COMPUTATIONAL GENOMICS AND STRUCTURAL BIOINFORMATICS IN MICROBIAL SCIENCE

EDITED BY: Saumya Patel, Dhaval K. Acharya and Mohammed Kuddus

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-625-5
DOI 10.3389/978-2-88974-625-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL GENOMICS AND STRUCTURAL BIOINFORMATICS IN MICROBIAL SCIENCE

Topic Editors:
Saumya Patel, Gujarat University, India
Dhaval K. Acharya, B N Patel Institute of Paramedical, India
Mohammed Kuddus, University of Hail, Saudi Arabia

Citation: Patel, S., Acharya, D. K., Kuddus, M., eds. (2022). Computational Genomics and Structural Bioinformatics in Microbial Science. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-625-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 11 February 2022
doi: 10.3389/fgene.2022.850397

Editorial: Computational Genomics and Structural Bioinformatics in Microbial Science

Dhaval Acharya[1], *Mohammed Kuddus*[2] *and Saumya Patel*[3]*

[1]Department of Microbiology, B. N. Patel Institute of Paramedical and Science, Gujarat, India, [2]College of Medicine, University of Hail, Hail, Saudi Arabia, [3]Department of Botany, Bioinformatics & Climate Change Impacts Management, School of Science, Gujarat University, Gujarat, India

Keywords: evolutionary and genomic microbiology, metagenomics, systems microbiology, microbiome data analytics, microbial bioinformatics

Editorial on the Research Topic

Computational Genomics and structural Bioinformatics in Microbial Science

Microbes play a crucial roles in the lives of hosts (plants, animals, humans) and in almost any environment one can think of. The goal of this Research Topic was to gather a collection of high-quality original papers on the general theme of Computational Genomics and Structural Bioinformatics in Microbial Science. This Research Topic collection from *Frontiers in Genetics* brings together 11 articles focused on computational analysis of genomic microbiology as well as computational analysis of nucleotide or amino acid sequences and structures from genomic and metagenomic data.

The first paper by Bharathi *et al.* provides new insight into the understanding of *Methanobrevibacter ruminantium* M1 (MRU) growth physiology and lifestyle in the ruminants, and its potential to reduce anthropogenic greenhouse gas emissions worldwide. They have predicted and assigned a precise function to hypothetical proteins (HPs) and categorized them as metabolic enzymes, binding proteins, and transport proteins using a combined bioinformatics approach. Moreover, they propose new methane mitigation interventions that target the key metabolic proteins to reduce methane emissions in ruminants.

In the next paper, Choure *et al.*, elaborate on a comparative metagenomic analysis of two alkaline hot springs, Chhoti Anhoni and Badi Anhoni of Madhya Pradesh, India, and decoded the extremophiles for industrial enzymes. The objective of this study was to undertake, analyze, and characterize the microbiome to find out the inhabitant microbial population, and their functional characteristics. The study showed the presence of different unassigned bacterial taxa with great abundance, which indicates the potential of novel genera or phylotypes. Furthermore, the functional analysis of microbiomes revealed that most of the genes are associated with functions related to metabolism and environmental information processing.

Joshi *et al.* sequenced and analyzed the total number of 502 SARS-CoV-2 genomes from Gujarat, India to understand its phylogenetic distribution and variants against global and national sequences to understand its role in pathogenesis. The SARS-CoV-2 genomes they found, namely C28854T (Ser194Leu), showed an allele frequency of 47.62 and 7.25 percent in patients who dies from Gujarat and worldwide datasets, respectively, among the missense mutations. They concluded that SARS-CoV-2 genomes from Gujarat are forming distinct clusters under the GH clade of GISAID. Rampelli *et al.*, developed G2S, a bioinformatic tool for taxonomic prediction of the human fecal microbiome directly from the oral microbiome data of the same individual. This tool can be used in retrospective studies, where fecal sampling was not performed, especially in the field of paleomicrobiology.

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

***Correspondence:**
Saumya Patel
saumya50@gmail.com
patelsaumya@gujaratuniversity.ac.in

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *07 January 2022*
Accepted: *17 January 2022*
Published: *11 February 2022*

Citation:
Acharya D, Kuddus M and Patel S (2022) Editorial: Computational Genomics and Structural Bioinformatics in Microbial Science. Front. Genet. 13:850397. doi: 10.3389/fgene.2022.850397

Liu *et al.* developed a novel algorithm called DRAGoM for family-based ncRNA homology searches against metagenomic sequencing data (Detection of RNA using Assembly Graph from Metagenomic data). This tool can improve taxonomic analysis through facilitating the use of ncRNA families as taxonomic biomarkers. Andreu-Sánchez et al., benchmarked seven bioinformatic tools for genetic variant, calling in metagenomics data and evaluating their performance. This benchmark showed probabilistic tools that can be used to call metagenomes and recommendations of GATK's tools as reliable variant callers for metagenomic samples.

Sevugapperumal et al. reported a draft genome sequence of B. amyloliquefaciens strain CB, which was isolated from the rhizospheric soil of a cotton plant, and which can be used as a reference sequence to explore and map specific genes related to antimicrobial peptide (AMP) genes and other important enzymes. The genome interpretation of B. amyloliquefaciens strain CB indicated antagonistic potential due to AMPs imparting various antifungal, antibacterial, and antiviral properties as well plant growth promotion, leading to strong prospects for uplifting sustainable agriculture.

Liu et al. attempted to reconstruct the biogeographical structure according to functional traits and the evolutionary lineage of B. amyloliquefaciens using comparative genomics analysis. Nimavat et al. analyzed 2,349 genome sequences of SARS-CoV-2 submitted in GISAID by a single institute pertaining to infections from the Gujarat state to know their variants and phylogenetic distributions with a major focus on the spike protein. The D614G variant in spike protein has been reported to have a very high frequency of >95% globally followed by the L452R and P681R.

Ahmad *et al.* modeled methyltransferase as antibiotics against foodborne pathogens. Its interactions were analyzed against a membrane protein of the Gram-positive and Gram-negative bacteria through *in silico* protein–protein interactions and established that it is a conclusively useful enzymobiotics agent.

The variety of the topic contributions by authors, including theoretical considerations and research articles, shed light on current advances in Computational Genomics and Structural Bioinformatics in Microbial Science and support further approaches for research in integrative microbial science.

AUTHOR CONTRIBUTIONS

SP and DA are responsible for background and analysis of contributions. MK is responsible for the structure and comments in the editorial.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 December 2020
doi: 10.3389/fgene.2020.593990

Functional Prediction and Assignment of *Methanobrevibacter ruminantium* M1 Operome Using a Combined Bioinformatics Approach

M. Bharathi[1], *N. Senthil Kumar*[2] *and P. Chellapandi*[1]*

[1] *Molecular Systems Engineering Lab, Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli, India,* [2] *Human Genetics Lab, Department of Biotechnology, School of Life Sciences, Mizoram University (Central University), Aizawl, India*

Methanobrevibacter ruminantium M1 (MRU) is a rod-shaped rumen methanogen with the ability to use H_2 and CO_2, and formate as substrates for methane formation in the ruminants. Enteric methane emitted from this organism can also be influential to the loss of dietary energy in ruminants and humans. To date, there is no successful technology to reduce methane due to a lack of knowledge on its molecular machinery and 73% conserved hypothetical proteins (HPs; operome) whose functions are still not ascertained perceptively. To address this issue, we have predicted and assigned a precise function to HPs and categorize them as metabolic enzymes, binding proteins, and transport proteins using a combined bioinformatics approach. The results of our study show that 257 (34%) HPs have well-defined functions and contributed essential roles in its growth physiology and host adaptation. The genome-neighborhood analysis identified 6 operon-like clusters such as *hsp*, TRAM, *dsr*, *cbs* and *cas*, which are responsible for protein folding, sudden heat-shock, host defense, and protection against the toxicities in the rumen. The functions predicted from MRU operome comprised of 96 metabolic enzymes with 17 metabolic subsystems, 31 transcriptional regulators, 23 transport, and 11 binding proteins. Functional annotation of its operome is thus more imperative to unravel the molecular and cellular machinery at the systems-level. The functional assignment of its operome would advance strategies to develop new anti-methanogenic targets to mitigate methane production. Hence, our approach provides new insight into the understanding of its growth physiology and lifestyle in the ruminants and also to reduce anthropogenic greenhouse gas emissions worldwide.

Keywords: methanobrevibacter, methane mitigation, hypothetical proteins, protein function, molecular machinery

OPEN ACCESS

Edited by:
Saumya Patel,
Gujarat University, India

Reviewed by:
Khanh N. Q. Le,
Taipei Medical University, Taiwan
Sailu Yellaboina,
CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, India

***Correspondence:**
P. Chellapandi
pchellapandi@gmail.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *12 August 2020*
Accepted: *17 November 2020*
Published: *16 December 2020*

Citation:
Bharathi M, Senthil Kumar N and Chellapandi P (2020) Functional Prediction and Assignment of Methanobrevibacter ruminantium M1 Operome Using a Combined Bioinformatics Approach. Front. Genet. 11:593990. doi: 10.3389/fgene.2020.593990

INTRODUCTION

Enteric methane emission from ruminants is of great concern not only for its impact on global warming potential but also for ensuring the long-term sustainability of ruminant-based agriculture. Methane emission from rumen methanogens (163.3 million metric tons of CO_2 equivalents) represents a loss of about 5–7% of dietary energy in ruminants (Hristov et al., 2013;

Chellapandi et al., 2017a, 2018; Chellapandi and Prathiviraj, 2020). Methanobrevibacter genus is a dominant rumen methanogenic archaea (61.6%) in which *Methanobrevibacter ruminantium* M1 (MRU) accounted for 27.3% (Janssen and Kirs, 2008). MRU is a hydrogenotrophic rumen methanogen that use H_2 to reduce CO_2 for methane biosynthesis. It also uses formate as a carbon source for its growth and energy metabolism (Kaster et al., 2011). This is the first genome sequence to be completed for rumen methanogen. It is a circular chromosome (2.93 Mbp) consisting of 2,278 coding-genes and 144 metabolic pathways with 722 reactions, 557 enzymes, and 751 metabolites (Leahy et al., 2010). However, the MRU genome consists of 756 coding-genes (73%) annotated as hypothetical proteins (HPs). It suggests that the entire proteome functions of this organism are not yet known and have to be elucidated to date.

The function of only 50–70% of coding-genes has been annotated with reasonable confidence in the most completely sequenced bacterial genomes using automated genome sequence analysis (Loewenstein et al., 2009). The characterization of proteins with unknown biological function is known as operome (Greenbaum et al., 2001; Chellapandi et al., 2017b; Prathiviraj and Chellapandi, 2019). Putative genes with known orthologs and no orthologs are termed as conserved hypothetical proteins and uncharacterized proteins, respectively (Mazandu and Mulder, 2012; Shahbaaz et al., 2013). Several approaches have been developed for assisting the function of operome from prokaryotic genomes using the information derived from sequence and structural motifs (Sivashankari and Shanmughavel, 2006; Chellapandi et al., 2017b; Singh and Singh, 2018; Prathiviraj and Chellapandi, 2020a; Sangavai et al., 2020). No one has been employed a combined bioinformatics prediction approach including sequence, structure, and literature confidences for functional assignment of operome and its contribution to metabolic subsystems and cellular machinery. A precise annotation of the operome of a particular genome leads to the discovery of new functions for the development of veterinary and human therapeutics (Ijaq et al., 2015).

The conserved domain-based functional assignment was done for HPs from *Pongo abeli*i and *Sus scrofa*. It has provided a hint for genome-wide annotation in poorly understood genomes (Jitendra et al., 2011). The structure-based approach has been applied to predict the function of operome from *Mycoplasma hyopneumoniae* (da Fonsêca et al., 2012). Functional and structural domain analysis (Namboori et al., 2004), integrated genomic context analysis (Yellaboina et al., 2007) and literature mining (Doerks et al., 2012), functional enrichment analysis (Mazandu and Mulder, 2012), and genome-scale fold-recognition (Mao et al., 2013) have been used to annotate the potential function of operome from *Mycobacterium tuberculosis* H37Rv. Sequence-based and structure-based approaches have been used to define and prioritize some HPs from *Candida dubliniensis*, *Vibrio cholerae* O139, and *Staphylococcus aureus* as therapeutic targets for the treatment of their infections in humans (McAdow et al., 2011, 2012; Bharat Siva Varma et al., 2015; Islam et al., 2015). Besides, only one HP (MJ_0577) was functionally annotated in *Methanococcus jannaschii* using a structural-based approach (Zarembinski et al., 1998).

Many *in silico* attempts have been focused on the functional prediction of operome from human pathogens and no reports on rumen methanogens. Several genome-scale metabolic networks have been reconstructed for methanogenic archaea with a low fraction of HPs functionally assigned by sequence similarity analysis (Chellapandi et al., 2018; Prathiviraj and Chellapandi, 2020a). Since, functional annotation of operome is a great concern not only for implementing our fragmentary knowledge on the potential drug targets but also for genome refinement and improved microbial genome-scale reconstructions (Poulsen et al., 2010; Mazandu and Mulder, 2012; Prathiviraj and Chellapandi, 2019). Thus, we have employed a combined bioinformatics approach for functional assignment, and categorization of operome from MRU with a biological knowledgebase. The predicted functions of operome allow us to comprehend its growth physiology and metabolic behavior in the rumen environment. Several methanogenic antibiotics, inhibitors, and vaccines have been currently available for enteric methane mitigation, but these are a narrow spectrum and species-specific activity (Pulendran and Ahmed, 2006). The present approach is used to predict new anti-methanogenic targets from its precisely annotated operome that resolves the current demand for veterinary therapeutics.

MATERIALS AND METHODS

Dataset Preparation

We retrieved protein sequences of 756 HPs in the MRU genome from the National Centre for Biotechnology Information (NCBI)[1] and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2018) using a simple text mining approach (Le and Huynh, 2019; Le et al., 2019). We used broad ranges of source types such as keywords, "*hypothetical proteins*, *unknown*, *uncharacterized*, and *putative*" to retrieve the protein sequences from the NCBI and KEGG (Chellapandi et al., 2017b). The FASTA sequences of all HPs were taken separately to carry out sequence analysis. For functional annotation and assignment of MRU operome, we used six different prediction tasks as detailed below (**Figure 1**). The overall information about similar or identical functions of HPs predicted from each task was manually evaluated to reasoning out the functional assignment of operome. The prediction tools used for each functional annotation were more robust and confident for our analysis similar to the previous works on archaeal and bacterial operome (Prathiviraj and Chellapandi, 2019; Sangavai et al., 2020). E-value is the number of expected hits of a similar score that could be found just by chance. Like *p*-value, we used e-value for the scoring of each prediction from the dataset and represented in **Supplementary Data**.

Conserved Motif Analysis

A motif is a short segment of a protein sequence or structure, which may be conserved in a large number of different proteins. It can be used to determine the function or conformation

[1] http://www.ncbi.nlm.nih.gov

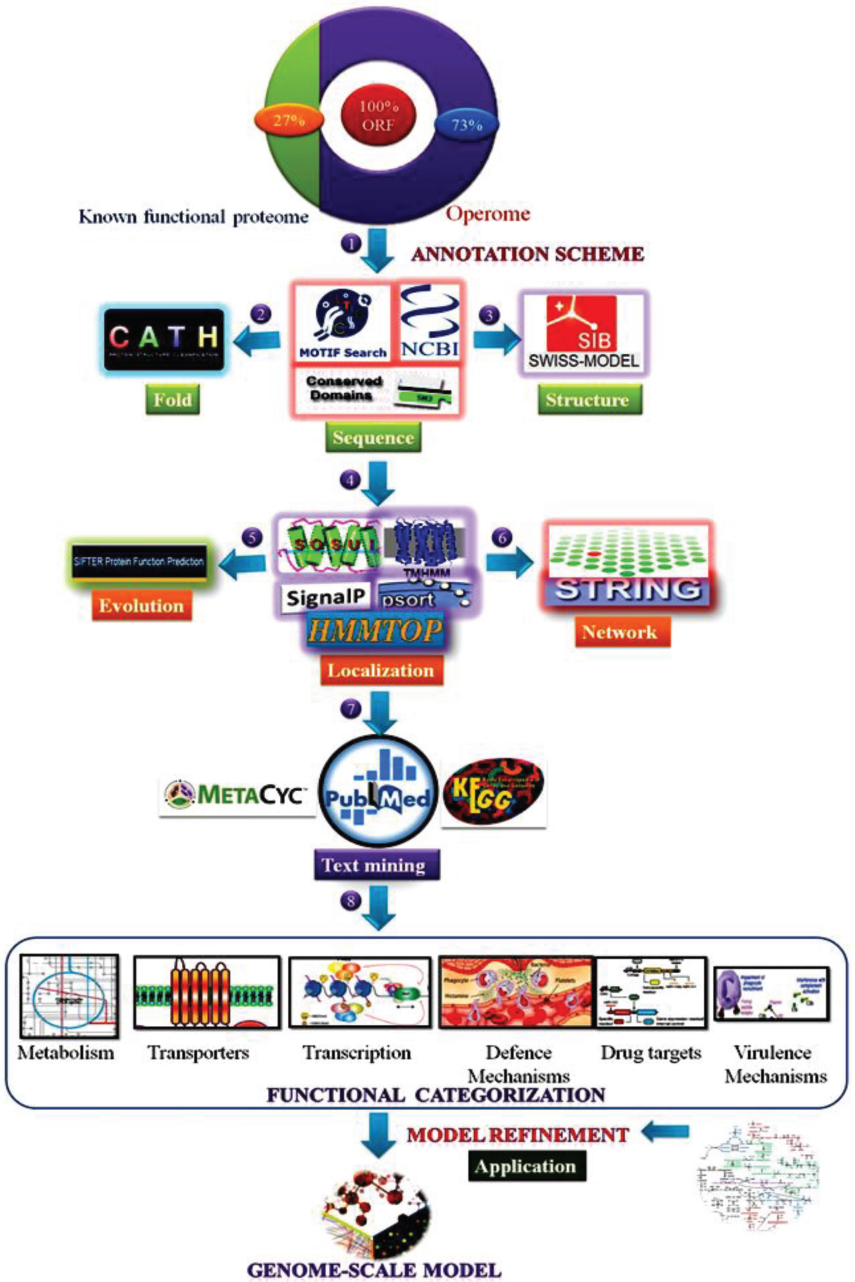

FIGURE 1 | Experimental workflow of a combined bioinformatics approach employed for functional annotation of operome from MRU.

of a protein. The conserved motifs in each protein were searched out against the KEGG-Motif search tool[2], InterProScan (Quevillon et al., 2005), and Pfam library (Finn et al., 2016). To improve the lineament of prediction, cut off value was set as 10^{-5} and DUF (domains with unknown functions) were removed from the dataset. We found motif similarity hits for 756 HPs out of which 257 HPs were chosen for further analysis.

[2]http://www.genome.jp/tools/motif/

Conserved Domain Analysis

Conserved domains in each protein were identified by the NCBI-CDD v3.16 search tool using the position-dependent weight matrices. Additionally, composition-based statistics adjustment was used to remove low complexity composition for statistical significance using the RPS-BLAST version 2.2.28 (Marchler-Bauer et al., 2015). The query sequence was compared with domain architecture and profiles in the domain databases, after that, the compositionally biased conserved region was identified by the SMART (Letunic et al., 2012). The PROSITE profile

was scanned for detection of the protein domains, families, and functional sites and associated patterns in the protein sequence using ScanProsite (de Castro et al., 2006). The probable function of HPs was predicted with the InterPro database based on the domain and important sites in the sequences (Finn et al., 2016).

Structural Analysis

The secondary structural elements (helix, sheets, extended coil, and loops) in each protein were predicted from the sequences using SOPMA (Geourjon and Deléage, 1995). We identified structural and functional characteristics by PSI-BLAST similarity searching against the protein data bank[3] (Altschul et al., 1997). The sequence similarity hits were selected for finding the alignment of functional residues of a protein of known function with the sequence of HPs using ClustalW (Thompson et al., 2002). Fold assignment, target-template alignment, model building, and model evaluation were carried out with the Swiss Model (Biasini et al., 2014). QMEAN was a composite scoring function describing the major geometrical aspects of protein structures as described below.

$$S_{\text{weighted}_{\text{average}}(x)} = \frac{\Sigma_i(\text{GDT}^*_{\text{TS(x,i)}}\text{QMEAN(i)})}{\Sigma_i\text{QMEAN(i)}}$$

where, the GDT_TS score as the target function. We evaluated the structural quality and accuracy of the resulted homology models based on the potential function as below (Benkert et al., 2008).

$$\text{QMEAN5 score} = 0.3 \times \text{Score}_{\text{torsion 3-residue}} + 0.17 \times \text{Score}_{\text{pairwise C}\beta/\text{SSE}} + 0.7 \times \text{Score}_{\text{solvation C}\beta} + 80 \times \text{Score}_{\text{SSE PSPIRED}} + 45 \times \text{Score}_{\text{ACCpro}}$$

Evolutionary Trace Analysis

The evolutionary relationships to deduce the functionality of operome were inferred using the SIFTER (Radivojac et al., 2013). It was used to predict the protein function and Gene ontology term using the following confidence score.

$$\text{Sg(f)} = 1 - \prod_{i=1}^{k}(1 - \text{Sg(f)})$$

where, Sg(f) confidence score as the default prediction for a query protein g, Sg_i(f) is the probability domain has function f (Sahraeian et al., 2015).

Analysis of Physicochemical Properties

The physiochemical properties including molecular weight, theoretical pI, instability index, aliphatic index, and grand average of hydropathicity of HPs were predicted from their sequences using the Expasy's Protparam server[4]. The instability index provides an estimate of the stability of a protein. An instability index <40 is predicted to be stable, and a value >40 is predicted to be unstable. The instability index uses the following weight values.

$$\text{II} = \left(\frac{10}{L}\right) * \sum_{\text{i=1}}^{\text{i=L-1}} \text{DIWV(x[i]x[i+1])}$$

where, L is the length of the sequence, DIWV(x[i]x[i+1]) is the instability weight value for the dipeptide starting in position I (Guruprasad et al., 1990). The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chain amino acids using the following equation.

$$\text{Aliphatic index X(Ala)} + \text{a} * \text{X(Val)} + \text{b} * (\text{X(Ile)} + \text{X(Leu)})$$

Where, X(Ala), X(Val), X(Ile), and X(Leu) are mole percent (100 X mole fraction) (Ikai, 1980). The GRAVY value for a protein is calculated as the sum of the hydropathy values of all of the amino acids divided by the number of residues in the sequence (Kyte and Doolittle, 1982).

Analysis of Protein Subcellular Localization

The subcellular localization of every protein was predicted with PSORTb version 3.0.2 based on the hydrophobicity index of amino acids (Yu et al., 2010). The propensity of a protein for being a membrane protein was predicted by SOSUI 2.0 based on the physicochemical parameters (Mitaku et al., 2002). The transmembrane helix and topology of each protein were detected by the TMHMM 2.0 (Krogh et al., 2001) and HMMTOP (Tusnády and Simon, 2001) using the Hidden Markov Model. The signal peptide and location of the cleavage site in the peptide chain were predicted with the SingnalP 4.0 based on a neural network model (Petersen et al., 2011).

Literature Search

The literature survey is the stepping-stone and an essential skill toward the accomplishment of structural and functional analysis provides of proteins (Hubbard and Dunbar, 2017). A process of uncovering useful knowledge from a collection of data from bioinformatics and literature databases is referred to as a knowledge-based discovery (Chellapandi et al., 2017b). Functional assessment of operome was strengthened by extracting relevant experimental supports from available literature in NCBI-PubMed[5]. A maximum confidence score was set as 12 levels (6 levels from predictions and 6 levels from the literature mining) in which 50% score systematically enumerated and assigned from overall prediction approaches. The rest of them was assigned by manual annotation based on the strength of the literature validation. For example, if the predicted function is similar or identical in all prediction approaches, a maximum confidence score will be assigned as 6. The literature-based confidence score for each predicted function of HPs assigned as; 6- MRU, 5- Phylogenetic neighbors, 4- Methanogens, 3- Archaea,

[3]www.rcsb.org/

[4]http://web.expasy.org/protparam/

[5]https://www.ncbi.nlm.nih.gov/pubmed/

2- Bacteria, and 1- Eukaryotes. We have set a confidence score interval as 3–6 for both computational prediction and biological knowledge base and then neglected the predicted function of a protein with a low confidence score (<3).

Functional Categorization

We classified the predicted function of HPs based on conserved domain, protein fold, family, and biological function using the CATH database (Knudsen and Wiuf, 2010). The genome-wide analysis was performed to identify the order of gene clusters covering the predicted function of HPs using a genomic context approach (Yellaboina et al., 2007). Gene-neighborhood or adjutant genes were identified by exploring the MRU genome in the KEGG database. Metabolic information of HPs was collected from the MetaCyc (Metabolic Pathways from all Domains of Life) database (Caspi et al., 2014). The resulted data were used to assign the functions of hypothetical proteins of the understudied genome. The overall structural and functional information was manually analyzed to categorize the molecular involvement of HPs in respective metabolic subsystems and the cellular process of the understudied organism.

RESULTS

Functional Classification and Categorization

All predicted protein functions were classified and categorized according to their protein folds, molecular function, subsystems, and transmembrane topologies as shown in **Figure 2**. About 20% of operome encompasses a Rossmann fold consisting of a nucleotide cofactor binding domain of some NAD^+-dependent dehydrogenases, in particular to ribonucleases (Barbas et al., 2013). Fourteen percent of operome belongs to rubrerythrin that constitutes non-haem iron proteins. This functional fold is responsible for oxidative stress protection in anaerobic bacteria and archaea (Prakash et al., 2018). The arcR repressor mutant fold occupies 4–5% of operome, which performs the functions of small homodimeric proteins involved in transcriptional regulation by sequence-specific DNA binding (Vershon et al., 1986; Homa and Brown, 1997). MRU operome contains phoA fold (3–4%) that fused with the cell surface glycoprotein signal sequence similar to *Haloferax volcanii* (Kandiba et al., 2013). It indicates the importance of some protein folds for conferring oxidative tolerance and cell wall assembly. We found 91 HPs involving in the metabolic reactions with a confidence score >5. A total of 23 HPs is entailed in the small molecule reactions and 15 HPs required for the biosynthesis of cofactors, prosthetic groups, and electron carriers. About 9 HPs are essential to the protein modification reactions whereas 4 HPs contributed to the formation of precursor metabolites for the energy-driven process of this organism. Approximately 50% of drug targets are transmembrane proteins as they play many roles in transport, cell signaling, and energy transduction processes (Terstappen and Reggiani, 2001). We predicted 91 HPs having transmembrane helixes based on their conservation of membranous helix ratios. The α-helix bundle and the β-barrel are predicted as fold classes in many membrane proteins. Archaeal transmembrane proteins have two or more α-helixes consisting of hydrophobic amino acids.

