase (OMT), and Flavonoid 3′-monooxygenase (F3'H) in biosynthesis of anti-cancerous compound Quercetin.

Chloroplast genome of *Moringa oleifera* reported 131 genes (Lin et al. 2019) and was found to have a length of 160,600 bp with a large single-copy (LSC) region of 88,577 bp, a small single-copy (SSC) region of 18,883 bp, separated by two inverted repeat (IR) regions of 26,570 bp each. Phylogenetic analysis of 71 protein-coding sequences of 13 plant plastomes showed that *Moringa oleifera* is closest to *Carica papaya*.

WRKY transcription factors are well known for their role in plant development, signal transduction, and stress responses (Zhang et al. 2019). This gene family has been characterized in a genomic scale in *Moringa oleifera* through bioinformatics tools through the analysis of gene structures, motif analysis, conserved motifs, and phylogenetic tree construction. Fifty-four WRKY genes were identified through HMM profile search performed using WRKY domains.

Phylogenetic analysis showed its close relation with *Arabidopsis thaliana*. Also, analysis of commonly occurring cis-acting elements in WRKY promoter regions reported hormone responsive elements (ABRE, CGTCA motif, and TGACG motif), a drought stress responsive element (MBS), a heat stress responsive element (HSE), and four light responsive elements (Sp1, Box 4, G box, and GT1 motif), respectively.

Further expression profiling of WRKY genes reported its significance in various abiotic stress conditions such as under drought, salt, cold, and heat stresses. In a similar study of genome-wide analysis of trehalose-6-phosphate synthase (TPS) family, Group II *Moringa oleifera TPS* genes have evolved under relaxed purifying selection or positive selection. Further, group I *TPS* genes closely relate to reproductive development, and Group II *TPS* genes closely relate to high temperature resistance in leaves, stem, stem tip, and roots. Expression pattern of WRKY genes and TPS genes is experimentally validated using reverse transcription polymerase chain reaction (RT-PCR) and quantitative RT-PCR (qRT-PCR) experiments under different stress conditions.

12.7 Computational Identification of *Moringa oleifera* miRNA

Highly enriched nutritive and medicinal value of miracle tree is the repository of bioactive phytochemicals which has the potential to improve the health conditions of prevailing malnutrition observed among children from poorer section of the society and also pregnant women. Understanding the regulatory role of plant is very

important to acquire the knowledge on the production of various phytochemicals under different stress conditions. MicroRNA (miRNA) is a small non-coding RNA of length about 22 nucleotides having an important role in regulating the gene expression at both post-transcriptional and translational level.

Moringa leaves and cold stressed callus (Pirrò et al. 2016) are characterized for the presence of conserved and novel microRNA families through RNA sequencing technology. Analysis using miRBase database and miRDeep2 tool predicted 431 conserved and 392 novel microRNAs, and it was confirmed using qRT-PCR analysis. Among the reported microRNAs, microRNA159 was majorly observed in leaf and callus, respectively. These miRNAs majorly targets the transcription factor that controls the plant growth, reproduction, and stress response.

Furthermore, plant microvesicles (MVs) possess similar features with that of mammalian exosomes, which are involved in cell–cell communication and miRNA transporters.

Moringa oleifera has shown enormous medicinal and pharmaceutical properties in a number of human diseases. In this context, ingestion of plant miRNA has shown regulation of gene activity in human which makes it a remarkable bioactive constituent for treating a number of human diseases. High-throughput sequencing of moringa seeds has reported miRNAs that has the potential to regulate the human gene expression at the post-transcriptional level (Pirrò et al. 2016). This regulation has the impact on exerting medicinal activity in a number of human diseases.

In silico analysis of miRNAs of moringa seeds for their contribution to medicinal value using MirCompare and combinatorial miRNA target prediction (COMIR) web tool identified *Moringa oleifera*–miR168a as a potential candidate for regulation of human genes. These genes are majorly involved in the cell–cycle regulation and p53 pathway. It is interesting that *SIRT1* (*Sirtuin*) gene was positively regulated by miR168a which was confirmed by transfection experiment.

12.8 Computational Screening of Potential Bioactive Compounds from *Moringa oleifera*

Computational screening of potential bioactive compounds against various biological protein targets for different diseases are gaining importance as complement to traditional drug design and discovery process nowadays. Success of this method depends on the identification of valid protein targets from the genomic region of the organism.