Operon-Like Organization

The genome-wide analysis discovered 32 coding genes for HPs, which are all clustered separately, form 6 operon-like organizations (*hsp, TRAM, dsr, cbs, anti-toxin*, and *cas*) in the MRU genome (**Figure 3**). Molecular chaperones such as hsp70, hsp60, and hsp80 resemble some bacterial genomes than the eukaryotic homologs (Gaywee et al., 2002). The *hsp* gene cluster is essential for chaperone-assisted protein folding in Achaea (Dokland, 1999; Benaroudj and Goldberg, 2000; Large et al., 2009). The assimilatory sulfite reductase (*dsrHFEBA*) gene cluster detected from this genome provides the importance for the oxidation of accumulated intracellular sulfide and thiosulfate in the diverse environmental niche. The presence of *cbs, anti-toxin*, and *cas* gene clusters confers host defense response (innate immunity) to this organism against foreign genetic elements in the rumen ecosystem (Louwen et al., 2014; Chellapandi and Ranjani, 2015). The anti-toxin system plays a vital role in toxicity neutralization (Unterholzner et al., 2013).

Cell Division Systems

In this study, we assigned the function of 9 HPs contributing a major role in the cell cycle process in which 8 HPs have shown new functions to this organism (**Table 1**). AAA^+ ATPase, cell division inhibitor, cell division control protein, DNA replication protein 6-2, and structural maintenance of chromosomes protein-1 is highly conserved within the archaeal domain and performs archaeal-specific cell cycle process, DNA repair, and replication fidelity (Kalliomaa-Sanford et al., 2012; Grogan, 2015). A proteasome is a central player in energy-dependent proteolysis and forms a nano-compartment where proteins are degraded into oligopeptides by processive hydrolysis. The 20S proteasome is a catalytic core responsible for this processing. AAA^+ ATPase plays several roles in mediating energy-dependent proteolysis by the proteasome (Forouhar et al., 2011; Maupin-Furlow, 2013). Moreover, it contains a P-loop motif involved in the origin of recognition during DNA replication initiation even if conventional C-terminal winged-helix DNA-binding elements lacked (He et al., 2008).

Transcriptional Regulatory Systems

A total of 26 HPs predicted as functional candidates in which 20 HPs have shown new functions to the transcriptional regulation process of this organism (**Table 2**). Transcriptional regulatory proteins identified from MRU operome can express a set of proteins that protect cellular proteins against a sudden heat-shock stress, copper and arsenic toxicities, protein folding, and nitrogen starvation (Thieringer et al., 1998; Giaquinto et al., 2007; Chang et al., 2014; Prathiviraj and Chellapandi, 2020a,b). Bro N-terminal domain protein has an N-terminal domain with ALI motif that influences host DNA replication and/or transcription (Makarova et al., 2009). HrcA repressor contains a motif of winged helix-turn-helix transcription repressor. It controls the transcription of heat-shock repressor proteins and

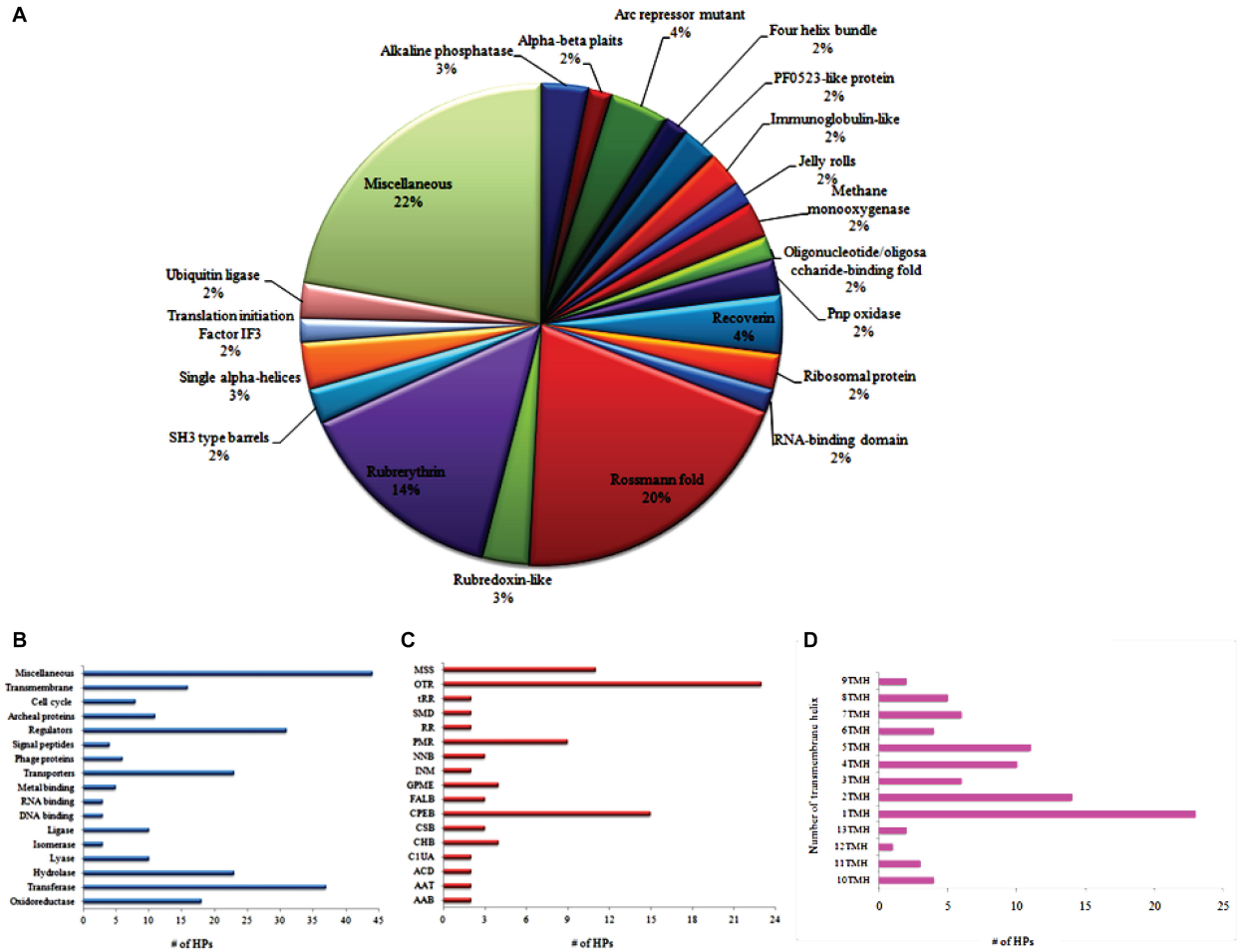

FIGURE 2 | Functional classification of MRU operome based on the protein fold **(A)**, functional category **(B)**, subpathway systems **(C)**, and transmembrane topologies **(D)**. AAB, Amino acid biosynthesis; AAT, Aminoacyl-tRNA charging metabolic clusters; ACD, Aromatic compounds degradation; C1UA, C1 Compounds utilization and assimilation; CHB, Carbohydrates biosynthesis; CSB, Cell structures biosynthesis; CPEB, Cofactors, prosthetic groups, electron carriers biosynthesis; FALB, Fatty acid and lipid biosynthesis; GPME, Generation of precursor metabolites and energy; INM, Inorganic nutrients metabolism; NNB, Nucleosides and nucleotides biosynthesis; PMR, Protein-modification reactions; RR, RNA-reactions; SMD, Secondary metabolites degradation; tRR, tRNA reactions; OTR, Other reactions.

protects cellular proteins from being denatured by heat (Liu et al., 2005; Prathiviraj and Chellapandi, 2020b). Hsp70 and Hsp80 from MRU operome perform renaturation of luciferase similar to that found in *M. mazei* (Zmijewski et al., 2004). Hsp60s are more similar to the type II chaperonins found in the eukaryotic cytosol involved in macromolecular assembly and protein folding (Large et al., 2009). TRAM protein regulates the RNA chaperone activity that is essential for MRU to grow and survive in a cold environment (Zhang et al., 2017).

Biosynthesis of Macromolecules

We predicted the function of 20 HPs exhibiting new metabolic roles in this organism and the rest of 76 HPs has shown known functions (**Table 3** and **Supplementary Table S1**). Saccharopine dehydrogenase (NAD/P, L-lysine-forming) (*lysA*) and succinylglutamate desuccinylase (*astE*) genes identified from MRU operome, which are responsible to mediate the biosynthesis of L-lysine and L-glutamate. LysA protein contains a motif of LOR/SDH bifunctional conserved region that converts L-saccharopine into L-lysine via l-α-aminoadipate pathway (Xu et al., 2007). Cheng et al. (2010), revealed a cross-talk between fungi and methanogens which may occur in host animals since the l-α-aminoadipate pathway is very specific to fungi. The second enzyme transforms N_2-succinylglutamate into succinate and glutamate. Therefore, both enzymes proposed to be involved in amino acid biosynthesis of MRU as reported earlier on other methanogens (Enzmann et al., 2018).

The 2-enoyl-CoA hydratase catalyzes the second step in the physiologically important β-oxidation pathway of fatty acid metabolism in MRU (Agnihotri and Liu, 2003). Glycogen phosphorylase catalyzes the phosphorolysis of α-1, 4 glycosidic bonds in glycogen to yield glucose-1-phosphate for glycolysis (Rath et al., 2000). Interestingly, MRU operome has the ability to synthesis enterobacterial-like common

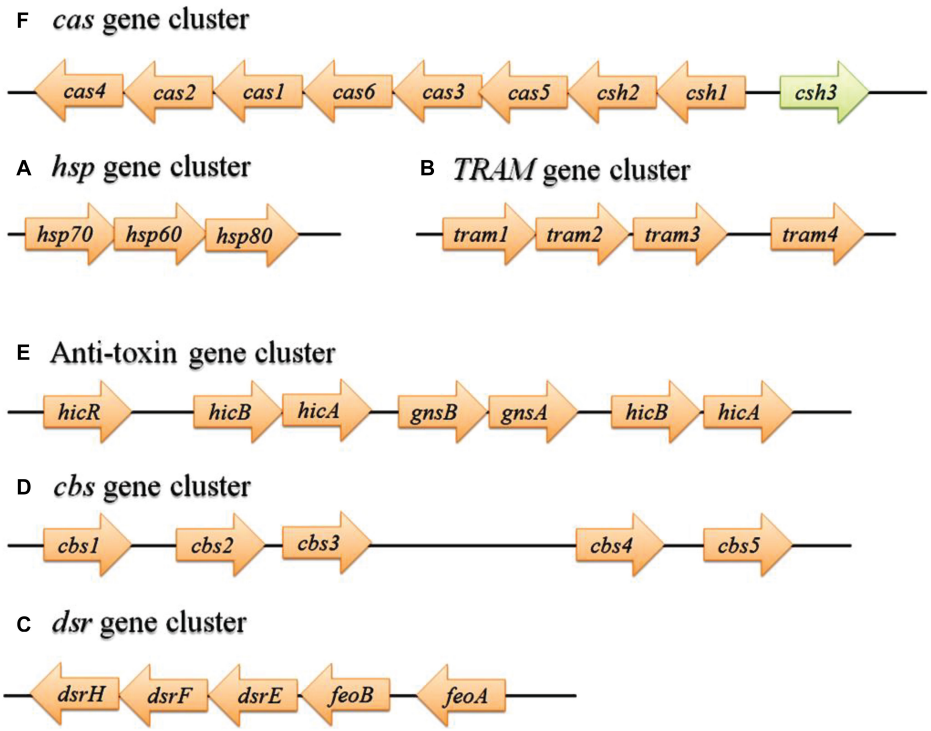

FIGURE 3 | Detection of gene clusters from MRU operome responsible for protein folding **(A)**, cold adaptation **(B)**, sulfite tolerance **(C)**, binding with adenosyl groups **(D)**, degradation of the labile antitoxin **(E)**, and defense/virulence system **(F)**. The green arrow represents a gene with a known function. hsp, Heat shock protein; TRAM, RNA modification protein; dsr, Dissimilatory sulfate reductase; cbs, cystathionine beta-synthase; cas, CRISPR-associated gene.

antigen as it contains dTDP-4-amino-4, 6-dideoxygalactose transaminase (*rffA*). This enzyme catalyzes the conversion of TDP-4-keto-6-deoxy-D-glucose to TDP-D-fucosamine similar to the enterobacteria family (Meier-Dieter et al., 1990; Hwang et al., 2004). The presence of phosphatidate cytidylyltransferase (*cdsA*) provides evidence of the biosynthesis of archaeal-specific phospholipids. It catalyzes *sn*-glycerol 3-phosphate into an L-1-phosphatidylglycerol-phosphate precursor-like *Escherichia coli* (Carter et al., 1968). We found an AMMECR1 motif in phosphomevalonate decarboxylase from MRU operome, which converts (R)-mevalonate 5-phosphate to isopentenyl diphosphate in the mevalonate pathway, as reported in *Methanocaldococcus jannaschii* (Grochowski et al., 2006). Results of our study revealed that the MRU genome has shown a metabolic potential for the biosynthesis of enterobacterial-like common antigen, archaeal-specific phospholipids, and isopentenyl diphosphate, a precursor required for cell wall biogenesis.

Cofactors, Prosthetic Groups, Electron Carrier Biosynthesis

We predicted the function of some HPs involving in the biosynthesis of coenzyme F_{420}, flavin, and electron carriers in MRU. F_{420}-0: L-glutamate ligase is a key enzyme identified from MRU operome, which converts multiple γ-linked L-glutamates to the polyglutamated F_{420}derivative in the biosynthesis of coenzyme F_{420} (Li et al., 2003). As reported in bacteria and plants, MRU operome has diamino hydroxy phosphoribosyl aminopyrimidine reductase (*ribD*) that converts 2, 5-diamino-6-(5-phospho-D-ribosylamino)pyrimidine-4(3H)-one into 5-amino-6-(5-phospho-D-ribosylamino)uracil in flavin biosynthesis pathway (Garfoot et al., 2014). Cytidylyltransferase belongs to the NTP transferase superfamily encoded by *mocA* gene (mru_1116) of the MRU genome. It catalyzes the cytidylation of the molybdenum cofactor demanded many functional enzymes (Fay et al., 2015). Energy-converting hydrogenase B subunit O consists of a conserved motif of IHPPAH, which generates low potential electrons required for autotrophic CO_2 assimilation as reported in *Methanococcus maripaludis* (Major et al., 2010).

Aromatic Compounds Degradation Systems

Pyrogallol hydroxytransferase (*athL*) detected from MRU operome has a carboxypeptidase regulatory-like domain. It is involved only in the regulation of peptidase catalyzing the conversion of pyrogallol into phloroglucinol. Phloroglucinolcan stimulates the gut microbiota and decreases the partial pressure of H_2 in the rumen. It suggests the capture of excess H_2 generated from methanogenesis inhibition can be promoted by phloroglucinol utilization in the rumen (Martinez-Fernandez et al., 2017). Interestingly, we assigned a precise function to HP

TABLE 1 | Functional annotation of operome involved in the cell division process of MRU.

Locus tag	Assigned function	Gene
0080| 0744| 0939| 1172| 1932	AAA^+ ATPase	*atad3A*
0647	Cell division inhibitor	*sepF*
1346	Cell division control protein	*minE*
1419	DNA replication protein 6-2	*cdc6-2*
1654	Structural maintenance of chromosomes protein 1	*smc1*

TABLE 2 | Functional annotation of operome involved in the transcriptional regulatory process of MRU.

Locus tag	Assigned function	Gene
0757	Bro N-terminal domain protein	*dxs*
0349	Nitrogen repressor	*nrpR*
1052	Heat-inducible transcriptional repressor	*hrcA*
1099	Translation initiation factor 3	*tif3*
1366| 2156	Arsenical resistance operon repressor	*arsR*
1862	Copper-sensing transcriptional repressor	*csoR*
0488| 0490| 0499| 0658| 0764| 0780| 0790| 0801| 0930| 1131| 1147| 1150| 1364| 1590| 1796	Transcription factor	*tf2B*
1185	Cold shock protein	
1108	DEAD/DEAH box helicase	*polB*
0877	Preprotein translocase	*secY*

Mru_0476 as phenylacetate-CoA oxygenase in phenylacetate catabolic pathway. This enzyme converts phenylacetyl-CoA to a 2-(1, 2-epoxy-1, 2-dihydrophenyl) acetyl-CoA. Archaea harboring key genes of this pathway are some members of the Halobacteria, which may have acquired a multitude of bacterial genes (Kennedy et al., 2001; Notomista et al., 2003). As shown by our analysis, MRU can degrade pyrogallol and phenylacetate produced by gut microbial in ruminants (Martinez-Fernandez et al., 2017).

Detoxification Systems

MRU operome plays a key role in formaldehyde, inorganic arsenate, and copper detoxification process. It contains 6-phosphogluconate dehydrogenase (*gntZ*) gene as homologous to methanotrophic bacteria such as *Methylophilus methylotrophus* and *Methylobacillus flagellates* (Chistoserdova et al., 2000). The presence of arsenate reductase (*arsC*) and Cu^+-exporting ATPase (*copA*) provides a defense system to its cells against inorganic arsenate and copper toxicities (Liu et al., 2007).

Macromolecule Modification Systems

MRU operome contains α-2, 3-sialyltransferase gene coding protein having a Rossmann fold with the architecture of the α-β complex. This enzyme catalyzes the transfer of sialic acid from CMP-N-acetyl-β-neuraminate to membrane proteins and lipids of the cell wall of MRU (Koga et al., 1993). Dolichyl-phosphate-mannose-protein mannosyltransferase is identified as carbohydrate carriers to transfer mannosyl residues to the hydroxy group of serine or threonine residues during the post-translational protein modification process of MRU (Podar et al., 2013).

Membrane Transport Systems

We observed 16 HPs contributing to the transport systems of this organism (**Supplementary Table S2**). MRU operome encompasses genes coding for transporter proteins responsible for maintenance of metal homeostasis in particular to magnesium and manganese ions and uptake/export of vitamin, sulfite, and tricarboxylate (Winnen et al., 2003; Weinitschke et al., 2007; Hattori et al., 2007, 2009; Rodionov et al., 2009; Rosch et al., 2009; Mayer et al., 2012; Karpowich et al., 2015). The presence of PurR-regulated permease regulon and Na^+/H^+ antiporter protein carries out the exchange Na^+ for H^+ across the cytoplasmic membrane of archaea (Rimon et al., 2012). Cell-cell communication and intra-species electron transfer can be mediated by preprotein translocase predicted from its operome, as described for hydrogenotrophic methanogens and *E. coli* (Cooper et al., 2017). Translocation sheath protein has an N-terminal domain that mediates the translocation of SPI-2 TTSS effector proteins in MRU (Nikolaus et al., 2001).

D-Gluconate Catabolic System

As shown by our analysis, we proposed a putative D-gluconate catabolic pathway exclusively present in MRU for the biosynthesis of archaeal membrane phospholipids (**Figure 4**). The presence of six HPs with predicted functions evidences the existence of this pathway in this organism. Klemm et al. (1996), identified a gntP gene to be involved in gluconate uptake by *E. coli*. *Haloferax volcanii* contains a DeoR/GlpR-type transcription factor, which has shown its potential role as a global regulator of sugar metabolism and to cotranscribe with the downstream phosphofructokinase (*pfkB*) gene (Rawls et al., 2010). As similar to *Pseudomonas aeruginosa*, MRU operome has D-gluconate kinase gene despite a membrane-bound D-gluconate dehydrogenase gene to synthesize phospholipids (Matsushita et al., 1979; Schlictman et al., 1995; Kulakova et al., 2001). As similar to archaea, the utilization of gluconate in MRU leads to a branch point for two central metabolic pathways: the Entner-Doudoroff pathway and phospholipids biosynthesis (Bräsen et al., 2014).

DISCUSSION

The function of operome is obscure and quite unsettling in prokaryotic genomes. Understanding important knowledge gaps in the unknown function of operome can unravel their cellular and molecular mechanisms. The functionality of proteins with unknown function have been identified, characterized, and validated with a broad spectrum of genetic and biochemical experiments (Mills et al., 2015). Several computational methods have been used to describe the physiological states of methanogens from the predicted functions of operome (Chellapandi and Prisilla, 2018;

TABLE 3 | Functional annotation of operome involved in different metabolic subsystems of MRU.

Locus tag	Assigned function	EC	Gene
Biosynthesis			
Amino acids biosynthesis			
1696	Carbamoyl-phosphate synthase (glutamine-hydrolyzing)	6.3.5.5	*carB*
1737	Saccharopine dehydrogenase (NAD/P, L-lysine-forming)	1.5.1.7| 1.5.1.8	*lys1*
Aminoacyl-tRNA charging metabolic clusters			
0488| 0490| 0499| 0764| 0780| 0790| 0801| 1131| 1147| 1150| 1364| 1590| 1796	Methionine—tRNA ligase	6.1.1.10	*metG*
1493	Tryptophanyl-tRNA synthetase (Membrane bound)	6.1.1.2	*trpS*
Carbohydrates and Cell structures biosynthesis			
1418	dTDP-4-amino-4,6-dideoxygalactose transaminase	2.6.1.59	*rffA*
1886	Glycogen Phosphorylase	2.4.1.1	*glgP*
1469	UDP-glucose 4-epimerase	5.1.3.2	*galE*
1462	Pantothenate synthase	6.3.2.1	*panC*
0480	Pyruvate kinase	2.7.1.40	*pykA*
Cell structures biosynthesis			
1065	CDP-glycerol glycerophosphotransferase	2.7.8.12	*tagF*
1589| 1957	Thiamine monophosphate synthase	2.7.7.39	*tagD*
Cofactors, Prosthetic groups, Electron carriers biosynthesis			
2219	Cobalamin biosynthesis protein CbiB	6.3.1.10	*cbiB*
0947	Coenzyme F420-0:L-glutamate ligase	6.3.2.31	*cofE*
1116	CTP: Molybdenum cofactor cytidylyltransferase	2.7.7.76	–
1450	Energy-converting hydrogenase B subunit O	1.6.5.3	*ehbO*
0776| 0785	Gamma-glutamyl cyclotransferase	2.3.2.4	*ykqA*
1937	Glutathione peroxidase	1.11.1.9	*gpxA*
0035	NUDIX hydrolase	3.6.1.22	*nadM*
0596	4-Hydroxybenzoate octaprenyltransferase	2.5.1.39	*ubiA*
1550	5-Formyltetrahydrofolate cyclo-ligase activity	6.3.3.2	*mthfs*
1831	Dihydroneopterin aldolase	4.1.2.25	*folB*
1209	Nicotinate-nucleotide pyrophosphorylase [carboxylating]	2.4.2.19	*nadC*
0277	NUDIX hydrolase	3.6.1.22	*nudC*
2172	Riboflavin kinase	2.7.1.161	*ribK*
0432	Tocopherol cyclase	5.5.1.24	*vte1*
1728	Phosphomevalonate decarboxylase	4.1.1.99	*pmd*
Fatty acid and lipid biosynthesis			
0460	Dolichol kinase	2.7.1.108	*dolk*
1693	Integral Membrane bound Phosphatidate cytidylyltransferase	2.7.7.41	*cdsA*
Metabolic regulators biosynthesis			
0939	6-Phosphofructo-2-kinase| Fructose-2,6-bisphosphate 2-phosphatase	2.7.1.105| 3.1.3.46	*pfkfb3*
0393 (K18532 adenylate kinase [EC:2.7.4.3])	Adenylate kinase	2.7.4.3	*adk*
Nucleosides and nucleotides biosynthesis			
0425	L-Threonylcarbamoyladenylate synthase	2.7.7.87	*yrdC/sua5/ywlC*
1890	Phosphoribosylaminoimidazole carboxylase	4.1.1.21	*purE*
0720	Uridylate kinase (DNA binding protein)	2.7.4.22	*pyrH*
Catabolism			
Alcohols degradation			
0528	Coenzyme B12-dependent diol dehydrase	4.2.1.28	*pduC*
Amino acids degradation			
2016| 0381	Succinylglutamate desuccinylase	3.5.1.96	*astE*
Aromatic compounds degradation			
1622	4-Carboxymuconolactone decarboxylase	4.1.1.44	*pcaC*
0476	Phenylacetate-CoA oxygenase	1.14.13.149	*paaJ*
0313	Pyrogallol hydroxytransferase	1.97.1.2	*athL*

(Continued)

TABLE 3 | Continued

Locus tag	Assigned function	EC	Gene
C_1 Compounds utilization and assimilation			
2132	Bifunctional formaldehyde-activating enzyme	4.2.1.147/4.1.2.43	*fae-hps*
1013	Phosphogluconate dehydrogenase (NAD+-dependent, decarboxylating)	1.1.1.343	*gntZ*
Inorganic nutrients metabolism			
1280| 1936	NADPH-dependent FMN reductase	1.5.1.38	*ssuE*
0224| 0376	Phosphonoacetate hydrolase (membrane bound)	3.11.1.2	*phnA*
Secondary metabolites degradation			
1330	Carbohydrate kinase (Integral membrane-bound)	2.7.1.4	*pfkB*
2120	Quercetin dioxygenase	1.13.11.24	*qodI*
Macromolecule modification			
0421	Alpha-2,3-sialyltransferase	2.4.99.4	*siat4a*
Small molecule reactions			
1938	Arsenate Reductase (Thioredoxin)	1.20.4.1	*arsC*
0134	Type I restriction-modification system M subunit HsdM	2.1.1.72	*hsdM*
0674| 1683| 1749	Succinate dehydrogenase (quinone)	1.3.5.1	*sdh*
1013	Phosphogluconate dehydrogenase (NAD+-dependent, decarboxylating)	1.1.1.343	*gntZ*
2194	2-Enoyl-CoA Hydratase	3.4.21.92	*clpP*
0747	2-Polyprenylphenol 6- hydroxylase	1.14.13.-	*ubiB2*
0202	Aconitate hydratase	4.2.1.3	*acnA*
2180| 2184| 2185	Acyltransferase	2.3.1.13	*glyat*
0496	ATP pyrophosphatase	3.6.1.8	*thiI*
0062| 0063| 1113| 1172	ATP-dependent DNA helicase	3.6.4.12	*ashA*
2196	Choloylglycine hydrolase	3.5.1.24	–
0156| 0041	DNA binding E3 SUMO-protein ligase	6.3.2.-	*piaS4*
0174 (K09723 DNA replication factor GINS)	DNA primase small subunit	2.7.7.-	*priA*
2069	DNA-3-methyladenine glycosylase	3.2.2.20	*tag*
1108| 2173	DNA-directed DNA polymerase	2.7.7.7	*polB*
1660| 1699| 1734	Flavin reductase	1.5.1.36	*hpaC*
1442	Geranylgeranyl reductase	1.3.1.83	*chlP*
1290| 1291	Lincosamide nucleotidyltransferase	2.7.7.-	*inuA*
0930	Manganese-dependent inorganic pyrophosphatase	3.6.1.1	*ppaC*
0223	Membrane-bound O-acyltransferase	2.3.1.-	*rimL*
1242	Nucleoside Triphosphate Pyrophosphohydrolase	3.6.1.8	*mazG*
1605| 0049	Nucleotide diphosphatase	3.6.1.9	*ENPP*
2146	Oligosaccharyl transferase	2.4.99.18	*STT3*
1588	Succinylglutamate desuccinylase /aspartoacylase	3.5.1.15	*aspA*
0100	Peptidoglycan-associated polymer biosynthesis	2.-.-.-	*csaB*
1555	Pseudouridine-5'-monophosphatase	3.1.3.-	*HDHD1*
1964	Sterol 3-beta-glucosyltransferase (Phosphorylating)	2.4.1.173	–
1631	UDP-N-acetylglucosamine 2-epimerase (non-hydrolyzing)	5.1.3.14	*wecB*
0835	*von* Willebrand/Integrin A Domains	3.6.4.-	*hepA*
Protein-modification reactions			
1344	Lysine carboxypeptidase	3.4.17.3	*CPN1*
1375	Membrane-bound dolichyl-phosphate-mannose-protein mannosyltransferase	2.4.1.109	*pomT*
0791	Methylated-DNA—[protein]-cysteine S-methyltransferase	2.1.1.63	–
1884	Nucleotide-activated 6-deoxyhexose biosynthesis	2.4.1.109	*pomT*
2158	Putative pyruvate formate-lyase	1.97.1.4	*pflX*
1801| 1867	Ribosomal-protein-alanine N-acetyltransferase	2.3.1.128	*rimI*
1389| 1514	S-Adenosyl-L-methionine-dependent methyltransferase	1.16.1.8	*mtrR*
1096	Serine/threonine protein kinase with TPR repeats	2.7.11.1	*bub1*
1563	Proteasome endopeptidase complex	3.4.25.1	*psmA*
1311| 0426	tRNA-splicing ligase	6.5.1.3	*rtcB*

(Continued)

TABLE 3 | Continued

Locus tag	Assigned function	EC	Gene
Energy metabolism			
Generation of precursor metabolites and energy			
2214	Fuculose 1-phosphate aldolase	4.1.2.17	*fucA*
1894	Fumarate hydratase	4.2.1.2	*fumA*

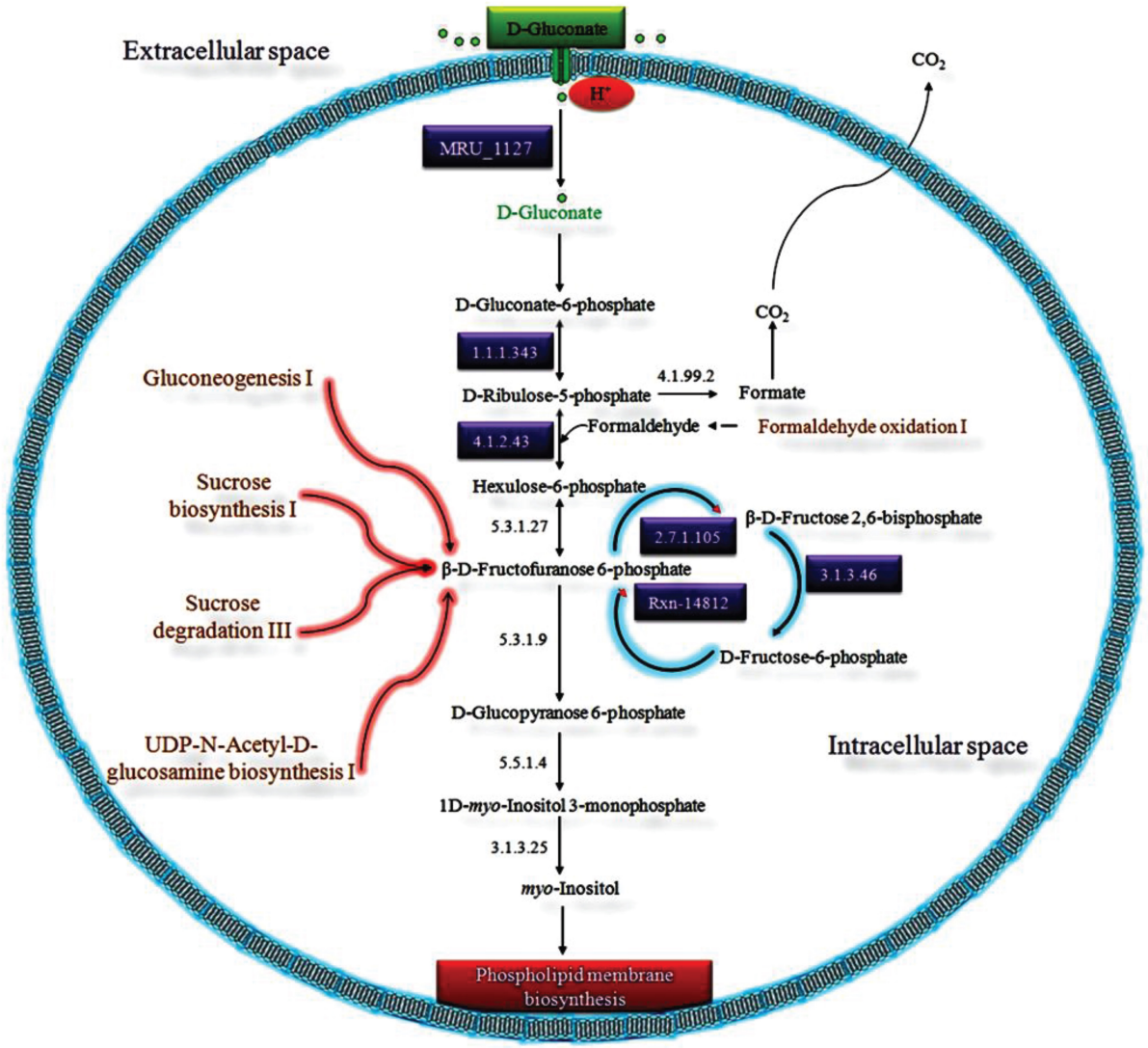

FIGURE 4 | The proposed D-gluconate catabolic pathway in MRU was discovered from the functional annotation of its operome. D-Gluconate is imported into the cytoplasm by the predicted gluconate transporter (*gntP*) gene. It can be phosphorylated to D-gluconate-6-phosphate by D-gluconate kinase (*gntK*), which is then converted to D-ribulose-5-phosphate by the catalytic action of NAD^+-dependent phosphogluconate dehydrogenase (*gntZ*). D-Ribulose-5-phosphate is next oxidized to hexulose-6-phosphate by 3-hexulose phosphate synthase (*hxlA*) and converted into β-D-fructofuranose 6-phosphate with phospho-3-hexuloisomerase (*phi1*). The 6-phosphofructose 2-kinase phosphorylates β-D-fructofuranose 6-phosphate into β-D-fructose 2, 6-bisphosphate, which then interconverted from D-fructose-6-phosphate to β-D-fructofuranose 6-phosphate by fructose-2, 6-bisphosphate 2-phosphatase. In an alternative way, β—D-fructofuranose 6-phosphate is phosphorylated to D-glucopyranose 6-phosphate by 6-phosphofructo-2-kinase. Glucopyranose 6-phosphate is converted to 1D-*myo*-inositol 3-monophosphate by D-glucose 6-phosphate cycloaldolase (*ino1*) and reduced to *myo*-inositol by inositol-phosphate phosphatase (*suhB*).

Prathiviraj and Chellapandi, 2019). There are several functional measures (structural and functional motifs) to be considered for computational predictions of operome from available microbial genomes. The present study employed to collect comprehensive information derived from sequence similarity, conserved domain, motif, structure, fold, protein-protein

interaction, subcellular localization, phylogenetic inference, and gene expression profile as the predictive measures to assign a precise molecular function to MRU operome. Collective information of them provides a hint to predict some distinct motifs and annotate the function of each protein accurately for studying growth physiology in the rumen ecosystem.

Generally, the protein sequence is less conserved than the tertiary structure of a protein (Illergård et al., 2009). In this study, experimentally solved structures and accurate protein folding offered the major importance to deduce some level of a functional description of a protein, as described by Nealon et al. (2017). Characterization of binding motifs and catalytic cores present in the proteins and functional categorization in the cell has been achieved by using the predictive measures derived from overall proteome information (Shapiro and Harris, 2000). Many protein domains have unknown functions, but they may contribute to the metabolic regulation of organisms (Kotze et al., 2013). It implied the possibility of finding a new domain and motif as well as discovers additional protein pathways and cascades from functionally annotated operome (Ijaq et al., 2015). Functional prediction and assignment of prokaryotic operome have been either only sequence-based or structure-based strategies. In our study, a combination of bioinformatics tools with 6 different prediction schemas and additional literature evidence with a 6-level confidence score was applied to improve the prediction accuracy of our functional assignment (**Figure 1**). Compared to earlier functional prediction approaches, our approach provides a strong emphasis to reveal its metabolic subsystems and cellular mechanisms from the assigned function of operome.

The mechanisms of molecular pathogenesis and virulence of many pathogenic organisms and drug targets discovery are being considered an accurate prediction of operome function as an important biological knowledgebase (Amavisit et al., 2003; Lamarche et al., 2008; Kumar et al., 2014). Several bioinformatics tools have been utilized for functional prediction of operome from different pathogenic organisms (Kumar et al., 2014, 2015; Singh et al., 2017; Shrivastava et al., 2017). It clearly described that all of them are pathogenic organisms but no reports on rumen methanogens yet. It was the first computational study to characterize the function of MRU operome, a potential methanogen for enteric methane emission in the ruminants via enteric fermentation.

The Rossmann was a novel and ancient fold found in 5, 10-methenyltetrahydromethanopterin hydrogenase, a key enzyme of hydrogenotrophic methanogenesis. It explains the possibility of hydrogenotrophic lifestyle in MRU, as described by Leahy et al. (2010). The reduction potentials of rubredoxin fold-containing proteins are known to be involved in biochemical processes including carbon fixation, detoxification, and fatty acid metabolism (Prakash et al., 2018; Prathiviraj and Chellapandi, 2020). Cofactors or other prosthetic groups are more attractive to stimulate enzyme activity in hydrolytic reactions of archaea. Transmembrane helixes are generally independently stable in a membrane or membrane-like environment, which are important for signal recognition, transport phenomena, energy translocation, and conservation in the living cell (von Heijne, 1988; Jennings, 1989). Concerning the functional importance, we classified and categorized the function of MRU operome in this study.

In this study, six operon-like clusters were identified from MRU operome. The functions of predicted gene clusters were contributed in chaperone-assisted protein folding, host defense response, and toxicity neutralization of MRU. Some transcriptional regulatory systems predicted from its operome have shown to protect cellular proteins against sudden heat-shock stress, nitrogen limitation, and heavy metal homeostasis. MRU genome contains many pathway holes, which hinder its accurate metabolic reconstruction at the genome-scale. In our study, we detected some key genes missing in the metabolic network of this organism. Consequently, complete metabolic subsystems were annotated for the biosynthesis of L-lysine, L-glutamate, enterobacterial-like common antigen, archaeal-specific phospholipids, and isopentenyl diphosphate. MRU operome can produce coenzyme F_{420} and flavin and electron carriers. Cell wall lipids and membrane proteins have been synthesized from the function of some HPs through macromolecule modification reactions. This organism has well-established transporter systems to maintain metal homeostasis and uptake/export of vitamin, gluconate, sulfite, and tricarboxylate. D-Gluconate catabolic pathway was uniquely discovered from MRU operome for the biosynthesis of archaeal membrane phospholipids.

CONCLUSION

The functional assignment of operome is a mandatory process for a better understanding of the metabolic and molecular processes of this organism. The predicted functional properties of its operome afford us not only for new structural information but also for new molecular functions essential for the lifestyle in the rumen ecosystem. A major operome covers all functional counterparts needed to perform diverse metabolic pathways and regulatory processes. Some imperative physiological functions (oxidative stress, archaeal-specific membrane phospholipids, etc.) of this organism are revealed from this study. The genome-neighborhood analysis found six main gene clusters (hsp, tram, dsr, cbs, anti-toxin, and gas), which are contributed to the energetic metabolism and defense systems. MRU operome contains 119 metabolic enzymes with 18 sub-pathways and 25 binding proteins that recognize the DNA, RNA, metal, and membrane for cellular function. Interestingly, we discovered a putative D-gluconate catabolic pathway for the biosynthesis of archaeal-specific membrane phospholipids. Several virulence-associated and vaccine targeted proteins have been identified from MRU operome. It suggests the development of new methane mitigation interventions that target the key metabolic proteins to reduce methane emissions in ruminants. Functional prediction and assignment of its operome are thus very important to comprehend the cellular machinery at the systems-level for

anti-methanogenic compounds discovery. Nevertheless, all of our predicted functions of its operome should be evaluated and validated experimentally with protein expression and purification, crystallization, and structure determination studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

PC: research design, concept, and manuscript writing. MB: data preparation and analysis. NS: data analysis and manuscript revision. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to thank the University Grants Commission (RA-2012-14-SC-TAM-1768) and Department of Biotechnology (BT/49/NE/2014), New Delhi, India for financial assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.593990/full#supplementary-material

Supplementary Table 1 | Functional annotation of operome involved in diverse metabolic systems of MRU.

Supplementary Table 2 | Functional annotation of operome involved in transporter mechanisms of MRU.

Supplementary Data | All predicted information for functional assignment of MRU operome.

REFERENCES

Agnihotri, G., and Liu, H. W. (2003). Enoyl-CoA hydratase. reaction, mechanism, and inhibition. *Bioorg. Med. Chem.*11, 9–20.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*25, 3389–3402. doi: 10.1093/nar/25.17.3389

Amavisit, P., Lightfoot, D., Browning, G. F., and Markham, P. F. (2003). Variation between pathogenic serovars within *salmonella* pathogenicity islands. *J. Bacteriol.*185, 3624–3635. doi: 10.1128/jb.185.12.3624-3635.2003

Barbas, A., Popescu, A., Frazão, C., Arraiano, C. M., and Fialho, A. M. (2013). Rossmann-fold motifs can confer multiple functions to metabolic enzymes: RNA binding and ribonuclease activity of a UDP-glucose dehydrogenase. *Biochem. Biophys. Res. Commun.*430, 218–224. doi: 10.1016/j.bbrc.2012.10.091

Benaroudj, N., and Goldberg, A. L. (2000). PAN, the proteasome-activating nucleotidase from archaebacteria, is a protein-unfolding molecular chaperone. *Nat. Cell Biol.*2, 833–839. doi: 10.1038/35041081

Benkert, P., Tosatto, S. C., and Schomburg, D. (2008). QMEAN: a comprehensive scoring function for model quality assessment. *Proteins*71, 261–277. doi: 10.1002/prot.21715

Bharat Siva Varma, P., Adimulam, Y. B., and Kodukula, S. (2015). Insilico functional annotation of a hypothetical protein from *Staphylococcus aureus*. *J. Infect. Publ. Health*8, 526–532. doi: 10.1016/j.jiph.2015.03.007

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modeling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*42, W252–W258.

Bräsen, C., Esser, D., Rauch, B., and Siebers, B. (2014). Carbohydrate metabolism in Archaea: current insights into unusual enzymes and pathways and their regulation. *Microbiol. Mol. Biol. Rev.*78, 89–175. doi: 10.1128/mmbr.00041-13

Carter, J. R., Fox, C. F., and Kennedy, E. P. (1968). Interaction of sugars with the membrane protein component of the lactose transport system of *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*60, 725–732. doi: 10.1073/pnas.60.2.725

Caspi, R., Altman, T., Billington, R., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*42, D459–D471.

Chang, F. M., Coyne, H. J., Cubillas, C., Vinuesa, P., Fang, X., Ma, Z., et al. (2014). Cu(I)-mediated allosteric switching in a copper-sensing operon repressor (*CsoR*). *J. Biol. Chem.*289, 19204–19217. doi: 10.1074/jbc.m114.556704

Chellapandi, P., Bharathi, M., Sangavai, C., and Prathiviraj, R. (2018). *Methanobacterium formicicum* as a target rumen methanogen for the development of new methane mitigation interventions-A review. *Veter.Anim. Sci.*6, 86–94. doi: 10.1016/j.vas.2018.09.001

Chellapandi, P., Bharathi, M., Prathiviraj, R., Sasikala, R., and Vikraman, M. (2017a). Genome-scale metabolic model as a virtual platform to reveal the ecological importance of methanogenic archaea. *Curr. Biotechnol.*6, 149–160. doi: 10.2174/2211550105666160901125353

Chellapandi, P., Mohamed Khaja, Hussain, M., and Prathiviraj, R. (2017b). CPSIR-CM: a database for structural properties of proteins identified in cyanobacterial C1 metabolism. *Algal. Res.*22, 135–139. doi: 10.1016/j.algal.2016.12.005

Chellapandi, P., and Prathiviraj, R. (2020). A systems biology perspective of *Methanothermobacter thermautotrophicus* strain ΔH for bioconversion of CO_2 to methane. *J.CO_2 Utiliz.*40:101210. doi: 10.1016/j.jcou.2020.101210

Chellapandi, P., and Prisilla, A. (2018). *Clostridium botulinum* type A-virulome-gut interactions: a systems biology insight. *Hum. Microb. J.*8, 15–22. doi: 10.1016/j.humic.2018.01.003

Chellapandi, P., and Ranjani, J. (2015). Knowledge-based discovery for designing CRISPR-CAS systems against invading mobilomes in thermophiles. *Syst. Synth. Biol.*9, 97–106. doi: 10.1007/s11693-015-9176-8

Cheng, A. G., McAdow, M., Kim, H. K., Bae, T., Missiakas, D. M., and Schneewind, O. (2010). Contribution of coagulases towards *Staphylococcus aureus* disease and protective immunity. *PLoS Pathog.*6:e1001036. doi: 10.1371/journal.ppat.1001036

Chistoserdova, L., Gomelsky, L., Vorholt, J. A., Gomelsky, M., Tsygankov, Y. D., and Lidstrom, M. E. (2000). Analysis of two formaldehyde oxidation pathways in *Methylobacillus flagellatus* KT, a ribulose monophosphate cycle methylotroph. *Microbiology*146, 233–238. doi: 10.1099/00221287-146-1-233

Cooper, R. M., Tsimring, L., and Hasty, J. (2017). Inter-species population dynamics enhance microbial horizontal gene transfer and spread of antibiotic resistance. *eLife*6:e25950.

da Fonsêca, M. M., Zaha, A., Caffarena, E. R., and Vasconcelos, A. T. (2012). Structure-based functional inference of hypothetical proteins from *Mycoplasma hyopneumoniae*. *J. Mol. Model*18, 1917–1925. doi: 10.1007/s00894-011-1212-3

de Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., et al. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.*34, W362–W365.

Doerks, T., van Noort, V., Minguez, P., and Bork, P. (2012). Annotation of the *M. tuberculosis* hypothetical orfeome: adding functional information to more than half of the uncharacterized proteins. *PLoS One7*:e34302. doi: 10.1371/journal.pone.0034302

Dokland, T. (1999). Scaffolding proteins and their role in viral assembly. *Cell Mol. Life Sci.*56, 580–603. doi: 10.1007/s000180050455

Enzmann, F., Mayer, F., Rother, M., and Holtmann, D. (2018). Methanogens: biochemical background and biotechnological applications. *AMB Exp.*8:1.

Fay, A. W., Wiig, J. A., Lee, C. C., and Hu, Y. (2015). Identification and characterization of functional homologs of nitrogenase cofactor biosynthesis protein *NifB* from methanogens. *Proc. Natl. Acad. Sci. U.S.A.*112, 14829–14833. doi: 10.1073/pnas.1510409112

Finn, R. D., Attwood, T. K., Babbitt, P. C., et al. (2016). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*45, D190–D199.

Forouhar, F., Saadat, N., Hussain, M., Seetharaman, J., Lee, I., Janjua, H., et al. (2011). A large conformational change in the putative ATP pyrophosphatase PF0828 induced by ATP binding. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.*67, 1323–1327. doi: 10.1107/s1744309111031447

Garfoot, A. L., Zemska, O., and Rappleye, C. A. (2014). *Histoplasma capsulatum* depends on de novo vitamin biosynthesis for intraphagosomal proliferation. *Infect. Immun.*82, 393–404. doi: 10.1128/iai.00824-13

Gaywee, J., Xu, W., Radulovic, S., Bessman, M. J., and Azad, A. F. (2002). The Rickettsia prowazekii invasion gene homolog (*invA*) encodes a Nudix hydrolase active on adenosine (5')-pentaphospho-(5')-adenosine. *Mol. Cell Proteom.*1, 179–185. doi: 10.1074/mcp.m100030-mcp200

Geourjon, C., and Deléage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.*11, 681–684. doi: 10.1093/bioinformatics/11.6.681

Giaquinto, L., Curmi, P. M., Siddiqui, K. S., Poljak, A., DeLong, E., DasSarma, S., et al. (2007). Structure and function of cold shock proteins in archaea. *J. Bacteriol.*189, 5738–5748. doi: 10.1128/jb.00395-07

Greenbaum, D., Luscombe, N. M., Jansen, R., Qian, J., and Gerstein, M. (2001). Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.*11, 1463–1468. doi: 10.1101/gr.207401

Grochowski, L. L., Xu, H., and White, R. H. (2006). *Methanocaldococcus jannaschii* uses a modified mevalonate pathway for biosynthesis of isopentenyl diphosphate. *J. Bacteriol.*188, 3192–3198. doi: 10.1128/jb.188.9.3192-3198.2006

Grogan, D. W. (2015). Understanding DNA repair in hyperthermophilic archaea: persistent gaps and other reasons to focus on the fork. *Archaea*2015;942605.

Guruprasad, K., Reddy, B. V., and Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence. *Prot. Eng.*4, 155–161. doi: 10.1093/protein/4.2.155

Hattori, M., Iwase, N., Furuya, N., Tanaka, Y., Tsukazaki, T., Ishitani, R., et al. (2009). $Mg^{(2+)}$-dependent gating of bacterial *MgtE* channel underlies $Mg^{(2+)}$ homeostasis. *EMBO J.*28, 3602–3612. doi: 10.1038/emboj.2009.288

Hattori, M., Tanaka, Y., Fukai, S., Ishitani, R., and Nureki, O. (2007). Crystal structure of the *MgtE* Mg^{2+} transporter. *Nature*448, 1072–1075. doi: 10.1038/nature06093

He, Z. G., Feng, Y., Wang, J., and Jiang, P. X. (2008). The regulatory function of N-terminal AAA+ ATPase domain of eukaryote-like archaeal *Orc1/Cdc6* protein during DNA replication initiation. *Arch. Biochem. Biophys.*471, 176–183. doi: 10.1016/j.abb.2008.01.007

Homa, F. L., and Brown, J. C. (1997). Capsid assembly and DNA packaging in herpes simplex virus. *Rev. Med. Virol.*7, 107–122. doi: 10.1002/(sici)1099-1654(199707)7:2<107::aid-rmv191>3.0.co;2-m

Hristov, A. N., Oh, J., Firkins, J. L., Dijkstra, J., Kebreab, E., Waghorn, G., et al. (2013). Special topics–Mitigation of methane and nitrous oxide emissions from animal operations: I. A review of enteric methane mitigation options. *J. Anim. Sci.*91, 5045–5069. doi: 10.2527/jas.2013-6583

Hubbard, K. E., and Dunbar, S. D. (2017). Perceptions of scientific research literature and strategies for reading papers depend on academic career stage. *PLoS One*12:e0189753. doi: 10.1371/journal.pone.0189753

Hwang, B. Y., Lee, H. J., Yang, Y. H., Joo, H. S., and Kim, B. G. (2004). Characterization and investigation of substrate specificity of the sugar aminotransferase *WecE* from *E. coli* K12. *Chem. Biol.*11, 915–925. doi: 10.1016/j.chembiol.2004.04.015

Ijaq, J., Chandrasekharan, M., Poddar, R., Bethi, N., and Sundararajan, V. S. (2015). Annotation and curation of uncharacterized proteins- challenges. *Front. Genet*6:119. doi: 10.3389/fgene.2015.00119

Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *J. Biochem.*88, 1895–1898.

Illergård, K., Ardell, D. H., and Elofsson, A. (2009). Structure is three to ten times more conserved than sequence-a study of structural response in protein cores. *Proteins*77, 499–508. doi: 10.1002/prot.22458

Islam, M. S., Shahik, S. M., Sohel, M., Patwary, N. I., and Hasan, M. A. (2015). *In silico* structural and functional annotation of hypothetical proteins of *Vibrio cholerae* O139. *Genom. Inform.*13, 53–59.

Janssen, P. H., and Kirs, M. (2008). Structure of the archaeal community of the rumen. *Appl. Environ. Microbiol.*74, 3619–3625. doi: 10.1128/aem.02812-07

Jennings, M. L. (1989). Topography of membrane proteins. *Annu. Rev. Biochem.*58, 999–1027. doi: 10.1146/annurev.bi.58.070189.005031

Jitendra, S., Narula, R., Agnihotri, S., and Singh, M. (2011). Annotation of hypothetical proteins orthologous in *Pongo abelii* and *Sus scrofa*. *Bioinformation*6, 297–299. doi: 10.6026/97320630006297

Kalliomaa-Sanford, A. K., Rodriguez-Castañeda, F. A., McLeod, B. N., Latorre-Roselló, V., Smith, J. H., Reimann, J., et al. (2012). Chromosome segregation in Archaea mediated by a hybrid DNA partition machine. *Proc. Natl. Acad. Sci. U.S.A.*109, 3754–3759. doi: 10.1073/pnas.1113384109

Kandiba, L., Guan, Z., and Eichler, J. (2013). Lipid modification gives rise to two distinct *Haloferax volcanii* S-layer glycoprotein populations. *Biochim. Biophys. Acta*1828, 938–943. doi: 10.1016/j.bbamem.2012.11.023

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2018). New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*47, D590–D595. doi: 10.1093/nar/gky962

Karpowich, N. K., Song, J. M., Cocco, N., and Wang, D. N. (2015). ATP binding drives substrate capture in an ECF transporter by a release-and-catch mechanism. *Nat. Struct. Mol. Biol.*22, 565–571. doi: 10.1038/nsmb.3040

Kaster, A. K., Moll, J., Parey, K., and Thauer, R. K. (2011). Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. *Proc. Natl. Acad. Sci. USA.*108, 2981–2986. doi: 10.1073/pnas.1016761108

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., and DasSarma, S. (2001). Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res.*11, 1641–1650. doi: 10.1101/gr.190201

Klemm, P., Tong, S., Nielsen, H., and Conway, T. (1996). The*gntP* gene of *Escherichia coli* involved in gluconate uptake. *J. Bacteriol.*178, 61–67. doi: 10.1128/jb.178.1.61-67.1996

Knudsen, M., and Wiuf, C. (2010). The CATH database. *Hum. Genom.*4, 207–212. doi: 10.1186/1479-7364-4-3-207

Koga, Y., Nishihara, M., Morii, H., and Akagawa-Matsushita, M. (1993). Ether polar lipids of methanogenic bacteria: structures, comparative aspects, and biosyntheses. *Microbiol. Rev.*57, 164–182. doi: 10.1128/mmbr.57.1.164-182.1993

Kotze, H. L., Armitage, E. G., Sharkey, K. J., Allwood, J. W., Dunn, W. B., Williams, K. J., et al. (2013). A novel untargeted metabolomics correlation-based network analysis incorporating human metabolic reconstructions. *BMC Syst. Biol.*7:107. doi: 10.1186/1752-0509-7-107

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*305, 567–580. doi: 10.1006/jmbi.2000.4315

Kulakova, A. N., Kulakov, L. A., Akulenko, N. V., Ksenzenko, V. N., Hamilton, J. T., and Quinn, J. P. (2001). Structural and functional analysis of the phosphonoacetate hydrolase (*phnA*) gene region in *Pseudomonas fluorescens 23F*. *J. Bacteriol.*183, 3268–3275. doi: 10.1128/jb.183.11.3268-3275.2001

Kumar, K., Prakash, A., Anjum, F., Islam, A., Ahmad, F., and Hassan, M. I. (2015). Structure-based functional annotation of hypothetical proteins from *Candida dubliniensis*: a quest for potential drug targets. *3 Biotech.*5, 561–576. doi: 10.1007/s13205-014-0256-3

Kumar, K., Prakash, A., Tasleem, M., Islam, A., Ahmad, F., and Hassan, M. I. (2014). Functional annotation of putative hypothetical proteins from *Candida dubliniensis*. *Gene*543, 93–100. doi: 10.1016/j.gene.2014.03.060

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*157, 105–132. doi: 10.1016/0022-2836(82)90515-0

Lamarche, M. G., Wanner, B. L., Crépin, S., and Harel, J. (2008). The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis. *FEMS Microbiol. Rev.*32, 461–473. doi: 10.1111/j.1574-6976.2008.00101.x

Large, A. T., Goldberg, M. D., and Lund, P. A. (2009). Chaperones and protein folding in the archaea. *Biochem. Soc. Trans.*37, 46–51. doi: 10.1042/bst0370046

Le, N. Q. K., and Huynh, T. T. (2019). Identifying SNAREs by Incorporating Deep Learning Architecture and Amino Acid Embedding Representation. *Front. Physiol.*10:1501. doi: 10.3389/fphys.2019.01501

Le, N. Q. K., Huynh, T. T., Yapp, E. K. Y., and Yeh, H. Y. (2019). Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. *Comput. Methods Progr. Biomed.*177, 81–88. doi: 10.1016/j.cmpb.2019.05.016

Leahy, S. C., Kelly, W. J., Altermann, E., Ronimus, R. S., Yeoman, C. J., Pacheco, D. M., et al. (2010). The genome sequence of the rumen methanogen *Methanobrevibacter ruminantium* reveals new possibilities for controlling ruminant methane emissions. *PLoS One*5:e8926. doi: 10.1371/journal.pone.0008926

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.*40, D302–D305.