Complete genome sequence of the *Moringa oleifera* (Tian et al. 2018) gives the possibility of exploring various biological targets for drug design and discovery to treat various diseases. The minimum requirement of inputs for computational screening of chemical compounds is availability of three-dimensional structure of the target protein. This structure could be retrieved from protein structure database called Protein Data Bank (PDB) (www.rcsb.org), which contains experimentally determined three-dimensional structures of the proteins.

Comparative homology modeling approach could help us to model the three-dimensional structure of the protein in the case of non-availability of experimental structures. Similarly, structure of phytochemicals can be retrieved from PubChem (https://pubchem.ncbi.nlm.nih.gov) database, which is an open database of chemical compounds maintained by National Institute of Health.

In recent years, many of the phytochemicals of *Moringa oleifera* have been virtually screened against various disease targets such as Diabetes Mellitus, different cancers, hypertension, COVID-19, antimicrobial activity, antioxidant defense systems, and HIV.

12.8.1 Diabetes Mellitus

Phytochemicals of *Moringa oleifera* such as anthraquinone, 2-phenylchromenylium (Anthocyanins), hemlock tannin, sitogluside (glycoside),

and A-phenolic steroid were reported as potential therapeutic agents against mutated insulin receptor using molecular docking approach (Zainab et al. 2020).

The strategy of toxicity screening, checking for Lipinski rule violation, and pharmacophore generation has been carried out along with docking study for the phytochemicals of *Moringa oleifera* in their study. Similarly, Yang et al. (2014) have performed virtual screening, docking, ADME prediction, and in-vitro analysis for the phytochemicals of *Moringa oleifera* to identify potential compounds for diabetics.

A total of 111 phytochemicals were screened in their analysis against Potential Dipeptidyl Peptidase (DPP)-IV, and it was reported O-Ethyl-4-[(α-L-rhamnosyloxy)-benzyl] carbamate has the activity with half-maximal inhibitory concentration [IC50] = 798 nM.

12.8.2 Carcinoma

Crude ethanolic extract (HF-CEE) of *Moringa oleifera* seeds reported for their inhibitory action on MCF7 breast cancer cell growth (Mansour et al. 2019). Methylelaidate, a phytochemical of *Moringa oleifera,* was tested for their binding efficiency with Bax and MDM2 apoptotic proteins using AutoDock. This showed a stable interaction with the target proteins and also has the potential of drug-like properties according to the Lipinski's rule of 5 Adebayo (Adebayo et al. 2018).

Another phytochemical, quinicacid, was reported for their better pharmacokinetic properties and suitable for further drug discovery and development cycle to control prostate cancer cell growth (Inbathamizh and Padmini 2013).

12.8.3 Hypertension

Some of the phytochemicals of *Moringa oleifera* such as Niazicin-A, Niazimin-A, and Niaziminin-B were tested for their binding efficiency with Angiotensin-converting enzyme (ACE) using AutoDockVina and for pharmacokinetics activity using ADME-Toxicity prediction. The above compounds showed good binding energy compared to the reference drug molecules such as Captopril and Enalapril (Khan et al. 2019; Aktar et al. 2019). They have reported that leaves methanolic extract (MOLME) of *Moringa oleifera* showed inhibitory activity against Angiotensin-converting enzyme using spectrophotometric method. It is also reported that substrate hippuryl-L-histidyl-L-leucine (HHL) inhibits ACE with an IC50 value of 226.37 μg/ml, in comparison to reference compound, captropril, which shows IC50 value of 0.0289 μM.

12.8.4 COVID-19

Currently, there is no potential drug or vaccine developed to target SARS-CoV-2 virus. It is mandatory to identify a drug to handle the pandemic situation prevailing worldwide. The multiple protein targets of SARS-CoV-2 could be accessible from publicly available genomic and proteomic databases. Phytochemical compounds of *Moringa oleifera* were virtually screened using AutoDockVinato to discover novel lead compounds against main protease (Mpro) and RNA-dependent RNA polymerase (RdRp) to treat COVID-19.

The compounds were also screened for drug-likeness properties using Swiss ADME. The scientists have reported that the compounds kaempferol, pterygospermin, morphine, and quercetin made a stable interaction with Mpro and RdRp target proteins. These compounds could be taken as potential lead molecules for further evaluation to treat COVID-19 (Shaji 2020).