Li, H., Xu, H., Graham, D. E., and White, R. H. (2003). Glutathione synthetase homologs encode alpha-L-glutamate ligases for methanogenic coenzyme F_{420} and tetrahydrosarcinapterin biosyntheses. *Proc. Natl. Acad. Sci. U.S.A.*100, 9785–9790. doi: 10.1073/pnas.1733391100

Liu, J., Huang, C., Shin, D. H., Yokota, H., Jancarik, J., Kim, J. S., et al. (2005). Crystal structure of a heat-inducible transcriptional repressor *HrcA* from *Thermotoga maritima*: structural insight into DNA binding and dimerization. *J. Mol. Biol.*350, 987–996. doi: 10.1016/j.jmb.2005.04.021

Liu, T., Ramesh, A., Ma, Z., Ward, S. K., Zhang, L., George, G. N., et al. (2007). *CsoR* is a novel *Mycobacterium tuberculosis* copper-sensing transcriptional regulator. *Nat. Chem. Biol.*3, 60–68. doi: 10.1038/nchembio844

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., et al. (2009). Protein function annotation by homology-based inference. *Genome Biol.*10:207. doi: 10.1186/gb-2009-10-2-207

Louwen, R., Staals, R. H. J., Endtz, H. P., van Baarlen, P., and van der Oost, J. (2014). Therole of CRISPR-cas systems in virulence of pathogenic bacteria. *Microbiol. Mol. Biol.*78, 74–88. doi: 10.1128/mmbr.00039-13

Major, T. A., Liu, Y., and Whitman, W. B. (2010). Characterization of energy-conserving hydrogenase B in *Methanococcus maripaludis*. *J. Bacteriol.*192, 4022–4030. doi: 10.1128/jb.01446-09

Makarova, K. S., Wolf, Y. I., van der Oost, J., and Koonin, E. V. (2009). Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol. Direct.*4:29. doi: 10.1186/1745-6150-4-29

Mao, C., Shukla, M., Larrouy-Maumus, G., Dix, F. L., Kelley, L. A., Sternberg, M. J., et al. (2013). Functional assignment of *Mycobacterium tuberculosis* proteome revealed by genome-scale fold-recognition. *Tuberculosis*93, 40–46. doi: 10.1016/j.tube.2012.11.008

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res.*43, D222–D226.

Martinez-Fernandez, G., Denman, S. E., Cheung, J., and McSweeney, C. S. (2017). Phloroglucinol degradation in the rumen promotes the capture of excess hydrogen generated from methanogenesis inhibition. *Front. Microbiol.*8:1871. doi: 10.3389/fmicb.2017.01871

Matsushita, K., Shinagawa, E., Adachi, O., and Ameyama, M. (1979). Membrane-bound D-gluconate dehydrogenase from *Pseudomonas aeruginosa*. Purification and structure of cytochrome-binding form. *J. Biochem.*85, 1173–1181.

Maupin-Furlow, J. A. (2013). Ubiquitin-like. proteins and their roles in archaea. *Trends Microbiol.*21, 31–38 doi: 10.1016/j.tim.2012.09.006

Mayer, J., Denger, K., Hollemeyer, K., Schleheck, D., and Cook, A. M. (2012). (R)-Cysteate-nitrogen assimilation by *Cupriavidus necator H16* with excretion of 3-sulfolactate: a patchwork pathway. *Arch. Microbiol.*194, 949–957. doi: 10.1007/s00203-012-0825-y

Mazandu, G. K., and Mulder, N. J. (2012). Function prediction and analysis of *Mycobacterium tuberculosis* hypothetical proteins. *Int. J. Mol. Sci.*13, 7283–7302. doi: 10.3390/ijms13067283

McAdow, M., DeDent, A. C., Emolo, C., Cheng, A. G., Kreiswirth, B. N., Missiakas, D. M., et al. (2012). Coagulases as determinants of protective immune responses against *Staphylococcus aureus*. *Infect. Immun.*80, 3389–3398. doi: 10.1128/iai.00562-12

McAdow, M., Kim, H. K., Dedent, A. C., Hendrickx, A. P., Schneewind, O., and Missiakas, D. M. (2011). Preventing *Staphylococcus aureus* sepsis through the inhibition of its agglutination in blood. *PLoS Pathog.*7:e1002307. doi: 10.1371/journal.ppat.1002307

Meier-Dieter, U., Starman, R., Barr, K., Mayer, H., and Rick, P. D. (1990). Biosynthesis of enterobacterial common antigen in *Escherichia coli*. Biochemical characterization of Tn10 insertion mutants defective in enterobacterial common antigen synthesis. *J. Biol. Chem.*265, 13490–13497.

Mills, C. L., Beuning, P. J., and Ondrechen, M. J. (2015). Biochemical functional predictions for protein structures of unknown or uncertain function. *Comput. Struct. Biotechnol. J.*13, 182–191. doi: 10.1016/j.csbj.2015.02.003

Mitaku, S., Hirokawa, T., and Tsuji, T. (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics*18, 608–616. doi: 10.1093/bioinformatics/18.4.608

Namboori, S., Mhatre, N., Sujatha, S., Srinivasan, N., and Pandit, S. B. (2004). Enhanced functional and structural domain assignments using remote similarity detection procedures for proteins encoded in the genome of *Mycobacterium tuberculosis H37Rv*. *J. Biosci.*29, 245–259. doi: 10.1007/bf02702607

Nealon, J. O., Philomina, L. S., and McGuffin, L. J. (2017). Predictive and experimental approaches for elucidating protein-protein interactions and quaternary structures. *Int. J. Mol. Sci.*18:E2623.

Nikolaus, T., Deiwick, J., Rappl, C., Freeman, J. A., Schroder, W., Millller, S. I., et al. (2001). SseBCD proteins are secreted by the type III secretion system of *Salmonella* pathogenicity island 2 and function as a translocon. *J. Bacteriol.*183, 6036–6045. doi: 10.1128/jb.183.20.6036-6045.2001

Notomista, E., Lahm, A., Di Donato, A., and Tramontano, A. (2003). Evolution of bacterial and archaeal multicomponent monooxygenases. *J. Mol. Evol.*56, 435–445. doi: 10.1007/s00239-002-2414-1

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*8, 785–786. doi: 10.1038/nmeth.1701

Podar, M., Makarova, K. S., Graham, D. E., Wolf, Y. I., Koonin, E. V., and Reysenbach, A. L. (2013). Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct.*8:9.

Poulsen, C., Akhter, Y., Jeon, A. H., Schmitt-Ulms, G., Meyer, H. E., Stefanski, A., et al. (2010). Proteome-wide identification of mycobacterial pupylation targets. *Mol. Syst. Biol.*6:386. doi: 10.1038/msb.2010.39

Prakash, D., Walters, K. A., Martinie, R. J., McCarver, A. C., Kumar, A. K., Lessner, D. J., et al. (2018). Toward a mechanistic and physiological understanding of a ferredoxin:disulfide reductase from the domains archaea and bacteria. *J. Biol. Chem.*293, 9198–9209. doi: 10.1074/jbc.ra118.002473

Prathiviraj, R., and Chellapandi, P. (2019). Functional annotation of operome from *Methanothermobacter thermautotrophicus* ΔH: An insight to metabolic gap filling. *Int. J. Biol. Macromol.*123, 350–362. doi: 10.1016/j.ijbiomac.2018.11.100

Prathiviraj, R., and Chellapandi, P. (2020). Comparative genomic analysis reveals starvation survival systems in Methanothermobacter thermautotrophicus ΔH. *Anaerobe*64:102216. doi: 10.1016/j.anaerobe.2020.102216

Prathiviraj, R., and Chellapandi, P. (2020a). Comparative genomic analysis reveals starvation survival systems in *Methanothermobacter thermoautotrophicus* ΔH. *Anaerobe*64:102216.

Prathiviraj, R., and Chellapandi, P. (2020b). Modeling a global regulatory network of *Methanothermobacter thermautotrophicus* strain ΔH. *Netw. Model. Anal. Health Inform. Bioinform.*9:17.

Pulendran, B., and Ahmed, R. (2006). Translating innate immunity into immunological memory: implications for vaccine development. *Cell*124, 849–863. doi: 10.1016/j.cell.2006.02.019

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., et al. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res.*33, W116–W120.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nat. Methods*10, 221–227.

Rath, V. L., Ammirati, M., LeMotte, P. K., Fennell, K. F., Mansour, M. N., Danley, D. E., et al. (2000). Activation of human liver glycogen phosphorylase by alteration of the secondary structure and packing of the catalytic core. *Mol. Cell.*6, 139–148. doi: 10.1016/s1097-2765(05)00006-7

Rawls, K. S., Yacovone, S. K., and Maupin-Furlow, J. A. (2010). GlpR represses fructose and glucose metabolic enzymes at the level of transcription in the haloarchaeon *Haloferax volcanii*. *J. Bacteriol.*192, 6251–6260. doi: 10.1128/jb.00827-10

Rimon, A., Kozachkov-Magrisso, L., and Padan, E. (2012). The unwound portion dividing helix IV of NhaA undergoes a conformational change at physiological pH and lines the cation passage. *Biochemistry*51, 9560–9569. doi: 10.1021/bi301030x

Rodionov, D. A., Hebbeln, P., Eudes, A., ter Beek, J., Rodionova, I. A., Erkens, G. B., et al. (2009). A novel class of modular transporters for vitamins in prokaryotes. *J. Bacteriol.*191, 42–51. doi: 10.1128/jb.01208-08

Rosch, J. W., Gao, G., Ridout, G., Wang, Y. D., and Tuomanen, E. I. (2009). Role of the manganese efflux system *mntE* for signalling and pathogenesis in *Streptococcus pneumoniae*. *Mol. Microbiol.*72, 12–25. doi: 10.1111/j.1365-2958.2009.06638.x

Sahraeian, S. M., Luo, K. R., and Brenner, S. E. (2015). SIFTER search: a web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res.*43, W141–W147.

Sangavai, C., Prathiviraj, R., and Chellapandi, P. (2020). Functional prediction, characterization and categorization of operome from *Acetoanaerobium sticklandii* DSM 519. *Anaerobe*61:102088. doi: 10.1016/j.anaerobe.2019.102088

Schlictman, D., Kubo, M., Shankar, S., and Chakrabarty, A. M. (1995). Regulation of nucleoside diphosphate kinase and secretable virulence factors in *Pseudomonas aeruginosa*: roles of *algR2* and *algH*. *J. Bacteriol.*177, 2469–2474. doi: 10.1128/jb.177.9.2469-2474.1995

Shahbaaz, M., Hassan, M. I., and Ahmad, F. (2013). Functional annotation of conserved hypothetical proteins from *Haemophilus influenzae Rd KW20*. *PLoS One*8:e84263. doi: 10.1371/journal.pone.0084263

Shapiro, L., and Harris, T. (2000). Finding function through structural genomics. *Curr. Opin. Biotechnol.*11, 31–35. doi: 10.1016/s0958-1669(99)00064-6

Shrivastava, A. K., Kumar, S., Sahu, P. S., and Mahapatra, R. K. (2017). *In silico* identification and validation of a novel hypothetical protein in *Cryptosporidium hominis* and virtual screening of inhibitors as therapeutics. *Parasitol Res.*116, 1533–1544. doi: 10.1007/s00436-017-5430-1

Singh, G., and Singh, V. (2018). Functional elucidation of hypothetical proteins for their indispensable roles toward drug designing targets from *Helicobacter pylori* strain HPAG1. *J. Biomol. Struct. Dyn.*1:13.

Singh, S., Singh, S. K., Chowdhury, I., and Singh, R. (2017). Understanding the mechanism of bacterial biofilms resistance to antimicrobial agents. *Open Microbiol. J.*11, 53–62. doi: 10.2174/1874285801711010053

Sivashankari, S., and Shanmughavel, P. (2006). Functional annotation of hypothetical proteins - A review. *Bioinformation*1, 335–338. doi: 10.6026/97320630001335

Terstappen, G. C., and Reggiani, A. (2001). *In silico* research in drug discovery. *Trends Pharmacol. Sci.*22, 23–26.

Thieringer, H. A., Jones, P. G., and Inouye, M. (1998). Cold shock and adaptation. *Bioessays*20, 49–57. doi: 10.1002/(sici)1521-1878(199801)20:1<49::aid-bies8>3.0.co;2-n

Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2002).) Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinform.*2, 2.3.

Tusnády, G. E., and Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics*17, 849–850. doi: 10.1093/bioinformatics/17.9.849

Unterholzner, S. J., Poppenberger, B., and Rozhon, W. (2013). Toxin-antitoxin systems: Biology, identification, and application. *Mob. Genet. Elements*3:e26219. doi: 10.4161/mge.26219

Vershon, A. K., Bowie, J. U., Karplus, T. M., and Sauer, R. T. (1986). Isolation and analysis of arc repressor mutants: evidence for an unusual mechanism of DNA binding. *Proteins*1, 302–311. doi: 10.1002/prot.340010404

von Heijne, G. (1988). Transcending the impenetrable: how proteins come to terms with membranes. *Biochim. Biophys. Acta*947, 307–333. doi: 10.1016/0304-4157(88)90013-5

Weinitschke, S., Denger, K., Cook, A. M., and Smits, T. H. (2007). The DUF81 protein *TauE* in *Cupriavidus necator H16*, a sulfite exporter in the metabolism of C2 sulfonates. *Microbiol*153, 3055–3060. doi: 10.1099/mic.0.2007/009845-0

Winnen, B., Hvorup, R. N., and Saier, M. H. (2003). The tripartite tricarboxylate transporter (TTT) family. *Res. Microbiol.*154, 457–465. doi: 10.1016/s0923-2508(03)00126-8

Xu, Z., Nie, P., Sun, B., and Chang, M. (2007). Molecular identification and expression analysis of tumor necrosis factor receptor-associated factor 2 in grass carp *Ctenopharyngodon idella*. *Acta Biochim. Biophys. Sin.*39, 857–868. doi: 10.1111/j.1745-7270.2007.00355.x

Yellaboina, S., Goyal, K., and Mande, S. C. (2007). Inferring genome-wide functional linkages in E. coli by combining improved genome context methods: comparison with high-throughput experimental data. *Genome Res.*17, 527–535. doi: 10.1101/gr.5900607

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*26, 1608–1615. doi: 10.1093/bioinformatics/btq249

Zarembinski, T., Hung, L. W., Mueller-Dieckmann, H. J., Kim, K. K., Yokota, H., Kim, R., et al. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl. Acad. Sci.U.S.A.*95, 15189–15193. doi: 10.1073/pnas.95.26.15189

Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., et al. (2017). RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.*45, D135–D138.

Zmijewski, M. A., Kwiatkowska, J. M., and Lipińska, B. (2004). Complementation studies of the *DnaK-DnaJ-GrpE* chaperone machineries from *Vibrio harveyi* and *Escherichia coli,* both *in vivo* and *in vitro*. *Arch. Microbiol.*182, 436–449. doi: 10.1007/s00203-004-0727-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 March 2021
doi: 10.3389/fgene.2021.643423

Comparative Metagenomic Analysis of Two Alkaline Hot Springs of Madhya Pradesh, India and Deciphering the Extremophiles for Industrial Enzymes

*Kamlesh Choure[1], Shreyansh Parsai[1], Rhitu Kotoky[2], Arpit Srivastava[1], Anita Tilwari[3], Piyush Kant Rai[1], Abhishek Sharma[4] and Piyush Pandey[2]**

[1] Department of Biotechnology, AKS University, Satna, India, [2] Department of Microbiology, Assam University, Silchar, India, [3] Centre of Excellence in Biotechnology, Madhya Pradesh Council of Science and Technology, Bhopal, India, [4] Amity Food and Agriculture Foundation, Amity University, Noida, India

OPEN ACCESS

Edited by:
Dhaval K. Acharya, B N Patel Institute of Paramedical, India

Reviewed by:
Digvijay Verma, Babasaheb Bhimrao Ambedkar University, India
Jitesh Kumar, University of Minnesota Twin Cities, United States

***Correspondence:**
Piyush Pandey
ppmicroaus@gmail.com; piyushddn@gmail.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *18 December 2020*
Accepted: *15 February 2021*
Published: *08 March 2021*

Citation:
Choure K, Parsai S, Kotoky R, Srivastava A, Tilwari A, Rai PK, Sharma A and Pandey P (2021) Comparative Metagenomic Analysis of Two Alkaline Hot Springs of Madhya Pradesh, India and Deciphering the Extremophiles for Industrial Enzymes. Front. Genet. 12:643423. doi: 10.3389/fgene.2021.643423

Hot springs are considered to be a unique environment with extremophiles, that are sources of industrially important enzymes, and other biotechnological products. The objective of this study was to undertake, analyze, and characterize the microbiome of two major hot springs located in the state of Madhya Pradesh explicitly, Chhoti Anhoni (Hotspring 1), and Badi Anhoni (Hotspring 2) to find out the inhabitant microbial population, and their functional characteristics. The taxonomic analysis of the microbiome of the hot springs revealed the phylum Proteobacteria was the most abundant taxa in both the hot-springs, however, its abundance in hot-spring 1 (∼88%) was more than the hot-spring 2 (∼52%). The phylum Bacteroides (∼10–22%) was found to be the second most abundant group in the hot-springs followed by Spirocheates (∼2–11%), Firmicutes (∼6–8%), Chloroflexi (1–5%), etc. The functional analysis of the microbiome revealed different features related to several functions including metabolism of organics and degradation of xenobiotic compounds. The functional analysis showed that most of the attributes of the microbiome was related to metabolism, followed by cellular processes and environmental information processing functions. The functional annotation of the microbiomes at KEGG level 3 annotated the sequences into 279 active features that showed variation in abundance between the hot spring samples, where hot-spring 1 was functionally more diverse. Interestingly, the abundance of functional genes from methanogenic bacteria, was higher in the hot-spring 2, which may be related to the relatively higher pH and temperature than Hotspring 1. The study showed the presence of different unassigned bacterial taxa with high abundance which indicates the potential of novel genera or phylotypes. Culturable isolates (28) were bio-prospected for industrially important enzymes including amylase, protease, lipase, gelatinase, pectinase, cellulase, lecithinase, and xylanase. Seven isolates (25%) had shown positive results for all the enzyme activities whereas 23 isolates (82%) produced Protease, 27 isolates (96%) produced lipase, 27 isolates produced amylase, 26 isolates

(92%) produced cellulase, 19 isolates (67%) produced pectinase, 19 isolates (67%) could produce lecithinase, and 13 isolates (46%) produced gelatinase. The seven isolates, positive for all the enzymes were analyzed further for quantitative analysis and identified through molecular characterization.

Keywords: microbiome, Hotsprings, extremophiles, microbial diversity, industrial enzymes

INTRODUCTION

Extremophilic microorganisms thrive in diverse and extreme conditions and constitute a major part of the biosphere (Mirete et al., 2016). The thermophiles and hyper-thermophiles live in high-temperature environments such as hot springs, though few of these can survive in co-existing, more than one extreme conditions, like acidic or alkaline hot springs. The accessibility of thermophiles to survive at high temperatures is related to their incredibly thermostable macromolecules present in them (Brock, 2001). These thermophilic microorganisms have been studied extensively for thermostable enzymes such as amylases, cellulases, chitinases, pectinases, xylanases, proteases, lipase, and DNA polymerases, etc. that has unique features of biotechnological processes (Singh et al., 2011).

Thermophilic microorganisms are an excellent source of thermostable enzymes and have been utilized in the greater part of industrial applications, for example, food, papers, pharmaceutical, cleansers, etc. (Schuler et al., 2017; Roy et al., 2020). Thermophilic microorganisms are also more stable than their mesophilic partners to natural solvents, cleansers, low and high pH, and other extreme conditions (Demirjian et al., 2001). Therefore, industrially important enzymes from thermophiles such as amylase (extracellular), protease (extracellular), lipase (extra/intracellular), gelatinase (extracellular), pectinase (extra/intracellular), cellulase (extra/intracellular), lecithinase (extracellular), and xylanase (extracellular) has been used extensively. Most of these enzymes are found to be optimally active at temperatures close to the host organism's optimal growth temperature. However, some of the extracellular and cell-bound hyperthermophilic enzymes were optimally active at temperatures above, sometimes far above than the host organism's optimum growth temperature (Vieille and Zeikus, 2001).

The Geological Survey of India has identified about 340 hot springs located in different parts of India, which are characterized by their orogenic activities (Chandrasekharam, 2005; Craig et al., 2013). All these hot springs have been classified and grouped into nine geothermal provinces based on their geo-tectonic setup that includes the Himalayas, Naga-Lushai province, Sohana, West coast, Andaman-Nicobar Islands, Cambay, Son-Narmada-Tapi (SONATA), Godavari, and Mahanadi valleys. Geothermal resources along Son-Narmada lineament viz. Choti and Badi Anhoni form the most promising resource base in central India (Shanker, 1986). The lineament is one of the most important lineaments/rifted structure of the sub-continent. It runs across the country in an almost East-West direction and has a long history of tectonic reactivation. It contains several known thermal spring areas, the most interesting one being those situated at Anhoni (Saxena et al., 2017). There are several hot springs situated in Madhya Pradesh at several locations like Anhoni in Chhindwara district, hot and boiling sulfur springs that flow along within the forest. Anhoni is particularly known for its 'boiling water *kund*' (*kund* means a small pond), Choti Anhoni near Pipariya, Badi Anhoni near Panchmarhi, Chavalpani at Pachmarhi, Anhoni Samoni (it is different from the aforementioned Anhoni springs), Babeha hot spring is in the Mandla district, and Dhuni Pani, Amarkantak. The alkaline hot springs have pH more than seven and can range from 8.5 to 12. Other alkaline hot-springs have also been studied, from other parts of world, such as the Great Rift Valley in northeastern Africa, which has been characterized to have high levels of carbonates, chlorides, and silica compounds (Jones et al., 1998). The organisms surviving in such alkaline hot springs acquire necessary adaptations. The bacteria present in such environments are either alkaliphilic or alkalitolerant, that are known as alkalithermophilic bacteria, and these organisms have enzymes to support their growth and survival in such extreme conditions. These alkalithermophiles are often reported to be chemolithoautotrophic (Sorokin and Kuenen, 2005).

Several studies have been done to analyze the microbial diversity of different hot-springs around the globe. The microorganisms growing in different ecological zones (e.g., hot springs and deep-sea) can be categorized into moderate thermophiles (growth optimum, 50–60°C), extreme thermophiles (growth optimum, 60–80°C), and hyperthermophiles (growth optimum, 80–110°C) (Gupta et al., 2014). The natural habitats of the thermophiles include continental solfataras, deep geothermally heated oil-containing stratifications, shallow marine, and deep-sea hot sediments, and hydrothermal vents. The hyperthermophiles have also been isolated from hot industrial environments. These hyperthermophiles with the highest growth temperatures are members of the genera Pyrobaculum, Pyrodictium, Pyrococcus, and Melanopyrus belonging to Archaea. However, the isolation and growth of pure cultures of novel hyperthermophiles has been a challenge, which mostly remains unculturable, and may be assessed using metagenomics and next-generation sequencing technologies (López-López et al., 2013).

The present study was taken to analyze the taxonomical and functional diversity of the microbiome of two alkaline hot-springs with idea to analyze the genetic pool of thermophilic microorganisms as a source of industrially important enzymes. This research describes the insights of their microbial diversity, including strategies followed by enzyme screening and quantifications.

MATERIALS AND METHODS

Collection of Samples

Water samples were collected from the Choti and Badi Anhoni Hot Springs (22.65°N latitude and 78.36°E longitude) situated in Panchmari, Madhya Pradesh (India). Physiological parameters of water samples were measured on-site using HANNA HI2300 EC/TDS/NaCl multi-probe system according to the manual. The sample of Choti Anhoni and Badi Anhoni are designated as Hotspring 1 and Hotspring 2, respectively.

Isolation and Characterization of Thermophilic Bacteria

The thermophilic bacteria from the water samples were isolated according to methods described by Adiguzel et al. (2009), through the serial dilution method. Thermophilic bacteria were isolated and cultured on Nutrient agar plates, the pH of the medium was adjusted to 7.0 before autoclaving and then incubated at 45°C for 24–48 h (Sikdar et al., 2015). Isolation of pure cultures was done using the streak plate method and the cultures were stored for enzyme screening analysis.

The selected bacterial isolates (positive for all the enzyme activity, as tested) were subjected to identification based on 16S rDNA gene sequencing. DNA isolated from the bacterial isolates was directly used for PCR amplification of 16S rRNA gene using 1492R (5′CGGTTACCTTGTTACGACTT3′) and 27F (5′AGAGTTTGATCMTGGCTCAG3′) universal primers. The sequence obtains after sequencing were used for the *in silico* study to obtain the highest similarity using online web server nucleotide BLASTN based on the BLAST alignment.

Analysis of Extracellular Enzymes

The isolates were analyzed for different extracellular enzymes of industrial importance like protease, lipase, amylase, xylanase, cellulase, pectinase, lecithinase, and gelatinase. The screening for protease activity was performed as described (Bragger et al., 1989), on skim milk agar containing 8 g/L nutrient broth, 10 g/L skim milk, and 17 g/L agar, then incubated for 36 h at 45°C. The presence of protease activity was confirmed by the appearance of clear zones around the well indicating degradation of casein milk.

The lipase activity of the isolates was performed according to the method described by Haba et al. (2000), on a medium containing 8 g/L nutrient broth, 0.25 g/L CaCl2.2H2O, 9 g/L agar dissolved in 500 mL deionized water, and 5 mL of Tween 20 dissolved in 500 mL deionized water autoclaved separately and add to the medium, then the medium with the cultures incubated for 2 days at 45°C. Clear zones that occur around the colonies indicated the presence of lipase activity. The screening of the amylase activity was performed as described (Bragger et al., 1989), on a medium containing 1 g/L yeast extract, 5 g/L soluble starch, and 17 g/L agar. Ingredients were dissolved in deionized water and sterilized by autoclaved and incubated for 1–2 days at 45°C. The presence of amylase activity was confirmed by the appearance of a clear halo around the well after the color with iodine.

The xylanase activity was performed according to the method described by Bragger et al. (1989), on a medium containing 1 g/L yeast extract, 5 g/L xylans, and 17 g/L agar, which was incubated for 3–4 days at 45°C. The activity of the xylanase enzyme was confirmed by the appearance of a clear zone around the tested strain following the staining with Congo Red. Similarly, the activity of cellulase was performed according to the method described by Bragger et al. (1989), on a medium containing 1 g/L yeast extract, 5 g/L carboxymethyl cellulose (CMC) salt, and 17 g/L agar then incubated for 3–4 days at 45°C. Cellulase activity resulted in the appearance of a clear zone around the tested strain after treatment with iodine. Identification of bacterial isolates displaying pectinase activity was performed according to Bragger et al. (1989), on a medium containing 1 g/L yeast extract, 2 g/L ammonium sulfate, 6 g/L Na2HPO4, 3 g/L KH_2PO_4, 5 g/L pectins, and 17 g/L agar. Ingredients were dissolved in deionized water and sterilized by autoclaved at 121°C for 15 min and incubated for 3–4 days at 45°C. Colonies with clear zones indicated pectinase activity.

Lecithinase production was tested on a modified medium as described previously (Oladipo et al., 2008). Lecithinase was detected according to the standard method (Sharaf et al., 2014), in which 1 ml of each bacterial culture, having cell density of 6 $\times$ 10^8 CFU/ml was inoculated into test tubes containing corn millet broth and incubated for 24 h at 37°C. After incubation, the cultures were centrifuged at 2500 rpm for 15 min to obtain a cell-free filtrate, and 100 μl of the filtrate was transferred into 10-mm wells made centrally in the egg-yolk agar plates and incubated for 24 h at 37°C. Opaque zones were measured as indicators of lecithinase production. Gelatinase production was detected by stab inoculating the test strain on nutrient agar supplemented with 3% gelatin kept at 37°C for 24 h followed by refrigeration at 4°C for 30 min. Liquefaction of gelatin was taken positive (Betty et al., 2007).

Quantitative Estimation of Enzyme Activities

The isolates positive for all the tested enzymatic production were further analyzed for the quantitative estimation of enzyme activities at ambient temperatures and pH (of respective site, as described in **Table 2**) i.e., 55/65°C and 8.5/9.5, respectively. To determine the cellulase activity, colorimetric assay by DNS (Dinitro salicylic acid) method was used (Miller, xbib1995). Samples were subjected to incubation for 30 min with CMC (Carboxymethyl cellulose) as substrate followed by the addition of DNS and boiled for 6–7 min, and absorbance was taken at 540 nm. Similarly for amylase activity, 1 ml of enzyme solution was incubated with substrate solution, containing 1% (w/v) (1ml) soluble starch at 55/65°C for 30 min followed by the addition of DNS to stop the reaction and kept at boiling water bath for 10 min (Bernfeld, 1955). Lipase enzyme activity was performed by using 1% tributyrin in basal salt media. P-Nitrophenol dodecanote was used as a substrate to determine lipase activity. The reaction mixture containing enzyme solution and P-Nitrophenol dodecanote was incubated for 30 min at 55/65°C. For the protease activity, casein is used as a substrate, and the reaction mixture was composed of 2.5 ml of the substrate and 1 ml cell-free extract enzyme solution followed by the incubation

TABLE 1 | Physicochemical analysis on water sample.

Sample	Temperature (area) °C	Temperature (sample) °C	pH	Humidity %	Electrical Conductivity ms	Turbidity NTU	DO mg/L	BOD mg/ml	COD mg/L	TDS (ppm)	NaCl Conc
Hot Spring 1	25.4	50-55	8.5	34	791	7.12	9	187	389	412	0.02
Hot Spring 2	33.2	60-65	9.5	58	827	5.96	17.1	120	1144	376	0.02

at 55/65°C for 30 min. Trichloroacetic acid is used for the termination of the reaction. Gelatinase activity was measured by using gelatin as a substrate, where 0.2 ml of 50% Trichloroacetic acid was used to terminate the reaction. The Lecithinase activity was performed on 10 ml 50% egg yolk in basal salt media. The activity was measured by using the method described by McLaughlin and (McLaughlin and Weiss, 1996). All the observations were recorded by taking absorbance at 540 nm using a spectrophotometer.

Microbiome Analysis of the Hot-Springs

The water samples from the hot springs were collected and analyzed for the microbiomes through metagenome analysis of the hypervariable V3–V4 region. The DNA was obtained from the water samples using the Nucleospin DNA kit. The amplicon libraries were prepared using the Nextera Index kit as per the 16S metagenomic sequencing library preparation protocol. For this, 16S rDNA specific primers were used for bacterial V3–V4. The libraries were sequenced on MiSeq using a 2×300 bp paired-end manner. The amplicons with the Illumina adapters were amplified by using i5 and i7 primers and purified by AMPureXP beads and quantified using a Qubit fluorometer. After that, the libraries were loaded onto Miseq at the appropriate concentration for cluster generation and sequencing (Kotoky and Pandey, 2020).

Quality Control was performed using the online FastQC tool v 0.11.7. Read quality was good with an average of more than 200,000 (2 lakh) reads per sample and a read length of 300 bp. High-quality reads were taken for further analysis. The fastq-Join tool was used to convert the overlapping paired-end reads into a consensus sequence of the V3–V4 region. It finds the overlap for each pair and combines them into a single read. In the Pre-processing step, Chimeric sequences were filtered out using the parameter reference_chimera_detection default implemented in the QIIME tool. OTU Picking and Taxonomic classification were performed using the UCLUST method in the QIIME. Reads from all samples were pooled and clustered into Operational Taxonomic Unit (OTU) based sequence similarity of >=97% with help of UCLUST method with reference to green gene database. Finally, 485 OTUs were identified at the species level.

After sequencing the paired-end sequences were analyzed as described by Kotoky and Pandey (2020). The Quantitative Insights into Microbial Ecology (QIIME2, version 2019.7) was used for the analysis of the samples (Bolyen et al., 2019). Sequences were clustered into operational taxonomic units (OTUs) using the Uclust algorithm at 97% sequence similarity (Edgar, 2010). The taxonomies were assigned to the OTUs by aligning the reads against the Greengenes Database (version 13_8) (McDonald et al., 2012) based on a threshold of 97% sequence similarity. The functional metagenomic profile and metagenomic content of the samples were predicted from the 16S rRNA profiles, and KEGG pathway functions were categorized at level 3 using the phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) tool (Langille et al., 2013) and visualized using STAMP (Software package for analyzing taxonomic or metabolic profiles) tool.

Statistical Analysis

Weighted and unweighted UniFrac distances analysis of the samples was done from the normalized OTU table. Alpha-Diversity values of the samples were calculated by the function using the Shannon method in QIIME2 and R to obtain the observed faith-pd, Shannon entropy, observed features, and pielou-evenness (Kotoky and Pandey, 2020).

RESULTS

Physicochemical Analysis of the Samples

The physicochemical parameters of water samples are mentioned in **Table 1**. The temperature of the sample Hot-spring2 was comparatively higher, but its turbidity was lesser than Hot-spring1. The pH was recorded higher with temperature range, variable conductivity, and salinity. Dissolved oxygen (DO), Biological oxygen demand (BOD), and chemical oxygen demand (COD) were measured to understand the level of oxygen concentration. Both samples Hot-spring1 and Hot-spring2 had shown BOD in the normal range but the COD of sample Hot-spring2 was found to be much higher than Hot-spring1 demonstrating the presence of more organics in the water. The Total Dissolved Solids (TDS) was also under the good range for both the samples. The hot springs were chosen for the study due to their different conditions of pH and temperature. Both the hot springs were found to be alkaline but with different temperatures (55 and 65°C). The sample Hot-spring2 had a relatively high concentration of salts than Hot-spring1.

Isolation of Thermophiles and Analysis of the Activity of Enzymes

From the two hot spring samples, 28 thermophilic bacterial isolates were isolated. The isolated bacterial strains were analyzed for the production of different industrial enzymes such as protease, lipase, amylase, cellulase, pectinase, xylanase, gelatinase, and lecithinase. From the isolates, seven isolates, including–CAP1, CAP3, CAP7, BAC18, BAC23, BAC26, BAC28 showed excellent potential for enzyme production.

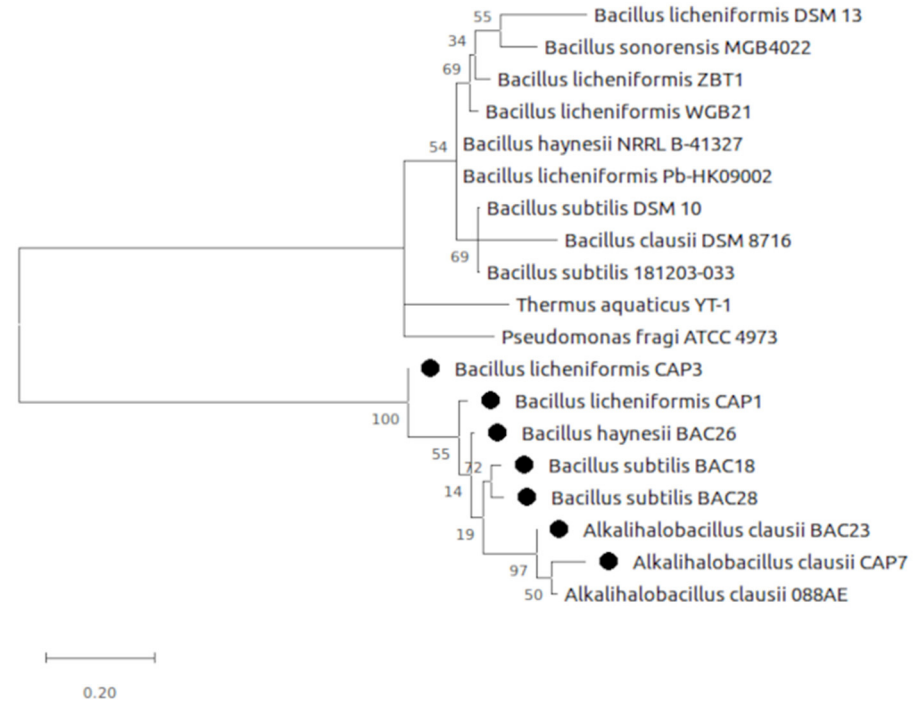

FIGURE 1 | Evolutionary analysis by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method and Tamura-Nei model (Tamura and Nei, 1993). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGA X (Kumar et al., 2018).

TABLE 2 | Production of enzyme at ambient temperature (T_{opt}) and pH (pH_{opt}).

Isolates	T_{opt} °C	pH_{opt}	Amylase U ml^{-1}	Lipase U ml^{-1}	Cellulase U ml^{-1}	Protease U ml^{-1}	Lecithinase U ml^{-1}	Gelatinase U ml^{-1}
Bacillus licheniformis CAP1	55	8.5	59.83	43.07	60.11	62.14	27.15	26.23
Bacillus licheniformis CAP3	55	8.5	61.26	36.10	39.26	36.28	31.19	28.21
Alkalihalobacillus clausii CAP7	55	8.5	61.19	44.23	43.28	53.12	28.54	33.23
Bacillus subtilis BAC18	65	9.5	58.17	47.36	51.18	31.15	33.33	35.71
Alkalihalobacillus clausii BAC23	65	9.5	55.23	51.25	38.20	42.18	37.39	41.26
Bacillus haynesii BAC26	65	9.5	51.85	33.09	48.51	51.67	28.19	51.23
Bacillus subtilis BAC28	65	9.5	49.87	38.67	51.19	45.83	29.24	24.18

Twenty-three isolates (82%) produced protease, 27 isolates (96%) produced lipase, 27 isolates produced amylase, 26 isolates (92%) produced cellulase, 19 isolates (67%) produced pectinase, 13 isolates (46%) produce gelatinase and 19 isolates (67%) could produce lecithinase. The study showed, all isolated thermophilic bacteria showed enzyme activities for at least three enzymes.

The selected bacterial isolates were characterized and identified as *Bacillus licheniformis* CAP1, *Bacillus licheniformis* CAP3, *Alkalihalobacillus clausii* CAP7, *Bacillus subtilis* BAC18, *Alkalihalobacillus clausii* BAC23, *Bacillus haynesii* BAC26, and *Bacillus subtilis* BAC28. The phylogenetic analysis of the isolates placed the organisms in at distinct branches of the dendogram (**Figure 1**). The quantitative analysis of the enzymes revealed that *B. licheniformis* CAP1 produces the highest amount of protease (62.14 U/ml) at given ambient temperature and pH, also showed good production of cellulase and amylase, 60.11 U/ml and 59.83 U/ml, respectively. Cellulase activity was also found to be maximum for isolate CAP1. *Bacillus licheniformis* CAP3 produced the highest amount of amylase (61.26 U/ml) and *Alkalihalobacillus clausii* BAC23 produced the highest activity of lipase (51.25 U/ml). The activity of lecithinase was found to be less than other enzymes and was in the range of 27–37 U/ml. Gelatinase activity was observed highest in *Bacillus haynesii* BAC26 (51.23 U/ml) while other isolates showed less production of the enzyme at given ambient temperature and pH. Conclusively, all the seven isolates were observed in amylase production ranges from 49 to 61 U/ml, lipase in the range 33–51 U/ml, cellulase in the range 38–60 U/ml, protease in 31–62 U/ml, and gelatinase 24–51 U/ml (**Table 2**).

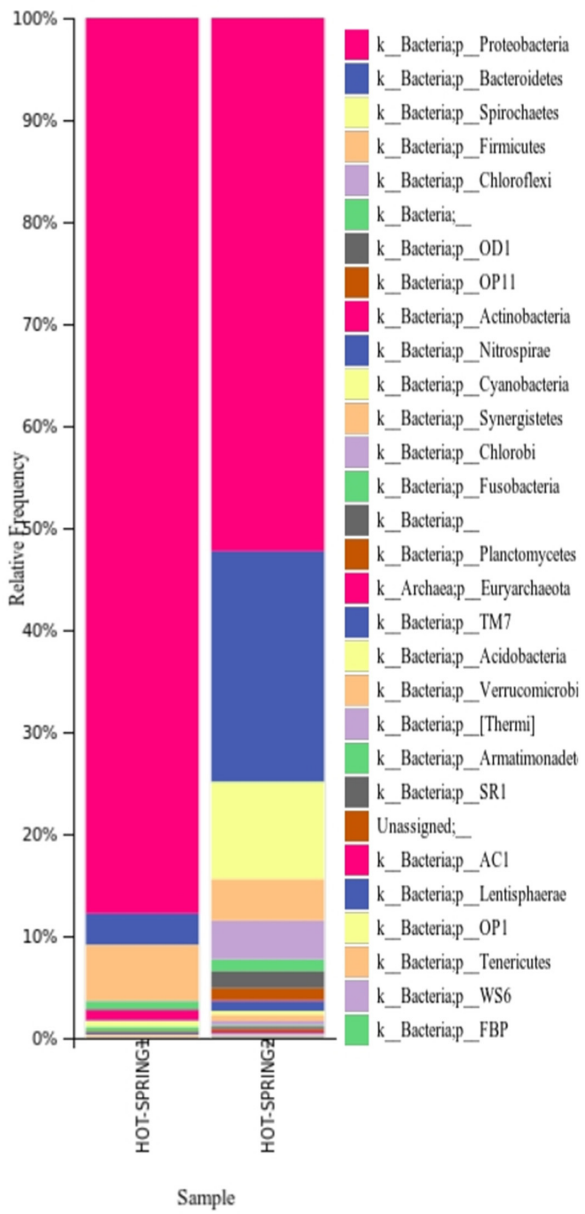

FIGURE 2 | Distribution of different taxa in the hot-spring samples analyzed at phylum level.

Composition of Microbial Community

The taxonomic analysis of the microbiome of the hot springs showed a predominance of bacteria and relatively very less proportion of archaea. In both the samples, the phylum Proteobacteria was found to be more abundant as plotted (**Figure 2**) however, the abundance of proteobacteria in hot-spring 1 (∼88% of total abundance) was more than the hot-spring 2 (∼52%). The phylum Bacteroides was found to be the second most abundant group (∼10–22%) in the hot-springs but very different from each other. The other phyla with more abundance were Spirocheates, Firmicutes, Chloroflexi, etc.

TABLE 3 | Alpha-diversity indices of the samples.

Samples	faith_pd	shannon_entropy	pielou_evenness
Hot-spring 1	72.33	4	0.44
Hot-spring 2	48.78	5.24	0.65

TABLE 4 | Alpha-diversity of the microbiome of the hot-springs of different part of world.

Sample origin	Temperature °C	pH	Number of distinct species	References
China	65	7	457.73	Menzel et al., 2015
Colombia	29	2.7	467.61	Jiménez et al., 2012
Iceland	85–90	5	196.14	MGRAST ID: mgm4530143.3
Italy	76	3	86.12	MGRAST ID: mgm4529716.3
Russia	61–64	5.8–6	615.97	MGRAST ID: 4544453.3
Spain	76	8.2	330.87	Lopez-Lopez et al., 2015
India	55–65	8.5–9.5	410.5	This study

The analysis at the genus level showed a very high abundance of an unknown genus from family commamonadaceae in both samples. The alpha-diversity analysis was done on the processed data and the faith_pd, shannon_entropy, and pielou_evenness of the samples have been calculated. The analysis showed that the hot-spring 1 was more diverse and had diversity richness (**Table 3**). The observed diversity in the microbiome was then compared with the alpha-diversity of different hot-spring samples of a different part of the world (**Table 4**). Which showed that the pH and geographical location of the hot-springs play a very crucial role in shaping their microbial diversity.

Functional Analysis of the Microbiome of the Hot Springs

The functional analysis of the microbiome revealed different features related to several functions categorized at different KEGG (Kyoto Encyclopedia of Genes and Genomes) level annotations. KEGG system analysis at level 1, significant differences in the abundance of genes for the different subsystems between the two samples. In hot-spring 1 the attributes related to cellular processes and environmental information processing were found to be significantly higher than hot-spring 2. However, the hot-spring 2 sample had greater attributes for genetic information, metabolism, and human diseases. The most of predicted protein sequences were associated with different functions related to metabolism (48–52%), environmental information processing (13–18%), genetic information processing (12–16%), and cellular processes (2–4%).

The functional prediction and annotation of the microbiomes at KEGG level 2, revealed a predominance of genes belonging

to carbohydrate metabolism, amino acid metabolism, and membrane transport. The clustering of the attributes was done using the UPGMA method with a threshold of 0.75, which clustered the similar abundant attributes in both samples.

The functional prediction and annotation of the microbiomes at KEGG level 3 annotated the sequences into 279 active features that showed variation in abundance between the samples. From the active features, 39 features were selected for analysis related to carbohydrate, protein, and fat metabolism and attributes related to the degradation of xenobiotic compounds. The functional analysis showed hot-spring 1 as more diverse functionally and have more abundance of attributes related to ABC transporters, amino acid metabolism, and genes for degradation of xenobiotic compound degradation. However, hot-spring 2 showed more abundance genes of metabolism of carbohydrates, lipids, and proteins, showing a greater abundance of functions related to industrial enzymes (**Figure 3**). The abundance of pathways related to ABC transporter (ko02010), bacterial motility proteins (ko02030), benzoate degradation (ko00362), starch and sucrose metabolism (ko00500), beta alanine metabolism (ko00410) were found to be significantly different between the two samples. On the other hand, the KEGG pathways related to calcium signaling pathway (ko04020), lipid biosynthesis process (ko00061), glycan biosynthesis (ko00510) etc. were found to be significantly low in abundance and less diverse between the samples.

Data Availability

The metagenomic sequences of the samples were deposited in NCBI, at Sequence Read Archive (SRA) under the accession number SRP13358614 and SRP13358615; Bioproject ID PRJNA688206 and BioSample ID- SAMN17170341 (Hot-spring 1) and SAMN17170342 (Hot-spring 2). The 16s rDNA sequences of the selected bacterial isolates were submitted to NCBI Genbank under the accession numbers MW527298 (*Bacillus licheniformis* CAP1), MW527299 (*Alkalihalobacillus clausii* CAP7), MW527300 (*Bacillus subtilis* BAC18), MW527301 (*Bacillus licheniformis* CAP3), MW527302 (*Alkalihalobacillus clausii* BAC23), MW527303 (*Bacillus haynesii* BAC26), and MW527304 (*Bacillus subtilis* BAC28).

DISCUSSION

The hot springs are considered to be the source of untapped microbial diversity, that are a source of enzymes of industrial importance. Therefore, the microbial diversity of two hot springs has been characterized and the functional roles of the microbiome were predicted using the metagenomics approach. Further, seven potential isolates were cultured and found efficient for industrially potential enzymes, active at high temperatures. Several studies of hot spring environments have focused on the relationship between microbial communities and different environmental factors especially temperature, which is believed to be the main factor that drives the community structure (Skirnisdottir et al., 2000). However, different other factors like available organic carbon, total dissolved solids, salt concentration also play a crucial role in shaping the microbial community structure as the microbial diversity do not have a monotonic relationship with temperature, where different other environmental or spatial factors may also be responsible for determining the microbial community (Purcell et al., 2007). It has been reported that the microbial community structures were different in the low- and high-sulfide hot spring mats with the same temperature (Skirnisdottir et al., 2000). Moreover, the hyperthermophilic archaeal communities are different in various hot springs. Therefore, the environmental and spatial variables play an important role in shaping microbial community compositions in natural ecosystems (Zhang et al., 2018). Power et al. (2018) and Uribe-Lorio et al. (2019) have reported that pH has a strong influence on the microbial community structure, where the influence of temperature was significant only at values above 70°C. Purcell et al. (2007) also reported that the high temperature (75–90°C) and alkaline pH (7.5–9) were the most influencing factors shaping the microbial community of the hot springs of Thailand.

Hot springs are the main source of microbial diversity to find industrially important enzymes (Sahay et al., 2017). The thermostable enzymes are stable and active even at temperatures higher than the optimal growth temperature showing potential for numerous industrial applications. Moreover, these enzymes have been reported to be more stable also against many solvents, detergents, and acidic and alkaline pH (Bhalla et al., 2013). Mohammad et al. (2017) also reported 10 thermophilic bacteria isolated from Jordanian Hot-spring could produce a wide range of thermostable enzymes leading to potential applications of the bio catalyzed processes in harsh conditions. Different thermostable bacterial enzymes like α-amylase, protease, and lipase have been used extensively in industrial processes. These thermophilic and hyperthermophilic enzymes are part of the enzyme category called extremozymes, which involve functions at extreme conditions like high salt levels, high alkaline conditions, or under extreme conditions of pressure or acidity (Vieille and Zeikus, 2001). The stability of the enzymes depends on the thermodynamic and kinetic stabilities. In the present study, the activity of the enzymes at ambient temperature was found to be high and have very good potential to be used for production.

The culture depended analysis of the bacterial population of the hot springs led to the isolation of 28 bacterial isolates that showed good enzyme activity of industrial importance. From the isolates, seven isolates were identified as having all the enzyme activity and were from phyla firmicutes. However, a culture-independent analysis of the microbiome of the hot-springs showed many unidentified classes and families, which are still left to be investigated. The taxonomic identification of the microbiome was done using Greengene classifier revealing many known and unknown bacterial taxa, and proteobacteria as most abundant. Different other studies also reported proteobacteria as dominant taxa in the hot springs with moderately high and very high temperatures (44–110°C) at various geographical locations, including India (Chan et al., 2015; Ghelani et al., 2015). Different earlier studies have suggested a decrease in diversity of the microbial community with increasing environmental temperature (Mathur et al., 2010; Valverde et al., 2012). Interestingly, the taxonomical analysis showed the hot springs

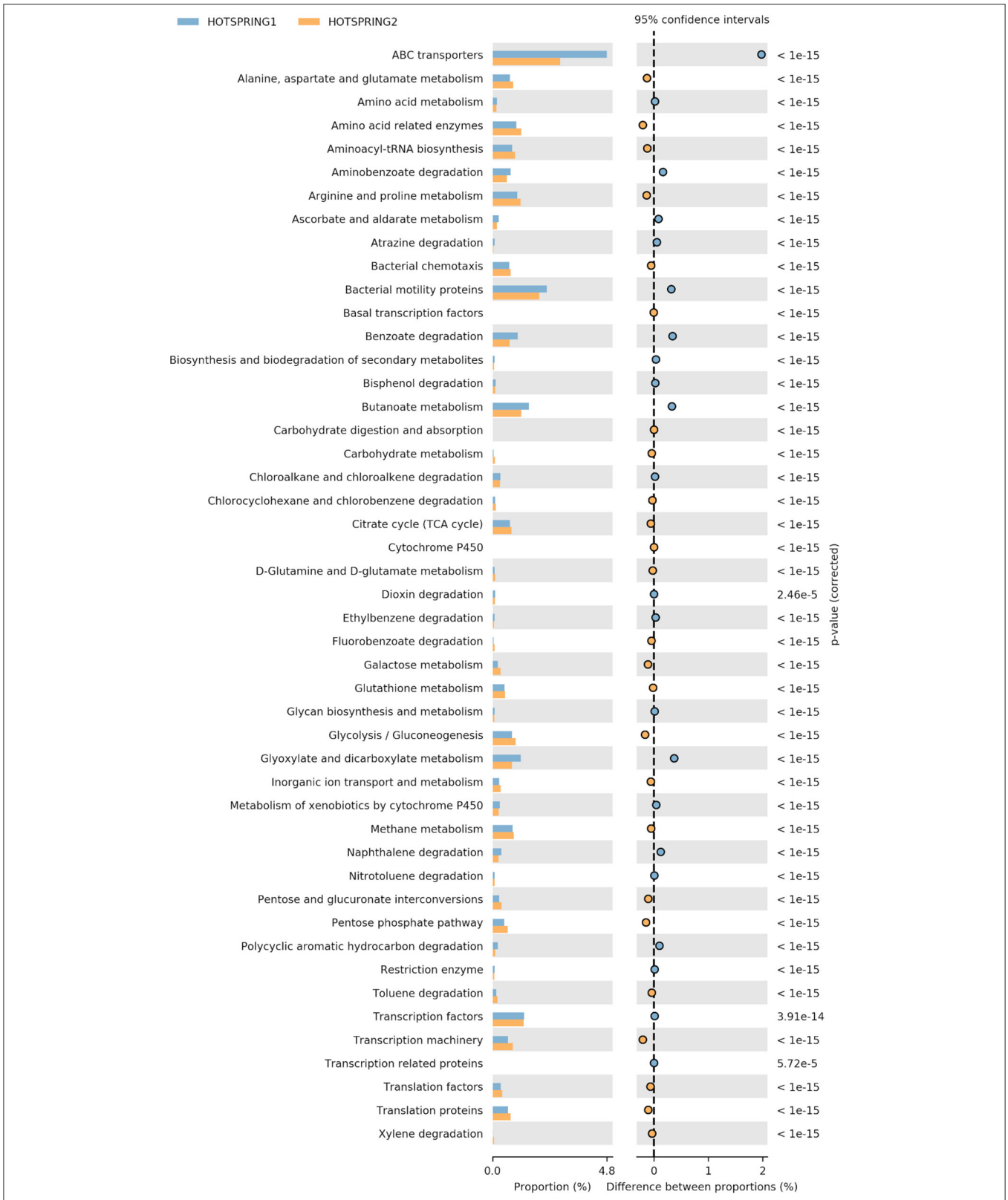

FIGURE 3 | Variation in abundance of selected attributes between the hot-spring samples annotated at KEGG level 3.

has a diverse and different pattern of abundance although both have different temperature, pH, and the influx of organic material. Thus, it can be assumed that the community structure is largely determined by a combination of environmental parameters, rather than geographical distance.