12.9 Future Prospects and Applications

In the present scenario of increase in the pandemic and life threating diseases, every human is in need of super food like *Moringa oleifera* which eradicates malnutrition as well as increase

immunity and longevity of life. It is very important to understand the molecular mechanism of biomolecules and their pathways to have future success stories on metabolomic and genetic engineering which will help in the production of value-added products in the pharmaceutical research. Enormous research studies are required to completely link the genes, proteins, metabolites, non-coding genes, and their interaction mechanisms in a concurrent way.

References

Adebayo IA, Arsad H, Samian MRJPM (2018) Methyl elaidate: A major compound of potential anticancer extract of Moringa oleifera seeds binds with bax and MDM2 (p53 inhibitor) In silico. Pharmacogn Mag 14 (59):554

Aktar S, Das PK, Asha SY, Siddika MA, Islam F, Khanam JA, Rakib MA (2019) Moringa oleifera leaves methanolic extract inhibits angiotensin converting enzyme activity in vitro which ameliorates hypertension. J Adv Biotechnol Exp Ther 2(2):73–77

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Anwar F, Latif S, Ashraf M, Gilani AH (2007) Moringa oleifera: a food plant with multiple medicinal uses. Phytother Res Int J Devoted Pharmacol Toxicol Eval Nat Prod Deriv 21(1):17–25

Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bursteinas B (2020) UniProt: the universal protein knowledgebase in 2021. Nucl Acids Res 49 (D1):D480–D489

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33(16):2583–2585

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y (2013) BLAST: a more efficient report with usability improvements. Nucl Acids Res 41(W1):W29–W33

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268 (1):78–94

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18(1):188–196

Chang Y, Liu H, Liu M, Liao X, Sahu SK, Fu Y, Song B, Cheng S, Kariba R, Muthemba S and Hendre PS (2019) The draft genomes of five agriculturally important African orphan crops GigaScience 8 (3):p. giy152.

Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucl Acids Res 39(suppl_2): W155–W159

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15–21

Du L, Zhang C, Liu Q, Zhang X, Yue B (2018) Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. Bioinformatics 34(4):681–683

Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011 Oct;7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20(1):1–14

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B (2018) The reactome pathway knowledgebase. Nucl Acids Res 46(D1):D649–D655

Fahey JW (2005) Moringa oleifera: a review of the medical evidence for its nutritional, therapeutic, and prophylactic properties. Part 1. Trees Life J 1(5):1–15

FastQC (2015). https://qubeshub.org/resources/fastqc

Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.5 c. Joseph Felsenstein

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang H-Y, Dosztányi Z, El-Gebali S, Fraser M (2017) InterPro in 2017—beyond protein family and domain annotations Nucl Acids Res 45 (D1):D190–D199

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J (2014) Pfam: the protein families database. Nucl Acids Res 42(D1):D222–D230

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N (2012) Phytozome: a comparative platform for green plant genomics. Nucl Acids Res 40(D1): D1178–D1186

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucl Acids Res 36(10):3420–3435

Goyal BR, Agrawal BB, Goyal RK, Mehta AA (2007) Phyto-pharmacology of Moringa oleifera Lam.—an overview. Nat Prod Radiance 6 (4)

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652

Hau J, Muller M, Pagni M (2007) HitKeeper, a generic software package for hit list management. Source Code Biol Med 2(1):1–8

Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73(1):237–244
https://doi.org/10.1007/s00122-002-1031-0

Inbathamizh L, Padmini E (2013) Quinic acid as a potent drug candidate for prostate cancer–a comparative pharmacokinetic approach. Asian J Pharm Clin Res 6(4):106–112

Jamil IN, Remali J, Azizan KA, Muhammad NAN, Arita M, Goh H-H, Aizat WM (2020) Systematic multi-omics integration (MOI) approach in plant systems biology. Front Plant Sci 11

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucl Acids Res 36(suppl_2): W5–W9

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, Rivas E (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families Nucleic Acids Research49(D1):D192-200

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucl Acids Res 45(D1):D353–D361

Kanehisa M, Sato Y, Morishima K (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428(4):726–731

Kang Y-J, Yang D-C, Kong L, Hou M, Meng Y-Q, Wei L, Gao G (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucl Acids Res 45(W1):W12–W16

Katoh K, Rozewicki J, Yamada KD (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20(4):1160–1166. https://doi.org/10.1093/bib/bbx108