The taxonomic and functional study of the microbial ecology in the hot-springs showed the influence of environmental factors like temperature, pH on the microbiome that boost the metabolism pattern and enhance the stress biology. The microbiome contains the functional groups that perform various metabolic functions. The metabolism of methane was found to be higher in hot-spring 2 with higher pH and temperature. The presence of a large number of phylogenetically diverse, metabolically divergent groups indicates a balanced complex community, where each group occupies its environmental niche. The temperature of the hot-springs was found to be in the range of 55–65°C, different from each other. However, several studies reported that the temperature is not a unique determinant of microbial diversity and its function in the hot springs (Huang et al., 2011; Wang et al., 2013). Importantly, the pH of the springs was found in the range of moderate alkaline (8.5–9.5). The higher pH ranges have been reported to significantly impact the biodiversity of certain biological niches leading to the association of different adapted microbial groups. As reported by Tyson et al. (2004), the acidic pH of mine drainage site in Iron Mountain, California, USA (pH 0.83, 42°C) led to the selection of a very simple community dominated by an extremophilic Leptospirillum and Ferroplasma. At alkaline pH range also, the effect is reported to be similar. Therefore, the microbiome of hot-spring 2 (pH 9.5) in the present study was found less diverse than hot-spring 1 (pH 8.5). It has been reported that alkaline hot springs with a lower temperature below 73°C are typically dominated by cyanobacteria (Pedersen and Miller, 2016). However, in contrast to that, the hot-springs of the present study the cyanobacteria phylum was not on the higher abundance side, instead, the phylum spirochaetes were found to be very high in abundance in the sample Hot-spring 2 which was not that abundant in sample Hot-spring 1. Therefore, the effect of pH also playing a very crucial role in microbiome function and taxonomy, where the effects are both direct and indirect.

Several previous studies have reported different type of microbial structure in alkaline hot springs. Lopez-Lopez et al. (2015) described the bacterial phyla Deinococcus-Thermus as the most dominant in a alkaline Hot Spring in Galicia (Spain), followed by Proteobacteria (13%), and Firmicutes (10%). The archaea phylum Thaumarchaeota (6%) was found to be most abundant. Similarly, other studies also reported high occurrence of Thaumarchaeota in the archaeal fraction in alkaline springs from Kamchatka and China (Huang et al., 2011; Wemheuer et al., 2013). Menzel et al. (2015) reported that relative abundance of Archaea in hot springs is higher in low pH and high temperature environments. However, in the present study, at higher pH and temperatures very low abundance of archaea was observed. Interestingly, it has been reported that the most common substrate in alkaline hot-spring is hydrogen and sulfur (Horikoshi, 1999). These alkaliphilic microbes have adapted to such conditions through different mechanisms including the presence of cytoplasmic polyamines with charged amino acids. In *Bacillus* spp., in addition to peptidoglycan, there are acidic compounds such as galacturonic acid, gluconic acid, glutamic acid, aspartic acid, and phosphoric acid that act as buffers to the alkaline environment, allowing uptake of hydronium ions and exclusion of hydroxide ions (Horikoshi, 1999).

CONCLUSION

The culture-dependent analysis of the water samples of the hot-springs led to the isolation of several bacterial strains having good enzymatic activities with significant industrial importance. The culture-independent analysis showed that the taxonomical and functional diversity of the hot springs were distinct and is possibly shaped by temperature, pH, and organic materials. The study showed the presence of different unassigned bacterial taxa with great abundance which indicates the potential of novel genera or phylotypes. Different taxa were found to be more prominent in higher temperature than others and it was observed that multiple factors like pH, salinity also play a great role in shaping a microbiome. The functional analysis of the microbiomes revealed that most of the genes are associated with functions related to metabolism and environmental information processing. The analysis showed the presence of metabolic and biosynthesis pathways of different primary substrates including carbohydrates, fats, proteins etc. which display its industrial importance. The microbiome study showed that the hotspring 1 with low temperature and pH was more diverse taxonomically and functionally.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

KC and SP did the sample collection and screening of the enzymes and wrote the first draft of the manuscript. RK did the metagenome analysis of the samples and wrote the second draft of the manuscript. ASri, AT, PR, and ASha did the characterization and shared ideas. PP did the conceptualization and revised the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Adiguzel, A., Ozkan, H., Baris, O., Inan, K., Gulluce, M., and Sahin, F. (2009). Identification and characterization of thermophilic bacteria isolated from hot springs in Turkey. *J. Microbiol. Methods* 79, 321–328. doi: 10.1016/j.mimet.2009.09.026

Bernfeld, P. (1955). Amylase α and β. *Methods Enzymol.* 1, 149–158. doi: 10.1016/0076-6879(55)01021-5

Betty, A. F., Daniel, L. S., and Weissfeld, A. S. (2007). *Bailey and Scott's Diagnostic Microbiology, 12th Edn*. Missouri, Mo: Mosby Elsevier.

Bhalla, A., Bansal, N., Kumar, S., Bischoff, K. M., and Sani, R. K. (2013). Improved lignocellulose conversion to biofuels with thermophilic bacteria and thermostable enzymes. *Bioresour. Technol.* 128, 751–759. doi: 10.1016/j.biortech.2012.10.145

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable, and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Bragger, J. M., Daniel, R. M., Coolbear, T., and Morgan, H. W. (1989). Enzymes from extreme environments. *Appl. Microbiol. Biotechnol.* 31, 556–561. doi: 10.1007/BF00270794

Brock, D. T. (2001). "The origins of research of the thermophiles," in *Thermophiles: Biodiversity, Ecology, and Evolution*, eds A.-L. Reysenbach et al. (New York, NY: Kluwer Academic/Plenum Publishers), 1–22.

Chan, C. S., Chan, K. G., Tay, Y. L., Chua, Y. H., and Goh, K. M. (2015). Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. *Front. Microbiol.* 6:177. doi: 10.3389/fmicb.2015.00177

Chandrasekharam, D. (2005). "Geothermal energy resources of India: past and the present," in *Proceedings World Geothermal Congress* (Antalya).

Craig, J., Absar, A., Bhat, G., Cadel, G., Hafiz, M., Hakhoo, N., et al. (2013). Hot springs and the geothermal energy potential of Jammu & Kashmir State, N.W. Himalaya, India. *Earth Sci. Rev.* 126, 156–177. doi: 10.1016/j.earscirev.2013.05.004

Demirjian, D. C., Moris-Varas, F., and Cassidy, C. S. (2001). Enzymes from extremophiles. *Curr. Opin. Chem. Biol.* 5, 144–151. doi: 10.1016/S1367-5931(00)00183-6

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Ghelani, A., Patel, R., Mangrola, A., and Dudhagara, P. (2015). Cultivation-independent comprehensive survey of bacterial diversity in Tulsi Shyam Hot Springs, India. *Genom Data* 4, 54–56. doi: 10.1016/j.gdata.2015.03.003

Gupta, G., Srivastava, S., Khare, S. K., and Prakash, V. (2014). Extremophiles: an overview of microorganism from extreme environment. *Int. J. Agric. Environ. Biotechnol.* 7, 371–380. doi: 10.5958/2230-732X.2014.00258.7

Haba, E., Bresco, O., Ferrer, C., Marques, A., Busquets, M., and Manresa, A. (2000). Isolation of lipase-secreting bacteria by deploying used frying oil as selective substrate. *Enzyme Microb. Technol.* 26, 40–44. doi: 10.1016/S0141-0229(99)00125-8

Horikoshi, K. (1999). Alkaliphiles: some applications of their products for biotechnology. *Am. Soc. Microbiol.* 63, 735–750. doi: 10.1128/MMBR.63.4.735-750.1999

Huang, Q., Dong, C. Z., Dong, R. M., Jiang, H., Wang, S., Wang, G., et al. (2011). Archaeal and bacterial diversity in hot springs on the Tibetan Plateau, China. *Extremophiles* 15, 549–563. doi: 10.1007/s00792-011-0386-z

Jiménez, D. J., Andreote, F. D., Chaves, D., Montaña, J. S., Osorio-Forero, C., Junca, H., et al. (2012). Structural and functional insights from the metagenome of an acidic hot spring microbial planktonic community in the Colombian Andes. *PLoS ONE* 7:e52069. doi: 10.1371/journal.pone.0052069

Jones, B. E., Grant, W., Duckworth, A. W., and Owenson, G. (1998). Microbial diversity of soda lakes. *Extremophiles* 2, 191–200. doi: 10.1007/s007920050060

Kotoky, R., and Pandey, P. (2020). Difference in the rhizosphere microbiome of *Melia azedarach* during removal of benzo(a)pyrene from cadmium co-contaminated soil. *Chemosphere* 258:127175. doi: 10.1016/j.chemosphere.2020.127175

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Langille, M. G., Zaneveld, J., Caporaso, J. G., Mcdonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

López-López, O., Cerdán, M. E., and González-Siso, M. I. (2013). Hot spring metagenomics. Life 3, 308–320. doi: 10.3390/life3020308

Lopez-Lopez, O., Knapik, K., Cerdán, M. E., and González-Siso, M. I. (2015). Metagenomics of an alkaline hot spring in Galicia (Spain): microbial diversity analysis and screening for novel lipolytic enzymes. *Front. Microbiol.* 6:1291. doi: 10.3389/fmicb.2015.01291

Mathur, E., Ortmann, A., Bateson, M., Geesey, G., and Frazier, M. (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS ONE* 5:e9773. doi: 10.1371/journal.pone.0009773

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

McLaughlin, B., and Weiss, J. B. (1996) Endothelial-cell-stimulating angiogenesis factor (ESAF) activates progelatinase A (72 kDa type IV collagenase), prostromelysin 1 and procollagenase and reactivates their complexes with tissue inhibitors of metalloproteinases. *Biochem. J.* 317, 739–745. doi: 10.1042/bj3170739

Menzel, P., Gudbergsdóttir, S. R., Rike, A. G., Lin, L., Zhang, Q., Contursi, P., et al. (2015). Comparative metagenomics of eight geographically remote terrestrial hot springs. *Microb. Ecol.* 70, 411–424. doi: 10.1007/s00248-015-0576-9

Miller, G. L. (1995). Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Anal. Chem.* 31, 426–428. doi: 10.1021/ac60147a030

Mirete, S., Morgante, V., and Gonzalez-Pastor, J. E. (2016). Functional metagenomics of extreme environments. *Curr. Opin. Biotechnol.* 38, 143–149. doi: 10.1016/j.copbio.2016.01.017

Mohammad, B. T., Al Daghistani, H. I., Jaouani, A., Abdel-Latif, S., and Kennes, C. (2017). Isolation and characterization of thermophilic bacteria from Jordanian hot springs: *Bacillus licheniformis* and *Thermomonas hydrothermalis* isolates as potential producers of thermostable enzymes. *Int. J. Microbiol.* 2017:6943952. doi: 10.1155/2017/6943952

Oladipo, I. C., Adebiyi, A. O., and Ayandele, A. A. (2008). Toxin production in food as influenced by pH, thermal treatment, and chemical preservatives. *Afr. J. Biotechnol.* 7, 1731–1739. doi: 10.5897/AJB08.188

Pedersen, D., and Miller, S. R. (2016). Photosynthetic temperature adaptation during niche diversification of the thermophilic cyanobacterium Synechococcus A/B clade. *ISME J.* 11, 1053–1057. doi: 10.1038/ismej.2016.173

Power, J. F., Carere, C. R., Lee, C. K., Wakerley, G. L. J., Evans, D. W., Button, M., et al. (2018). Microbial biogeography of 925 geothermal springs in New Zealand. *Nat. Commun.* 9:2876. doi: 10.1038/s41467-018-05020-y

Purcell, D., Sompong, U., Yim, L. C., Barraclough, T. G., Peerapornpisal, Y., and Pointing, S. B. (2007). The effects of temperature, pH, and sulphide on the community structure of hyperthermophilic streamers in hot springs of northern Thailand. *FEMS Microbiol. Ecol.* 60, 456–466. doi: 10.1111/j.1574-6941.2007.00302.x

Roy, C., Rameez, M. J., Haldar, P. K., Peketi, A., Mondal, N., Bakshi, U., et al. (2020). Microbiome and ecology of a hot spring-microbialite system on the Trans-Himalayan Plateau. *Sci Rep.* 10:5917. doi: 10.1038/s41598-020-62797-z

Sahay, H., Yadav, A. N., Singh, A. K., Singh, S., Kaushik, R., Saxena, A. K. (2017). Hot springs of Indian Himalayas: potential sources of microbial diversity and thermostable hydrolytic enzymes. *3 Biotech.* 7:118. doi: 10.1007/s13205-017-0762-1

Saxena, R., Dhakan, D. B., Mittal, P., Waiker, P., Chowdhury, A., Ghatak, A., et al. (2017). Metagenomic analysis of hot springs in Central India reveals hydrocarbon degrading thermophiles and pathways essential for survival in extreme environments. *Front. Microbiol.* 7:2123. doi: 10.3389/fmicb.2016.02123

Schuler, C. G., Havig, J. R., and Hamilton, T. L. (2017). Hot spring microbial community composition, morphology, and carbon fixation: implications for interpreting the ancient rock record. *Front. Earth Sci.* 5:97. doi: 10.3389/feart.2017.00097

Shanker, R. (1986). Scope of utilisation of geothermal energy for area development in backward, Hilly, and Tribal regions of India. *Indian Miner.* 40, 49–61.

Sharaf, E. F., El-Sayed, W. S., and Abosaif, R. M. (2014). Lecithinase-producing bacteria in commercial and home-made foods: evaluation of toxic properties and identification of potent producers. *J. Taibah Univ. Sci.* 8, 207–215. doi: 10.1016/j.jtusci.2014.03.006

Sikdar, A., Raziuddin, A., and Gupta, K. K. (2015). Isolation and characterisation of thermophilic bacteria of a hot water spring source, Balbal. *Int. J. Adv. Res. Biol. Sci.* 2, 106–111.

Singh, G., Bhalla, A., Kaur, P., Capalash, N., and Sharma, P. (2011). Laccase from prokaryotes: a new source for an old enzyme. *Rev. Environ. Sci. Biotechnol.* 10, 309–326. doi: 10.1007/s11157-011-9257-4

Skirnisdottir, S., Hreggvidsson, G. O., Hjorleifsdottir, S., Marteinsson, V. T., Petursdottir, S. K., Holst, O., et al. (2000). Influence of sulfide and temperature on species composition and community structure of hot spring microbial mats. *Appl. Environ. Microbiol.* 66, 2835–2841. doi: 10.1128/AEM.66.7.2835-2841.2000

Sorokin, D. Y., and Kuenen, J. G. (2005). Alkaliphilic chemolithotrophs from soda lakes. *FEMS Microbiol. Ecol.* 52, 287–295. doi: 10.1016/j.femsec.2005.02.012

Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43. doi: 10.1038/nature02340

Uribe-Lorio, L., Brenes-Guillen, L., Hernandez-Ascencio, W., Mora-Amador, R., Gonzalez, G., Ramirez-Umana, C. J., et al. (2019). The influence of temperature and pH on bacterial community composition of microbial mats in hot springs from Costa Rica. *Microbiologyopen* 8, 1–26. doi: 10.1002/mbo3.893

Valverde, A., Tuffin, M., and Cowan, D. (2012). Biogeography of bacterial communities in hot springs: a focus on the actinobacteria. *Extremophiles* 16, 669–679. doi: 10.1007/s00792-012-0465-9

Vieille, C., and Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* 65, 1–43. doi: 10.1128/MMBR.65.1.1-43.2001

Wang, S., Hou, W., Dong, H., Jiang, H., Huang, L., Wu, G., et al. (2013). Control of temperature on microbial community structure in hot springs of the Tibetan Plateau. *PLoS ONE* 8:e62901. doi: 10.1371/journal.pone.0062901

Wemheuer, B., Taube, R., Akyol, P., Wemheuer, F., and Daniel, R. (2013). Microbial diversity and biochemical potential encoded by thermal spring metagenomes derived from the Kamchatka peninsula. *Archaea* 2013:136714. doi: 10.1155/2013/136714

Zhang, Y., Wu, G., Jiang, H., Yang, J., She, W., Khan, I., et al. (2018). Abundant and rare microbial biospheres respond differently to environmental and spatial factors in Tibetan hot springs. *Front. Microbiol.* 9:2096. doi: 10.3389/fmicb.2018.02096

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 March 2021
doi: 10.3389/fgene.2021.586569

Genomic Variations in SARS-CoV-2 Genomes From Gujarat: Underlying Role of Variants in Disease Epidemiology

*Madhvi Joshi[1], Apurvasinh Puvar[1], Dinesh Kumar[1], Afzal Ansari[1], Maharshi Pandya[1], Janvi Raval[1], Zarna Patel[1], Pinal Trivedi[1], Monika Gandhi[1], Labdhi Pandya[1], Komal Patel[1], Nitin Savaliya[1], Snehal Bagatharia[2], Sachin Kumar[3] and Chaitanya Joshi[1]**

[1] Gujarat Biotechnology Research Centre (GBRC), Department of Science & Technology (DST), Gandhinagar, India, [2] Gujarat State Biotechnology Mission, Gandhinagar, India, [3] Indian Institute of Technology Guwahati, Guwahati, India

Humanity has seen numerous pandemics during its course of evolution. The list includes several incidents from the past, such as measles, Ebola, severe acute respiratory syndrome (SARS), and Middle East respiratory syndrome (MERS), etc. The latest edition to this is coronavirus disease 2019 (COVID-19), caused by the novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of August 18, 2020, COVID-19 has affected over 21 million people from 180 + countries with 0.7 million deaths across the globe. Genomic technologies have enabled us to understand the genomic constitution of pathogens, their virulence, evolution, and rate of mutation, etc. To date, more than 83,000 viral genomes have been deposited in public repositories, such as GISAID and NCBI. While we are writing this, India is the third most affected country by COVID-19, with 2.7 million cases and > 53,000 deaths. Gujarat is the 11th highest affected state with a 3.48% death rate compared to the national average of 1.91%. In this study, a total of 502 SARS-CoV-2 genomes from Gujarat were sequenced and analyzed to understand its phylogenetic distribution and variants against global and national sequences. Further variants were analyzed from diseased and recovered patients from Gujarat and the world to understand its role in pathogenesis. Among the missense mutations present in the Gujarat SARS-CoV-2 genomes, C28854T (Ser194Leu) had an allele frequency of 47.62 and 7.25% in deceased patients from the Gujarat and global datasets, respectively. In contrast, the allele frequency of 35.16 and 3.20% was observed in recovered patients from the Gujarat and global datasets, respectively. It is a deleterious mutation present in the nucleocapsid (N) gene and is significantly associated with mortality in Gujarat patients with a *p*-value of 0.067 and in the global dataset with a *p*-value of 0.000924. The other deleterious variant identified in deceased patients from Gujarat (*p*-value of 0.355) and the world (*p*-value of 2.43E-06) is G25563T, which is located in Orf3a and plays a potential role in viral pathogenesis. SARS-CoV-2 genomes from Gujarat are forming distinct clusters under the GH clade of GISAID. This study will shed light on the viral haplotype in SARS-CoV-2 samples from Gujarat, India.

Keywords: genomic surveillance, mutation analysis, SARS-CoV-2 (2019-nCoV), COVID-19, viral epidemiology, haplotyping

OPEN ACCESS

Edited by:
Dhaval K. Acharya, B N Patel Institute of Paramedical, India

Reviewed by:
Jeremy W. Prokop, HudsonAlpha Institute for Biotechnology, United States
Saumya Patel, Gujarat University, India

****Correspondence:***
Chaitanya Joshi
dir-gbrc@gujarat.gov.in; director.gbrc@gmail.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *23 July 2020*
Accepted: *15 February 2021*
Published: *19 March 2021*

Citation:
Joshi M, Puvar A, Kumar D, Ansari A, Pandya M, Raval J, Patel Z, Trivedi P, Gandhi M, Pandya L, Patel K, Savaliya N, Bagatharia S, Kumar S and Joshi C (2021) Genomic Variations in SARS-CoV-2 Genomes From Gujarat: Underlying Role of Variants in Disease Epidemiology. Front. Genet. 12:586569. doi: 10.3389/fgene.2021.586569

INTRODUCTION

As per the recent situation report-209 released by the World Health Organisation (WHO), as accessed on August 18, 2020, the total confirmed positive cases of COVID-19 across the globe are 21,294,845 resulting in 761,779 deaths. In many countries, such as China, Spain, Australia, Japan, South Korea, and the United States, the second wave of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections has started (Evenett and Winters, 2020; Leung et al., 2020; Strzelecki, 2020; Xu and Li, 2020). India is the third most affected country by coronavirus disease 2019 (COVID-19) after the United States and Brazil, with 2,771,958 cases and 53,046 deaths, respectively. Gujarat is located in the western part of India. It is the 11th highest affected state in India, with 80,942 cases and 2,820 deaths as per the https://www.covid19india.org/. However, the death rate in the state of Gujarat is 3.48% with a recovery rate of 78.83%, which is 5% higher than the existing recovery rate in India. Therefore, understanding the pathogen evolution and virulence through genome sequencing will be key to understanding its diversity, variation, and its effect on pathogenesis and disease severity. Global repositories, such as the GISAID and NCBI databases, are flooded with SARS-CoV-2 genomes with an average of 381 genomes per day being added from across the globe. SARS-CoV-2 genome size is 29–30.6 kb. The genome includes 10 genes that encode 4 structural and 16 non-structural proteins (NSPs). Structural proteins are encoded by the four structural genes, including spike (S), envelope (E), membrane (M), and nucleocapsid (N) genes. The ORF1ab is the largest gene in SARS-CoV-2, which encodes the pp1ab protein and 15 NSPs. The ORF1a gene encodes for pp1a protein, which also contains 10 NSPs (Du et al., 2009; Shereen et al., 2020).

In the present study, the whole genome of 502 SARS-CoV-2 from Gujarat has been sequenced and analyzed against 2,121 SARS-CoV-2 genomes across the globe with known patient status. The overall dataset comprises 361 confirmed positive COVID-19 patients, which included 122 female and 239 male patients from Gujarat, India. Furthermore, a total of 502 viral genomes were sequenced from 361 samples based on the dominant and recessive allelic frequencies. These genomes were studied against a total of 79,518 complete viral genome sequences as accessed on August 18, 2020 to characterize their clades and variant distribution. Further statistical tools were applied to understand the differences in the variants with respect to disease epidemiology. In the absence of clinically approved drugs and other possible therapies in treating COVID-19, tracking pathogen evolution through whole genome sequencing is instrumental in understanding the progression of the pandemic locally as well as globally. This will further help in devising better strategies for vaccine development, identifying potential drug targets, and understanding host–pathogen interactions.

MATERIALS AND METHODS

Sample Collection and Processing

Nasopharyngeal and oropharyngeal swabs from a total of 361 individuals who tested positive for COVID-19 from 46 locations representing 20 districts of Gujarat were collected after obtaining informed consent and appropriate ethics approval. The numbers of samples from these locations were selected on the basis of disease spread and incidence rate in Gujarat. The details of samples collected from each location are shown in **Supplementary Table 1**. Samples were transported as per standard operating procedures as prescribed by the World Health Organisation (WHO) and Indian Council of Medical Research (ICMR, New Delhi; SoP No. ICMR-NIV/2019-nCoV/Specimens_01) to a research laboratory and further stored at −20°C till processed.

Whole Genome Sequencing of SARS-CoV-2

Total genomic RNA from the samples was isolated using the QIAamp Viral RNA Mini Kit (Cat. No. 52904; Qiagen, Germany) following the prescribed biosafety procedure. cDNA from the extracted RNA was made using the SuperScript™ III Reverse Transcriptase first strand kit (Cat. No. 18080093; Thermo Fisher Scientific, United States) as per the procedures prescribed. SARS-CoV-2 genome was amplified by using the Ion AmpliSeq SARS-CoV-2 Research Panel (Thermo Fisher Scientific, United States) that consists of two pools with amplicons ranging from 125 to 275 bp in length and covering >99% of the SARS-CoV-2 genome, including all serotypes. Amplicon libraries were prepared using the Ion AmpliSeq™ Library Kit Plus (Cat. No. A35907; Thermo Fisher Scientific, United States). These libraries were quantified using the Ion Library TaqMan™ Quantitation Kit (Cat. No. 4468802; Thermo Fisher Scientific, United States). The quality of the library was checked using the DNA high sensitivity assay kit on Bio-analyser 2100 (Agilent Technologies, United States) and was sequenced on the Ion S5 Plus sequencing platform using a 530 chip.

Raw Data Quality Assessment and Filtering

The quality of data was assessed using the FASTQC v. 0.11.5 (Andrews, 2010) toolkit. All raw data sequences were processed using the PRINSEQ-lite v.0.20.4 (Schmieder and Edwards, 2011) program for filtering the data. All sequences were trimmed from the right to where the average quality of 5 bp window was lower than QV25, 5 bp from the left end was trimmed, and sequences with length lower than 50 bp and sequences with average quality QV25 were removed.

Genome Assembly, Variant Calling, and Global Dataset

Quality filtered data were assembled using reference-based mapping with CLC Genomics Workbench 12. Mapping was done using stringent parameters with a length fraction of 0.99 and a similarity fraction of 0.9. Mapping tracks were used to call and annotate variants. Variants were called using Ion Torrent variant caller with a minimum allele frequency of 30% with a minimum coverage of 10 reads considered. For comparative analysis with the global dataset, 79,518 complete viral genomes and 1,821 viral genome isolates from India were downloaded

from the GISAID flu server[1]. Considering haplotypes (a, b) based on allelic frequency, a total of 502 genomes were sequenced from a total of 361 patients as mentioned in **Supplementary Table 2**.

Phylogenetic Analysis

A total of 502 SARS-CoV-2 whole genomes were sequenced and analyzed for their phylogenetic distribution at different locations from Gujarat. The reference genome, Wuhan/Hu-1/2019 (EPI_ISL_402125), sampled on December 31, 2019, from Wuhan, China was downloaded from the GISAID flu server. Additionally, three viral genomes from the seafood market from Wuhan, China were included in the phylogenetic analysis (EPI_ISL_406798, EPI_ISL_402124, and EPI_ISL_406801). The multiple sequence alignment was performed using MAFFT (Katoh and Standley, 2013) implemented *via* a phylodynamic analysis pipeline provided by Augur[2]. The subsequent alignment output files were checked, visualized, and verified using PhyloSuite (Zhang et al., 2020). Afterward, the maximum likelihood phylogenetic tree was built using the Augur tree implementation pipeline with the IQ-TREE 2 (Minh et al., 2020) with default parameters. The selected metadata information plotted in the time-resolved phylogenetic tree was constructed using TreeTime (Sagulenko et al., 2018) and annotated and visualized in the FigTree (Rambaut, 2018).

Statistical Analysis

The non-parametric chi-square test of significance was used to check the difference of variables, such as the effect on age, gender, and mutations on mortality for the Gujarat, India, and global datasets for the deceased versus recovered patients.

RESULTS

Samples were collected based on COVID-19 incidence rate across Gujarat from 16 different originating laboratories representing 46 different geographical locations from 20 districts of Gujarat, India as mentioned in **Supplementary Table 1**. The geographical distributions of the top three locations of viral isolates are represented by Ahmedabad (n = 172), Vadodara (n = 92), and Surat (n = 86), respectively. A total of 502 viral genomes from 361 patients have been sequenced in the study from which 122 were from females, whereas 239 were from males. These patients were from 1 to 86 years old group with an average age of 47.91 years. Most of the COVID-19 positive patients had symptoms of fever, diarrhea, cough, and breathing problems, whereas some of them had comorbid conditions, such as hypertension, diabetes, etc. The final outcomes of these patients were classified as deceased, recovered, hospitalized, or unknown status for further data analysis based on the available metadata. These details are presented in **Supplementary Table 2**. Chi-square test was performed to test the effect of gender and age group for the Gujarat and global datasets. The female patients (at p-value of 2.7E-08) in the Gujarat dataset were observed to be at a significantly higher death rate than those in the global dataset in deceased and recovered patients. The genomic dataset was further divided into different age groups of up to 40, 41–60, and over 60 years. The results indicated a significantly higher mortality rate at the age groups of 41–60 (at p-value of 0.03783) and over 60 years in the Gujarat dataset (at p-value of 0.2084) than at the age groups in the global dataset. Life expectancy in India is 68.7 years as per the National Health Profile 2019 report released by the Central Bureau of Health Intelligence (CBHI), Ministry of Health and Family Welfare, Government of India. The mutation frequency profile of the Gujarat genome with the mutation spectrum is highlighted in **Figure 1** including synonymous and missense mutations.

Genome Sequencing and Haplotyping

Out of 361 patients, 141 had mixed infections. Mixed infections were judged by the frequency of heterozygous mutations. The heterozygous mutation was considered only if it was supported by forward and reverse reads of an amplicon. Genomes were observed for heterozygous allele frequencies and were manually divided into two genomes. As a result, from 141 patients, a total of 282 viral genomes were classified as two different haplotypes and annotated with suffixes "a" and "b." All major alleles having read frequency ranging from 60 to 80% were included in the "a" haplotype, whereas minor alleles having read frequency ranging from 20 to 40% were included in the "b" haplotype as provided in **Supplementary Table 2**.

Phylogenetic Analysis

Phylogenetic analysis of 502 genomes was done as per the definitions of the PANGOLIN lineage and GISAID clades. The overall lineages distribution highlighted the dominant occurrence of B.1.36 (n = 214), B.1 (n = 182), A (n = 18), B.6 (n = 12), B.1.1 (n = 9), and B (n = 4), whereas clade distribution highlighted the dominant prevalence of GH (n = 278), G (n = 180), O (n = 18), S (n = 18), GR (n = 7), and L (n = 1) as mentioned in **Supplementary Table 3**. While none of the genomes from Gujarat belonged to clade V, in the global perspective, the distribution of the GISAID clades as of 18th August 2020 from viral genome sequences indicates the dominance of GR clade (32.14%), G clade (23.72%), GH clade (22.66%), S clade (6.73%), L clade (5.13%), V clade (6.17%), and O clade (3.45%). The maximum likelihood time-resolved phylogeny tree in **Figure 2** was constructed using the TreeTime pipeline and Augur bioinformatics pipeline and annotated and visualized in the FigTree (Hadfield et al., 2018; Rambaut, 2018; Mercatelli and Giorgi, 2020). Similarly, genomes classified into GISAID clades across the globe and Gujarat are highlighted in **Figure 3**.

Comparative Analysis of Mutation Profile in SARS-CoV-2 Genomes

To understand the significance of the mutations in the SARS-CoV-2 genome isolates from the Gujarat, India, and global dataset, we have analyzed and compared the mutation

[1]https://www.gisaid.org/

[2]https://github.com/nextstrain/augur

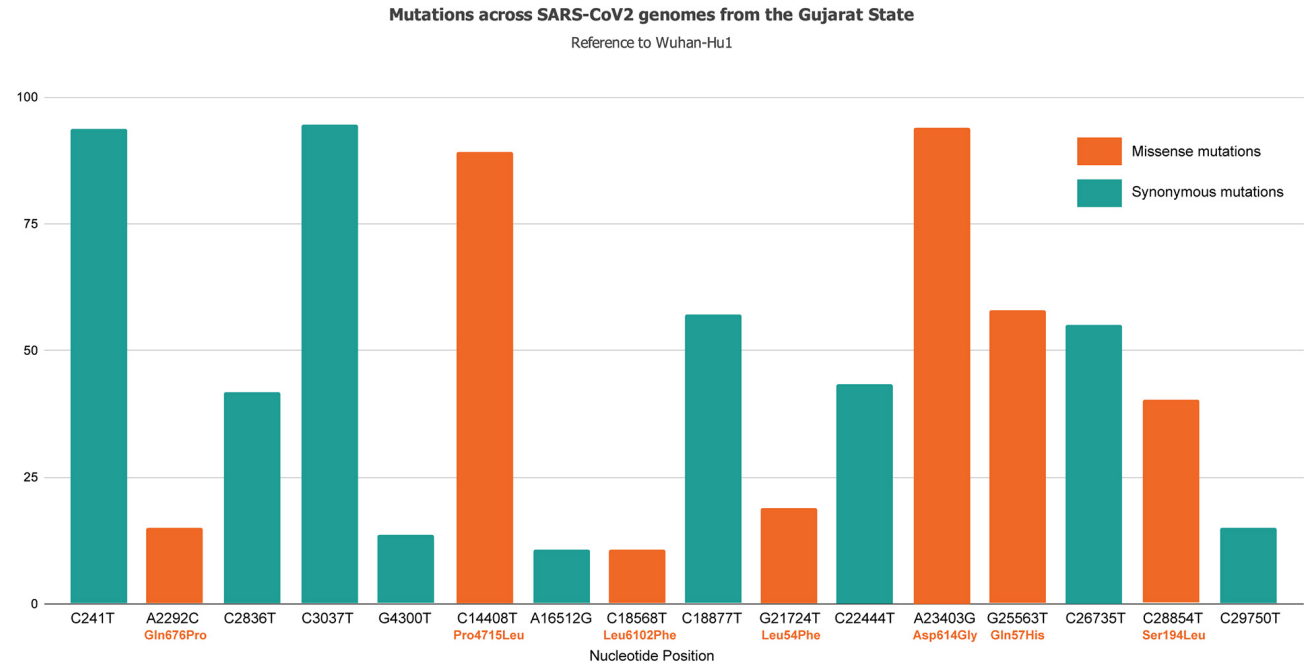

FIGURE 1 | Mutation spectrum profile of 502 SARS-CoV-2 genomes from 46 locations representing 20 districts of Gujarat, India including synonymous and missense mutations. The top mutations included C241T, C3037T, C14408T/Pro314Leu, C18877T, A23403G/Asp614Gly, G25563T/Gln57His, and C26735T with frequency >55%.

profile of the 502 viral isolates from Gujarat along with the global dataset of 79,518 viral genomes and 1,821 viral genomes from India obtained from the GISAID server. A total of 27,455 mutations were observed in the global viral genome sequences (n = 79,518) of SARS-CoV-2 from GISAID wherein 3,478 mutations were observed from viral genomes from the Indian isolates (n = 1,821), whereas 752 mutations were observed in genomes sequenced from the Gujarat isolates (n = 502). Out of these mutations, 111 mutations were novel to viral isolates from the Gujarat genomes, and 1,164 were novel to the Indian genomes. The bar chart displaying the comparative mutation analysis is represented as **Figure 4**, with a frequency of >5% from the global, Indian, and Gujarat viral genomes including missense and synonymous mutations.

A Venn diagram represents the overall mutations shared between viral genome sequences analyzed from the global, Indian, and Gujarat isolates as given in **Figure 5**. A comparison of the mutation profile analysis with p-value significance, Sorting Intolerant from Tolerant (SIFT) score functional effect, frequency >5%, and absolute count of the number of genomes with prevalence is represented in **Table 1**. Further, frequencies of all the mutations were calculated by subtracting variants of the Gujarat genomes from the Indian and global genomes with statistical significance.

Mutations (C241T, C3037T, A23403G, and C14408T) were dominant with frequency (>60%) in all the genomes (Gujarat, India, and global), whereas mutations (G11083T, C13730T, C28311T, C6312A, C313T, C5700A, G29868A, and C23929T) dominated (>19%) in the Indian genomes compared with the Gujarat and global genomes. The multi-nucleotide variant (MNV) GGG28881AAC is dominant in Indian (35.25%) and global genomes (32.72%), but in the context of Gujarat, it is present with a frequency of 2.19%. The mutations G25563T, C26735T, and C18877T (>55%), followed by C2836T, C22444T, and C28854T (>40%), followed by G21724T, C29750T, C18568T, G4300T, and A2292C (>13%) in viral genomes were sequenced from Gujarat. The detailed mutation frequency profile is provided as **Supplementary Table 4**. With reference to viral isolates from India, GGG28881AAC, G11083T, C28311T, C6312A, C23929T, and C13730T were found to be occurring at greater than 19% frequencies (p-value <0.001). Mutations G11083T and C6312A lie in the region of Orf1a encoding Nsp6, whereas mutation GGG28881AAC is present in the N gene. Further, deceased versus recovered patient mutation profile analysis of the known patient's status dataset from Gujarat and global is represented in **Figure 6** and **Supplementary Tables 5, 6**. Similarly, comparison of missense mutation profile of deceased versus recovered patients with genome count, frequency >5%, and p-value for the global dataset is represented in **Table 2** and for the Gujarat dataset in **Table 3**; additionally, metadata for deceased and recovered patients is provided as **Supplementary Tables 7, 8**. The statistical significance association of gender and age of the deceased and recovered patients from the Gujarat and global dataset patients in both datasets was considered for analysis. Similarly, for age group 41–60 years, it highlighted the higher observation of death rate in patients with known status as given in **Table 4**.

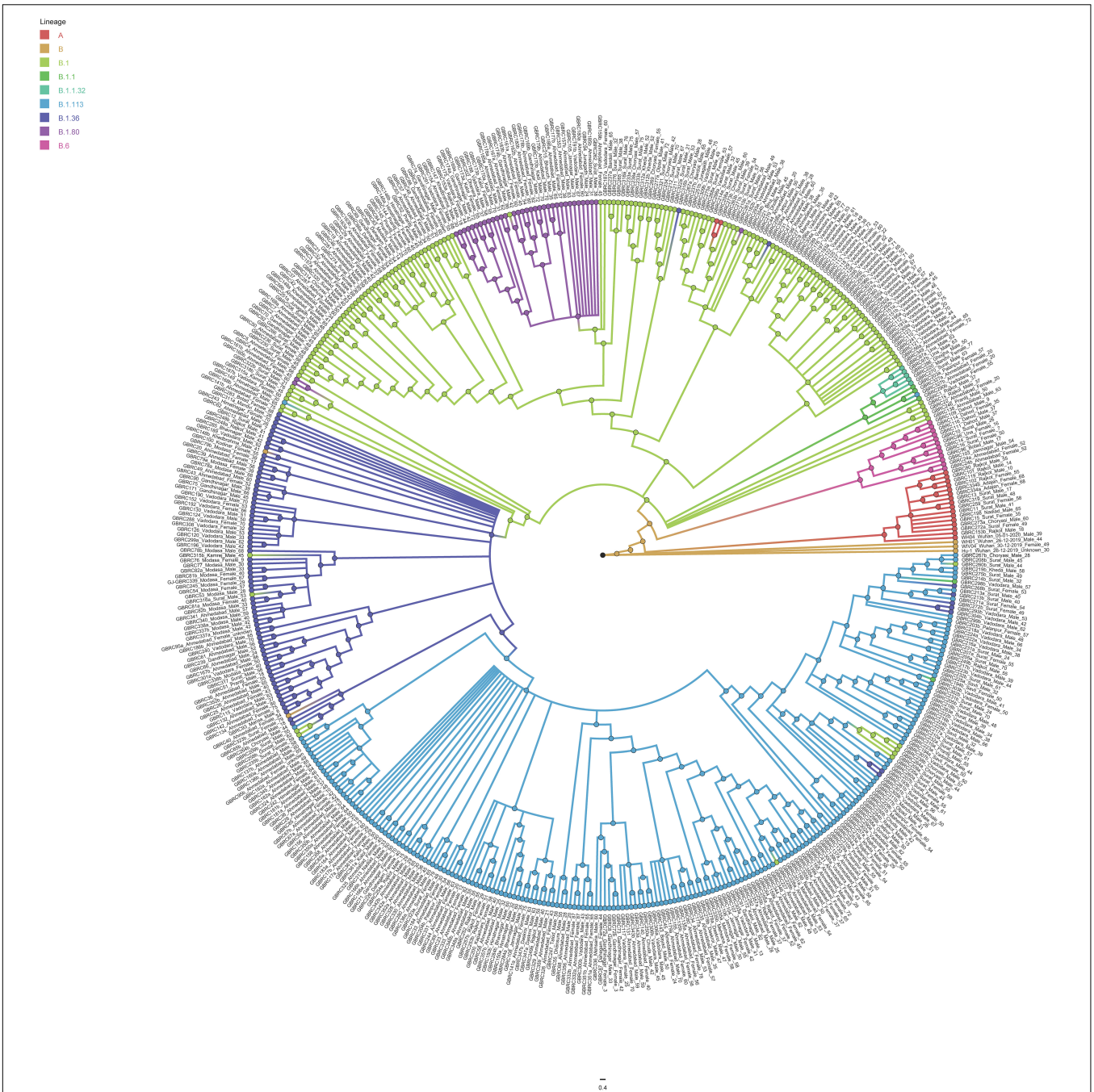

FIGURE 2 | Phylogenetic distribution of lineage from 502 SARS-CoV-2 viral genomes from Gujarat, India with reference to the Wuhan/Hu-1/2019 (EPI_ISL_402125). Maximum likelihood phylogenetic tree was built using the Augur tree implementation pipeline with the IQ-TREE 2 with default parameters. The selected metadata information plotted in the time-resolved phylogenetic tree was constructed using TreeTime and visualized in the FigTree.

DISCUSSION

India is a densely populated country and needs to tackle the challenges of the COVID-19 pandemic through management strategies and the stringent implementation of policies. The genome sequencing efforts have been enormously useful in understanding the pathogenic and adaptive behavior of viruses in the Indian population. The epidemiological approach-based method in a resource-poor setting, such as Telangana and Andhra Pradesh states, revealed that the case-fatality ratios spanned 0.05% at ages 5–17 years to 16.6% at ages ≥85 years (Laxminarayan et al., 2020). Similarly, immune response, food habits, and gut–microbiome dynamics might also play key roles in the SARS-CoV-2 viral outbreak that should further

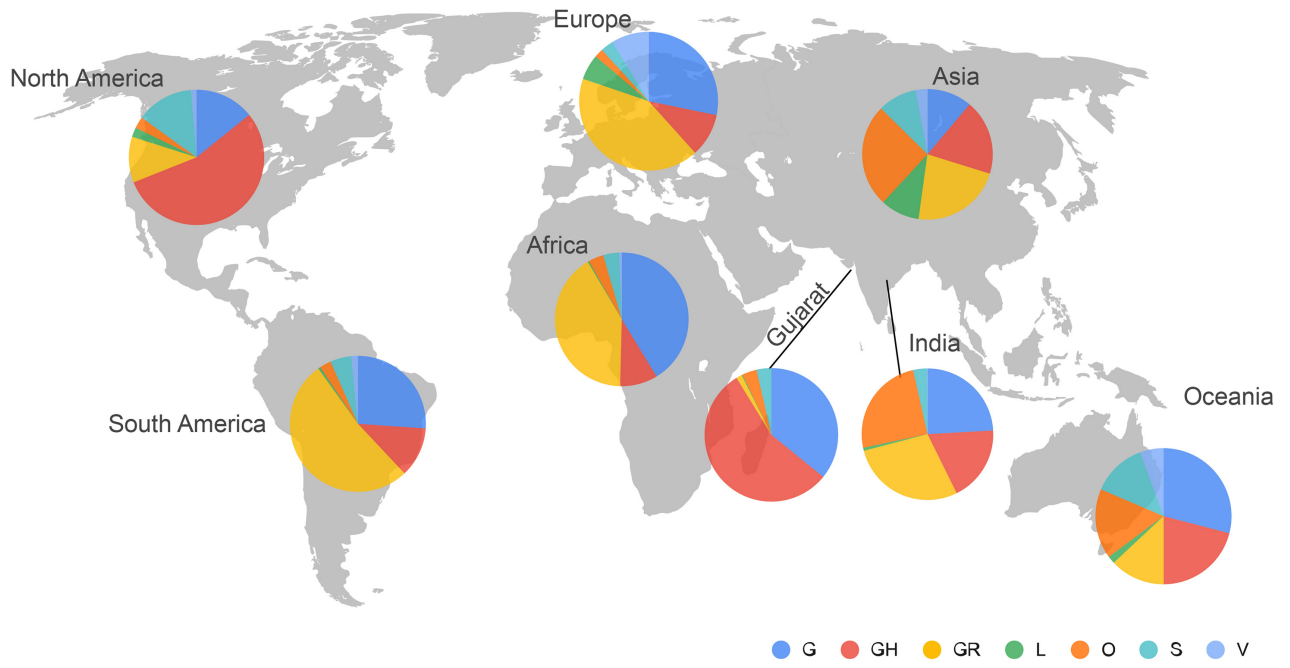

FIGURE 3 | Distribution of the GISAID clades of SARS-CoV-2 genomes from global and Gujarat datasets as of 18th August 2020. The majority of the viral genomes from Gujarat are falling under GH (*n* = 278) and G (*n* = 180) clades.

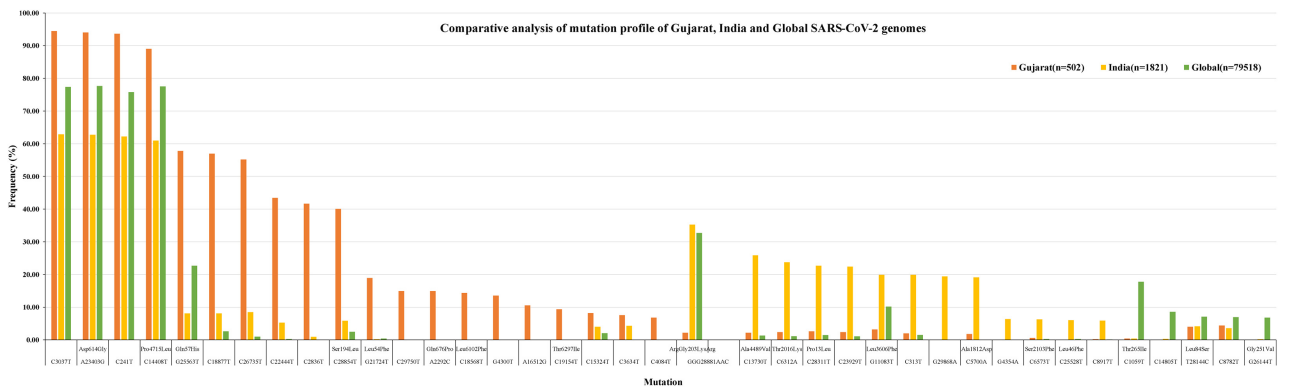

FIGURE 4 | Synonymous and missense mutation profiles of the SARS-CoV-2 viral genomes, Gujarat (*n* = 502), India (*n* = 1,821), and global (*n* = 79,518). Only mutations with frequency >5% are plotted.

help in identifying host-related responses and better control strategies (Bajaj and Purohit, 2020; Shastri et al., 2020). Furthermore, to understand virus pathogenesis dynamics in the populations of Gujarat, genome sequencing of the SARS-COV-2 clinical positive samples was carried out. SARS-CoV-2 viral genome analysis from Gujarat highlights the distinct genomic attributes, geographical distribution, age composition, and gender classification. These features also highlight unique genomic patterns in terms of synonymous and missense variants associated with the prevalence of dominant clades and lineages with distinct geographical locations in Gujarat. Our research study highlights the most comprehensive genomic resources available so far from Gujarat, India. Therefore, identifying variants specific to the deceased and recovered patients would certainly aid in better treatment and COVID-19 containment strategy. The fatality rate compared with different geographical locations may point toward the higher virulence profile of certain viral strains with lethal genetic mutations, but this remains to be clinically unestablished. Perhaps the onset of clinical features in symptomatic patients helps in prioritizing the diagnosis and testing strategy.

The first case report of complete genome sequence information from India is from a patient in Kerala with a direct travel history to Wuhan, China. Similarly, other isolates from India cluster with Iran, Italy, Spain, England, United States, and Belgium, and probably similar isolates

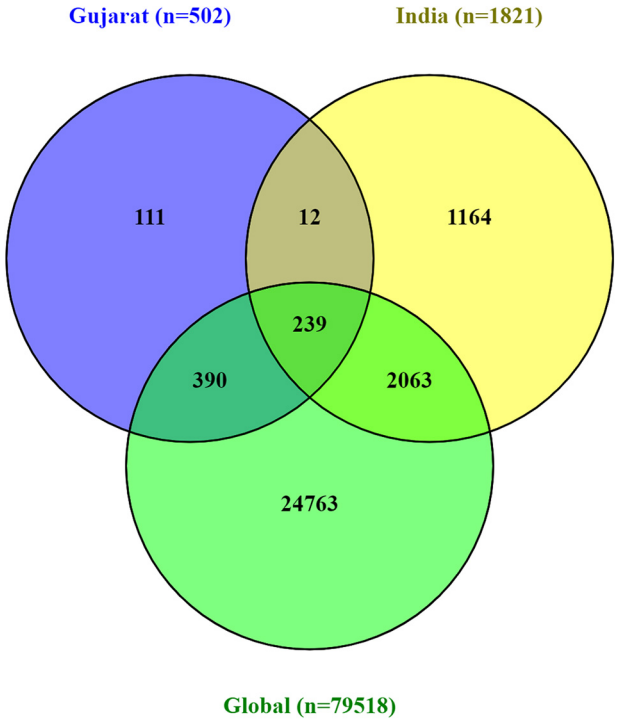

FIGURE 5 | Venn diagram representing the mutually common and exclusive synonymous and missense mutations among SARS-CoV-2 viral genomes, Gujarat (n = 502), India (n = 1,821), and global (n = 79,518).

are transmitting in India and may have variable mutation profile (Mondal et al., 2020; Potdar et al., 2020; Yadav et al., 2020). The dominance of a particular lineage or clade at a particular location merely does not establish the biological function of the virus type isolate in terms of higher death rate but the epidemiological factors, such as clinically diagnosed co-morbidity, age, gender, or asymptomatic transmission, that are the most likely influencing factor in transmission. Sampling biases could certainly influence the prediction models, but it would narrow down to particular types of isolates and unique mutations that can further be experimentally validated to establish their clinical significance.

The geographical distribution of the viral isolates is denoted in the phylogeny with the maximum SARS-CoV-2 positive samples sequenced from Ahmedabad (n = 172), followed by Vadodara (n = 92), Surat (n = 86), and Gandhinagar (n = 30). The distribution of dominant lineages in Ahmedabad is steered by occurrences of B.1.36 (n = 75), B.1 (n = 55), and B.6 (n = 2). The concept of lineages, clades, haplotypes, or genotypes is slightly perplexing and overlapping in terms of definitions with respect to different repositories and analytics. Therefore, it is most important to define mutations in the isolates that determine their unique position in phylogeny in terms of geographical distribution, age, gender, and locations of the genotypes, etc. Phylogenetic distribution of the viral genomes across different geographical locations along with metadata information should help in the evaluation of the posterior distribution, virulence, divergence times, and evolutionary rates in viral populations (Drummond and Rambaut, 2007). The recurrent mutations occurring independently multiple times in the viral genomes are hallmarks of convergent evolution in viral genomes with significance in host adaptability, spread, and transmission, even though contested in terms of mechanisms driving the pathogenicity and virulence across different hosts and specifically to human populations across different geographical locations (Grifoni et al., 2020; van Dorp et al., 2020).

Incidence of Mutations in Deceased and Recovered Patients

In the context of the globally prevalent mutations across different geographical locations, we have analyzed viral genome isolates with the most frequent mutations present in the patients from those who have suffered casualties. The higher death rate, especially in Ahmedabad, India, became a cause of serious concern and remains elusive to be identified with enough scientific evidence. We have identified differentially dominant and statistically significant mutations prevalent in the viral genome isolates in Gujarat, India.

The dominant mutations in the deceased patients represented by the change in A23403G were observed at a frequency of 98.41% in the Gujarat genomes (p-value of 0.1640) and 74.28% in the global genomes with known patient status (p-value of 0.5223). These missense mutations are found to be observed in the spike protein of the SARS-CoV-2 genome. The well-known function of the viral spike protein is in mediating the infection by interacting with the angiotensin-converting enzyme 2 (ACE2) receptor (Li et al., 2005; Chu et al., 2020; Guan et al., 2020; Guo et al., 2020) of the human host species. Another mutation, C14408T with a frequency of 96.83%, is present in the Orf1b gene encoding RNA-directed RNA polymerase (RDRP) non-structural protein (nsp12) with a p-value of 0.1440 in deceased versus recovered patients from Gujarat, while also being observed to be statistically significant in the global dataset with a p-value of 8.28E-05 with a frequency of 88.77%. The comparative analysis of the deceased patients (n = 63) and recovered patients (n = 256) in Gujarat as highlighted in **Figure 6** is represented by a Venn diagram. In contrast, the functional role of the RDRP enzyme activity is necessary for the viral genome replication and transcription of most RNA viruses (Imbert et al., 2006; Velazquez-Salinas et al., 2020). The MNV GGG28881AAC that is a missense mutation with a change in ArgGly203LysArg in the N gene is a deleterious mutation and is dominant in global viral genomes with a frequency in deceased (39.45%) and recovered patients (31.38%).

The exclusive dominant mutations present in the population of Gujarat were G25563T and C28854T in the Orf3a and N genes, respectively. The Orf3a gene encodes a protein involved in the regulation of inflammation, antiviral responses, and apoptosis. Mutation in these regions alters the functional profile of the nuclear factor-kβ (NF-kβ) activation and nucleotide-binding domain leucine-rich repeat containing (NLRP3) inflammasome. One of the main features of Orf3a protein is having the presence of a cysteine-rich domain, which participates in the

TABLE 1 | The overall comparison of missense 478 and synonymous mutation frequency profiles of Gujarat-502, India-1821, and Global-79518 datasets.