Khan H, Jaiswal V, Kulshreshtha S, Khan A (2019) Potential angiotensin converting enzyme inhibitors from Moringa oleifera. Recent Pat Biotechnol 13 (3):239–248

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucl Acids Res 47(D1):D807–D811

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35 (6):1547–1549

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189

Lin W, Dai S, Chen Y, Zhou Y, Liu X (2019) The complete chloroplast genome sequence of Moringa oleifera Lam.(Moringaceae). Mitochondrial DNA Part B 4 (2):4094–4095

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y (2015) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 4(1):s13742-13015–10069-13742

Mansour M, Mohamed MF, Elhalwagi A, El-Itriby HA, Shawki HH, Abdelhamid IA (2019) Moringa peregrina leaves extracts induce apoptosis and cell cycle arrest of hepatocellular carcinoma. BioMed research international 1;2019.

Matic I, Guidi A, Kenzo M, Mattei M, Galgani A (2018) Investigation of medicinal plants traditionally used as dietary supplements: A review on *Moringa oleifera* J Public Health Afr. 9(3):841.

Pasha SN, Shafi KM, Joshi AG, Meenakshi I, Harini K, Mahita J, Sajeevan RS, Karpe SD, Ghosh P, Nitish S (2020) The transcriptome enables the identification of candidate genes behind medicinal value of Drumstick tree (Moringa oleifera). Genomics 112(1):621–628

Pirrò S, Zanella L, Kenzo M, Montesano C, Minutolo A, Potestà M, Sobze MS, Canini A, Cirilli M, Muleo R (2016) MicroRNA from Moringa oleifera: identification by high throughput sequencing and their potential contribution to plant medicinal value. PLoS One 11 (3):e0149495

QIAGEN CLC Genomics Workbench 20.0 (https://digitalinsights.qiagen.com/)

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10(4):516–522

Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, Connor R, Fiorini N, Funk K, Hefferon T, Holmes JB (2019) Database resources of the national center for biotechnology information Nucleic acids research.47(Database issue):D23.

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. In: Gene prediction. Springer, pp 227–245

Shaji D (2020) Computational Identification of Drug Lead Compounds for COVID-19 from Moringa Oleifera. ChemRxiv. Cambridge: Cambridge Open Engage; This content is a preprint and has not been peer-reviewed.

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucl Acids Res 33(suppl_2): W465-7.

Stothard P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28 (6):1102–1104

Szcześniak MW, Rosikiewicz W, Makałowska I (2016) CANTATAdb: a collection of plant long non-coding RNAs. Plant Cell Physiol 57(1):e8–e8

Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P

(2020) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucl Acids Res

Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, Wei S, Preece J, Geniza MJ, Jiao Y (2018) Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucl Acids Res 46(D1):D1181–D1189

Tian W, Chen C, Lei X, Zhao J, Liang J (2018) CASTp 3.0: computed atlas of surface topography of proteins. Nucl Acids Res 46(W1):W363–W367. https://doi.org/10.1093/nar/gky473

Tian Y, Zeng Y, Zhang J, Yang C, Yan L, Wang X, Shi C, Xie J, Dai T, Peng L, Huan YZ (2015) High quality reference genome of drumstick tree (Moringa oleifera Lam.), a potential perennial crop Science China Life Sciences 58(7):627-38.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucl Acids Res 41(6):e74–e74

Xiyuan L, Dechao B, Liang S, Yang W, Shuangsang F, Hui L, Haitao L, Chunlong L, Wenzheng F, Runsheng C and Yi Z (2017) Using the NONCODE database resource Current protocols in bioinformatics 58(1):12-16.

Yang X, Chen X, Bian G, Tu J, Xing Y, Wang Y, Chen Z (2014) Proteolytic processing, deubiquitinase and interferon antagonist activities of Middle East respiratory syndrome coronavirus papain-like protease. J Gen Virol 95(3):614–626. https://doi.org/10.1099/vir.0.059014-0

Zainab B, Ayaz Z, Alwahibi MS, Khan S, Rizwana H, Soliman DW, Alawaad A, Abbasi AM (2020) In-silico elucidation of Moringa oleifera phytochemicals against diabetes mellitus. Saudi J Biol Sci 27(9):2299–2307

Zhang J, Yang E, He Q, Lin M, Zhou W, Pian R, Chen X (2019) Genome-wide analysis of the WRKY gene family in drumstick (*Moringa oleifera* Lam.). PeerJ 7: e7063