			Genome count			Frequency					
Gene	**NT position**	**AA position**	**Gujarat (*n* = 502)**	**India (*n* = 1,821)**	**Global (*n* = 79,518)**	**Gujarat**	**India**	**Global**	**SIFT score**	**Functional effect**	***p*-Value**
5' UTR	C241T		470	1,133	60,265	93.63	62.22	75.79	#N/A	#N/A	1.23505E-58
ORF1ab	C313T		10	362	1,178	1.99	19.88	1.48	0.84	Benign/tolerated	0
	C1059T	Thr265Ile	2	7	14,114	0.40	0.38	17.75	0.03	Deleterious	3.3988E-104
	A2292C	Gln676Pro	75	0	0	14.94	0.00	0.00	0.05	Deleterious	0
	C2836T		209	17	21	41.63	0.93	0.03	0.17	Benign/tolerated	0
	C3037T		474	1,145	61,503	94.42	62.88	77.34	0.66	Benign/tolerated	3.45605E-65
	C3634T		38	78	26	7.57	4.28	0.03	0.40	Benign/tolerated	0
	C4084T		34	1	35	6.77	0.05	0.04	0.72	Benign/tolerated	0
	G4300T		68	1	41	13.55	0.05	0.05	0.84	Benign/tolerated	0
	G4354A		0	116	0	0.00	6.37	0.00	1.00	Benign/tolerated	0
	C5700A	Ala1812Asp	9	348	8	1.79	19.11	0.01	0.38	Benign/tolerated	0
	C6312A	Thr2016Lys	12	432	882	2.39	23.72	1.11	0.03	Deleterious	0
	C6573T	Ser2103Phe	3	114	206	0.60	6.26	0.26	0.36	Benign/tolerated	0
	C8782T		22	65	5,526	4.38	3.57	6.95	0.67	Benign/tolerated	1.08234E-08
	C8917T		1	107	90	0.20	5.88	0.11	1.00	Benign/tolerated	0
	G11083T	Leu3606Phe	16	362	8,060	3.19	19.88	10.14	0.01	Deleterious	1.98676E-46
	C13730T	Ala4489Val	11	471	1,034	2.19	25.86	1.30	0.00	Deleterious	0
	C14408T	Pro4715Leu	447	1,110	61,641	89.04	60.96	77.52	0.31	Benign/tolerated	9.93477E-70
	C14805T		0	5	6,799	0.00	0.27	8.55	1.00	Benign/tolerated	2.09768E-45
	C15324T		41	73	1,588	8.17	4.01	2.00	1.00	Benign/tolerated	2.2731E-28
	A16512G		53	0	13	10.56	0.00	0.02	1.00	Benign/tolerated	0
	C18568T	Leu6102Phe	72	1	50	14.34	0.05	0.06	0.01	Deleterious	0
	C18877T		286	147	2,075	56.97	8.07	2.61	1.00	Benign/tolerated	0
	C19154T	Thr6297Ile	47	0	5	9.36	0.00	0.01	0.21	Benign/tolerated	0
	A20268G		0	3	4,650	0.00	0.16	5.85	1.00	Benign/tolerated	1.27368E-30
S	G21724T	Leu54Phe	95	4	304	18.92	0.22	0.38	0.69	Benign/tolerated	0
	C22444T		218	96	201	43.43	5.27	0.25	1.00	Benign/tolerated	0
	A23403G	Asp614Gly	472	1,142	61,751	94.02	62.71	77.66	0.30	Benign/tolerated	2.08832E-67
	C23929T		12	408	858	2.39	22.41	1.08	1.00	Benign/tolerated	0
ORF3a	C25528T	Leu46Phe	0	110	194	0.00	6.04	0.24	0.00	Deleterious	0
	G25563T	Gln57His	290	147	18,045	57.77	8.07	22.69	0.00	Deleterious	1.1597E-125
	G26144T	Gly251Val	0	4	5,385	0.00	0.22	6.77	0.00	Deleterious	1.93496E-35
M	C26735T		277	154	797	55.18	8.46	1.00	1.00	Benign/tolerated	0
ORF8	T28144C	Leu84Ser	20	75	5,636	3.98	4.12	7.09	0.37	Benign/tolerated	1.70788E-07
N	C28311T	Pro13Leu	13	413	1,151	2.59	22.68	1.45	0.00	Deleterious	0
	C28854T	Ser194Leu	201	106	1,948	40.04	5.82	2.45	0.05	Deleterious	0
	GGG28881AAC	ArgGly203LysArg	11	642	26,021	2.19	35.25	32.72	0.00	Deleterious	5.3828E-48
3' UTR	C29750T		75	0	42	14.94	0.00	0.05	#N/A	#N/A	0
	G29868A		0	353	42	0.00	19.38	0.05	#N/A	#N/A	0

TABLE 2 | Comparison of missense mutation frequency in deceased 481 vs recovered patients from global dataset.

NT mutation	AA mutation	Global mutation count (genomes)		Global frequency (%)		SIFT score	Functional effect	*p*-Value
		Deceased (*n* = 276)	Recovered (*n* = 1,845)	Deceased	Recovered			
C14408T	Pro4715Leu	245	1,450	88.77	78.59	0.31	Benign/tolerated	8.28E-05
A23403G	Asp614Gly	205	1,403	74.28	76.04	0.3	Benign/tolerated	0.522342
G25563T	Gln57His	112	495	40.58	26.83	0.00	Deleterious	2.43E-06
GGG28881AAC	ArgGly203LysArg	101	579	39.45	31.38	0.00	Deleterious	0.083557
C1059T	Thr265Ile	23	206	8.33	11.17	0.03	Deleterious	0.157376
C28854T	Ser194Leu	20	59	7.25	3.20	0.05	Deleterious	0.000924
G25088T	Val1176Phe	27	5	9.78	0.27	#N/A	#N/A	1.19E-33
T28144C	Leu84Ser	13	148	4.71	8.02	0.37	Benign/tolerated	0.052701
T12503C	Tyr4080His	0	109	0.00	5.91	0.00	Deleterious	3.38E-05
G11083T	Leu3606Phe	7	94	2.54	5.09	0.01	Deleterious	0.062656
G25770T	Arg126Ser	0	79	0.00	4.28	0.00	Deleterious	0.000459

TABLE 3 | Comparison of missense mutation frequency in deceased 485 vs recovered patients from Gujarat dataset.

NT mutation	AA mutation	Gujarat mutation count (genomes)		Gujarat frequency (%)		SIFT score	Functional effect	*p*-Value
		Deceased (*n* = 63)	Recovered (*n* = 256)	Deceased	Recovered			
A23403G	Asp614Gly	62	241	98.41	94.14	0.30	Benign/tolerated	0.164016
C14408T	Pro4715Leu	61	234	96.83	91.41	0.31	Benign/tolerated	0.144062
G25563T	Gln57His	39	142	61.90	55.47	0.00	Deleterious	0.355651
C28854T	Ser194Leu	30	90	47.62	35.16	0.00	Deleterious	0.067355
G16078A	Val5272Ile	7	10	11.11	3.91	0.00	Deleterious	0.022562
G23311T	Glu583Asp	5	10	7.94	3.91	0.33	Benign/tolerated	0.175819
C23277T	Thr572Ile	4	5	6.35	1.95	0.57	Benign/tolerated	0.059057
G21724T	Leu54Phe	3	39	4.76	15.23	0.69	Benign/tolerated	0.027646
C18568T	Leu6102Phe	2	33	3.17	12.89	0.01	Deleterious	0.027074
A2292C	Gln676Pro	2	31	3.17	12.11	0.05	Deleterious	0.036972

TABLE 4 | Chi-square test analysis of the deceased and recovered 490 patients for gender and age group.

		Gujarat (*n* = 319)		Global (*n* = 2,121)		
		Deceased	Recovered	Deceased	Recovered	*p*-Value
Total sample		63	256	276	1,845	0.00118
Gender	Male	37	178	203	1,002	0.89596
	Female	26	78	73	843	2.7E-08
Age (years)	0–40	2	94	18	865	0.97648
	41–60	28	115	101	675	0.03783
	> 60	33	47	157	305	0.20849

enzymatic nucleophilic substitution reactions. This protein is expressed abundantly in infected and transfected cells, which localizes to the intracellular and plasma membranes and induces apoptosis in transfected and infected cells (Issa et al., 2020). This enzyme mediates the extensive proteolytic processing of two overlapping replicase polyproteins, pp1a and pp1ab, to yield the corresponding functional polypeptides that are essential for coronavirus replication and transcription processes (Kohlmeier and Woodland, 2009; Benvenuto et al., 2020). While in the case of mutation at position C28311T leading to change of amino acid proline to leucine, the enzyme lies in the N gene that has a role in virion assembly and release and plays a significant role in the formation of replication–transcription complexes (Alsaadi and Jones, 2019; Liu, 2019; Wu et al., 2020; Yin, 2020). Similarly, the N protein is a highly basic protein that could modulate viral RNA synthesis (Millet and Whittaker, 2015; Hassan et al., 2020; Sarif Hassan et al., 2020).

The SIFT scores of these mutations were determined and also signify the functional effect change in whether an amino acid substitution affects protein function or not in terms of the deleterious effect or benign tolerated (Vaser et al., 2016). The predicted SIFT score of the mutation G25563T in the Orf3a and C28854T in the N gene was classified to be deleterious in nature. Similarly, a comparison analysis of the global (*n* = 79,518),

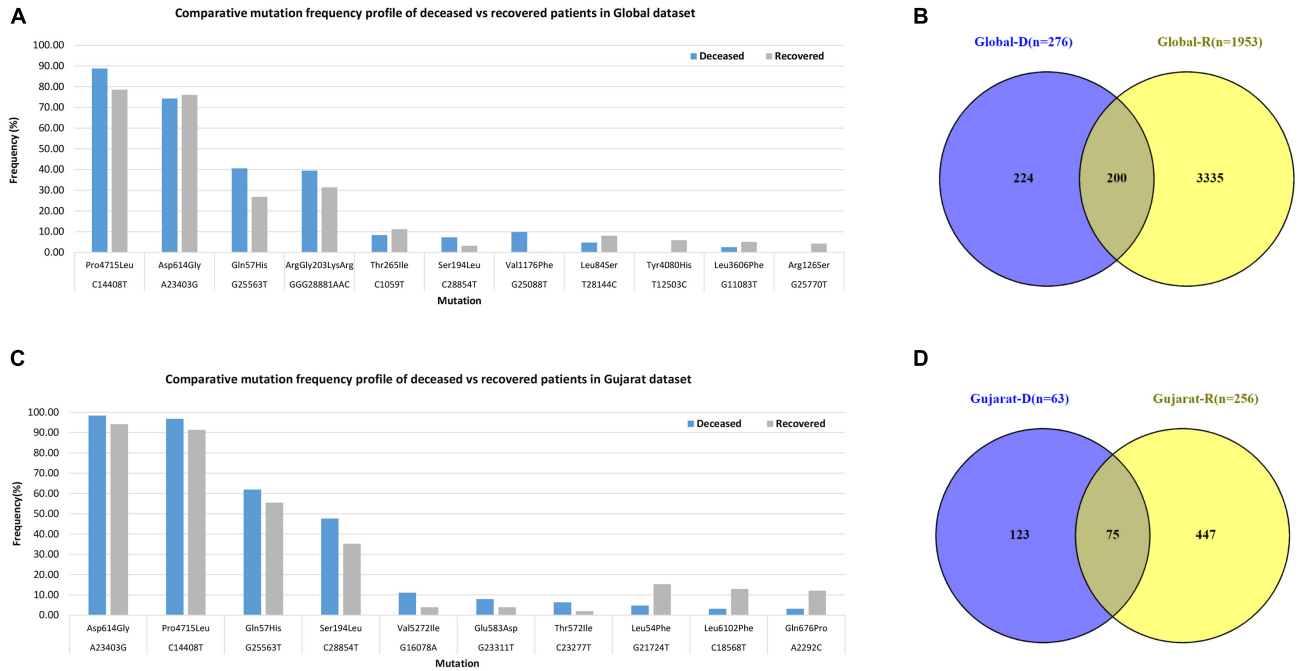

FIGURE 6 | Frequency of missense mutations in SARS-CoV-2 viral genome from global dataset. **(A)** Bar chart for global deceased versus recovered patients. **(B)** Venn diagram of the global deceased versus recovered patients. **(C)** Bar chart for the Gujarat deceased versus recovered patients. **(D)** Venn diagram of the Gujarat deceased versus recovered patients.

India (n = 1,821), and Gujarat datasets (n = 502), where the "n" is the number of genomes included in the analysis, indicates the overall dominance of C241T, C3037T, A23403G, C14408T, and G25563T. Furthermore, it is suggestive of the comparative dominant mutation profile, including non-synonymous and missense mutations. The analysis of the dataset from the global deceased (n = 276) and recovered patients (n = 1,845) with known status from the metadata information available on the GISAID server with the complete genome sequences considered in the analysis indicates the dominance of the missense mutations at A23403G, C14408T, C1059T, and G25563T. The overall comparison of the mutation profile of the patient dataset of deceased and recovered samples is highlighted in **Figure 7**, from global and Gujarat.

Mutation in the N gene at C28854T and mutation in the Orf3a gene at G25563T were found to be dominant among deceased patients from Gujarat. Moreover, C28854T is forming a distinct sub-cluster under 20A (A2a as per the old classification of the next strain) clade, highlighted in **Figure 8**. The same is proposed as a new sub-clade 20D in the next strain and GHJ in GISAID. This sub-clade is also present in genomes sequenced from Bangladesh and Saudi Arabia. Both these proteins play significant roles in viral replication and pathogenesis (Luan et al., 2020; Pachetti et al., 2020; Peter and Schug, 2020). The association of the mutations with the viral transmission and mortality rate remains a mystifying puzzle for the global scientific community. The identification and validation of these mutations should pave the way forward for the development of treatment and diagnostics of coronavirus disease. The evading host immune response and defense mechanism sufficiently improve the adaptive behavior of the pathogenic species, thus making them highly contagious. Further, laboratory and experimental studies need to be carried out to validate the exact role of this particular mutation with respect to the molecular pathways and

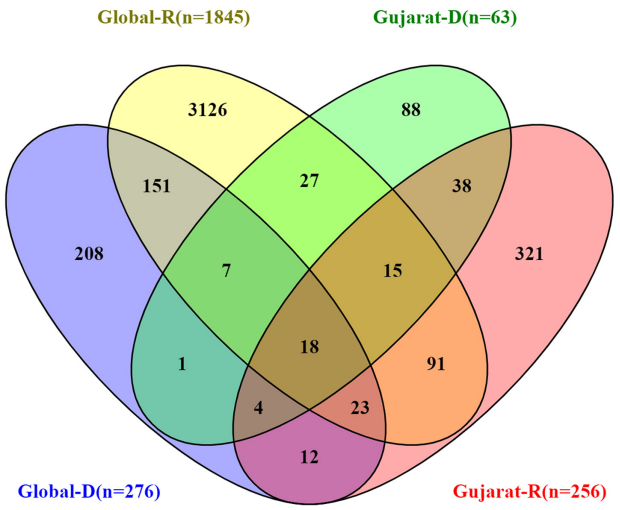

FIGURE 7 | Overall comparison of the missense mutations in SARS-CoV-2 genome. Gujarat (R = 256, D = 63) and Global (R = 1,845, D = 276), where "R" is the number of genomes from recovered patients, and "D" is the number of genomes from deceased patients.

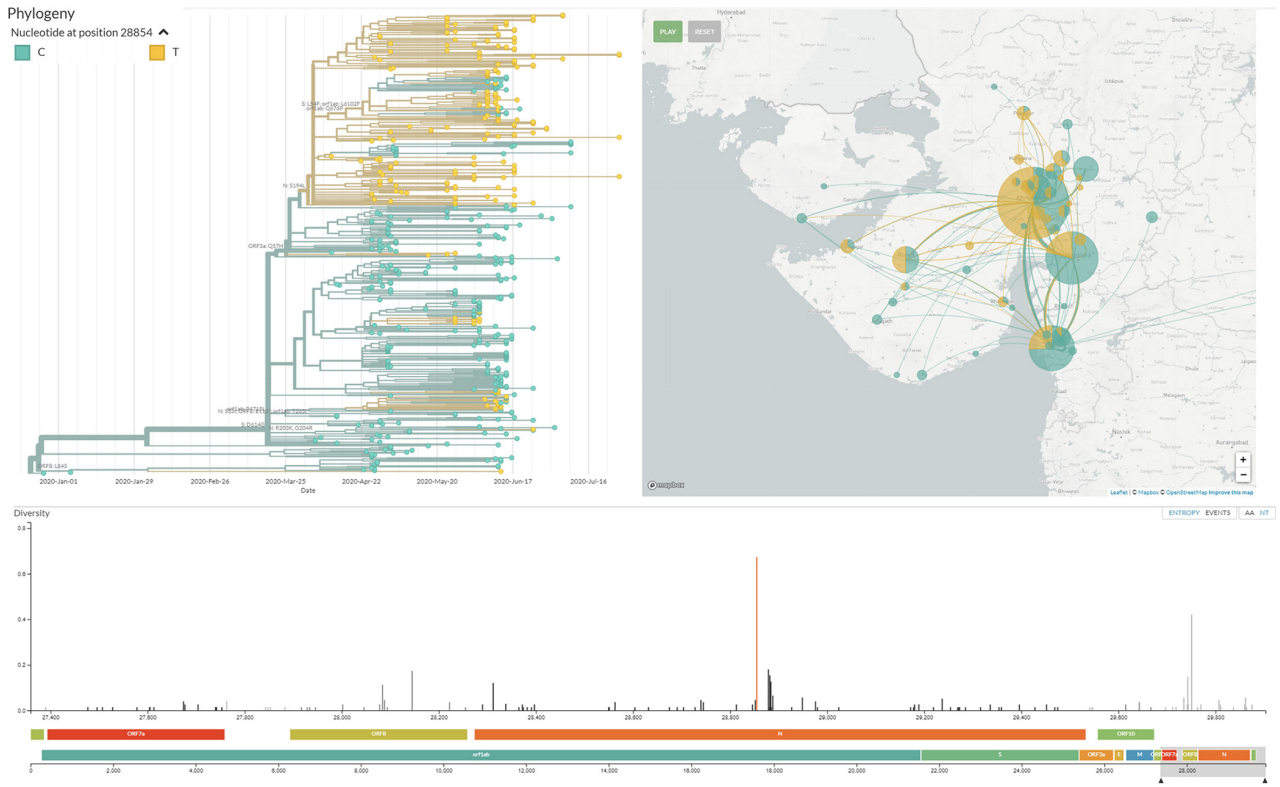

FIGURE 8 | Distinct cluster of the viral isolate with mutation C28854T/Ser194Leu/N gene in Gujarat SARS-CoV-2 genomes. This cluster is visualized at http://covid.gbrc.org.in/nextstrain.php using the Nextstrain virus genome analysis pipeline.

interactions in the biological systems despite being a strong possible mutation candidate found in the Gujarat region.

The genomics-based approach has been a useful resource to identify and characterize virulence, pathogenicity, and host adaptability. Further, identification and characterization of the frequently mutated positions in the SARS-CoV-2 genome will certainly help in the deeper understanding of the infection biology of coronaviruses, development of vaccines and therapeutics, and potential drug repurpose candidates using predictive computational biology and experimental validations. The present study highlights the genome sequencing, haplotyping, and mutation profile of the 502 SARS-CoV-2 viral genome isolates from 46 different locations representing 20 districts across Gujarat, India. Furthermore, we have reported significant variants associated with mortality in the Gujarat and global viral genomes. As the pandemic is progressing, the virus is also diverging into different clades. This also provides adaptive advantages to viruses in the progression of the disease and its pandemic potential. In this study, we have reported a distinct cluster of coronavirus under 20A clade of Nextstrain and proposed it as 20D as per the next strain analysis or GHJ as per the GISAID analysis, predominantly present in the Gujarat genomes. Understanding the SARS-CoV-2 genome and tracking its evolution will help in devising better strategies for the development of diagnosis, treatment, and vaccine candidates in response to the global pandemic.

DATA AVAILABILITY STATEMENT

The raw data generated in this study have been submitted to the NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA625669. **Supplementary Dataset** to this manuscript are also available at Mendeley Data with DOI: 10.17632/pc38m6mwxt.1 (https://data.mendeley.com/datasets/pc38m6mwxt/draft?a=1aa66c2a-5b93-456f-816c-3f26a482dc2a). All datasets of COVID-19 are also provided on GBRC-COVID portal (http://covid.gbrc.org.in/).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Ethical Committee of GBRC and respective government medical colleges. Gujarat Biotechnology Research Centre, Gandhinagar B.J. Medical College and Civil hospital, Ahmedabad Banas Medical College and Research Institute, Palanpur Department of MicroBiology, Government Medical College, Surat Dr. N. D. Desai Medical College & Hospital, Nadiad Dr. RSS Hospital, Modasa GAIMS & G K General Hospital, Bhuj GMERS Medical College and Hospital, Gandhinagar GMERS Medical College and Hospital, Gotri, Vadodara GMERS Medical College and Hospital, Himmatnagar Government Medical College, Bhavnagar

Government Medical College, Vadodara Pandit Deendayal Upadhyay Government Medical College, Rajkot Saikrishna Hospital, Mehsana Sardar Vallabhbhai Patel Institute of Medical Sciences & Research. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MJ, SB, and CJ conceptualized the work plan and guided it for analysis of primary data, interpretation of data, and editing of the manuscript. AP, DK, AA, and MJ retrieved and analyzed the data and generated the tables and figures under supervision of CJ. MJ, DK, and AP wrote the manuscript. MP, JR, ZP, PT, and MG did the sample processing and RNA isolation. LP, KP, and NS did the genome sequencing. SK did the data analysis and manuscript editing. All authors contributed to the article and approved the submitted version.

FUNDING

Department of Science and Technology (DST), Government of Gujarat, Gandhinagar, India.

ACKNOWLEDGMENTS

The authors are grateful to the Secretary, Department of Science and Technology (DST), and Health Commissioner, Government of Gujarat, Gandhinagar, Gujarat, India. The authors are also thankful to the clinical staff for extending support in the sample collection. The authors would like to acknowledge Dr. Manish Kumar, IIT-Gandhinagar for critically reviewing and providing essential inputs in the writing of the manuscript and Dr. Raghawendra Kumar and Mr. Zuber Saiyed for providing additional support to the genome assembly of viral genomes. The authors gratefully acknowledge the authors from the originating laboratories responsible for obtaining the specimens, as well as the submitting laboratories where the genome data were generated and shared *via* GISAID, on which this research is based for comparison of the Indian and global genome analyses. All submitters of data may be contacted directly *via* www.gisaid.org.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.586569/full#supplementary-material

REFERENCES

Alsaadi, E. A. J., and Jones, I. M. (2019). Membrane binding proteins of coronaviruses. *Future Virol.* 14, 275–286. doi: 10.2217/fvl-2018-2144

Andrews, S. (2010). *FastQC: a Quality Control Tool for High Throughput Sequence Data*. Babraham: Babraham institute.

Bajaj, A., and Purohit, H. J. (2020). Understanding SARS-CoV-2: genetic diversity, transmission and cure in human. *Indian J. Microbiol.* 60, 1–4. doi: 10.1007/s12088-020-00869-864

Benvenuto, D., Angeletti, S., Giovanetti, M., Bianchi, M., Pascarella, S., Cauda, R., et al. (2020). Evolutionary analysis of SARS-CoV-2: how mutation of non-structural protein 6 (NSP6) could affect viral autophagy. *J. Infect.* 81, e24–e27. doi: 10.1016/j.jinf.2020.03.058

Chu, D. K. W., Pan, Y., Cheng, S. M. S., Hui, K. P. Y., Krishnan, P., Liu, Y., et al. (2020). Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia. *Clin. Chem.* 555, 549–555. doi: 10.1093/clinchem/hvaa029

Drummond, A. J., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

Du, L., He, Y., Zhou, Y., Liu, S., Zheng, B. J., and Jiang, S. (2009). The spike protein of SARS-CoV - a target for vaccine and therapeutic development. *Nat. Rev. Microbiol.* 7, 226–236. doi: 10.1038/nrmicro2090

Evenett, S. J., and Winters, L. A. (2020). *Preparing for a Second Wave of Covid-19: A Trade Bargain to Secure Supplies of Medical Goods. Global Trade Alert*. Available online at: https://blogs.sussex.ac.uk/uktpo/publications/preparing-for-a-second-wave-of-covid-19-a-trade-bargain-to-secure-supplies-of-medical-goods/ (accessed July 20, 2020).

Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R. H., Peters, B., and Sette, A. (2020). A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe* 27, 671—-680.e2. doi: 10.1016/j.chom.2020.03.002

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* 382, 1708–1720. doi: 10.1056/NEJMoa2002032

Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., et al. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak- a n update on the status. *Mil. Med. Res.* 7:11. doi: 10.1186/s40779-020-00240-0

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). NextStrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123. doi: 10.1093/bioinformatics/bty407

Hassan, S. S., Choudhury, P. P., Basu, P., and Jana, S. S. (2020). Molecular conservation and differential mutation on ORF3a gene in Indian SARS-CoV2 genomes. *Genomics* 112, 3226–3237. doi: 10.1016/j.ygeno.2020.06.016

Imbert, I., Guillemot, J. C., Bourhis, J. M., Bussetta, C., Coutard, B., Egloff, M. P., et al. (2006). A second, non-canonical RNA-dependent RNA polymerase in SARS coronavirus. *EMBO J.* 25, 4933–4942. doi: 10.1038/sj.emboj.7601368

Issa, E., Merhi, G., Panossian, B., Salloum, T., and Tokajian, S. (2020). SARS-CoV-2 and ORF3a: nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems* 5:e00266-20. doi: 10.1128/msystems.00266-220

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kohlmeier, J. E., and Woodland, D. L. (2009). Immunity to respiratory viruses. *Annu. Rev. Immunol.* 27, 61–82. doi: 10.1146/annurev.immunol.021908.132625

Laxminarayan, R., Wahl, B., Dudala, S. R., Gopal, K., Mohan, B. C., Neelima, S., et al. (2020). Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science* 370, 691–697. doi: 10.1126/science.abd7672

Leung, K., Wu, J. T., Liu, D., and Leung, G. M. (2020). First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet* 395, 1382–1393. doi: 10.1016/S0140-6736(20)30746-30747

Li, W., Zhang, C., Sui, J., Kuhn, J. H., Moore, M. J., Luo, S., et al. (2005). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J.* 24, 1634–1643. doi: 10.1038/sj.emboj.7600640

Liu, D. X. (2019). Human coronavirus: host-pathogen interaction. *Annu. Rev. Microbiol.* 73, 529–557. doi: 10.1146/annurev-micro-020518

Luan, J., Lu, Y., Jin, X., and Zhang, L. (2020). Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection. *Biochem. Biophys. Res. Commun.* 526, 165–169. doi: 10.1016/j.bbrc.2020.03.047

Mercatelli, D., and Giorgi, F. M. (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. *Front. Microbiol.* 11:800. doi: 10.3389/fmicb.2020.01800

Millet, J. K., and Whittaker, G. R. (2015). Host cell proteases: critical determinants of coronavirus tropism and pathogenesis. *Virus Res.* 202, 120–134. doi: 10.1016/j.virusres.2014.11.021

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Mondal, M., Lawarde, A., and Somasundaram, K. (2020). Genomics of Indian SARS-CoV-2: implications in genetic diversity, possible origin and spread of virus. *Medrxiv* [preprint] doi: 10.1101/2020.04.25.20079475

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* 18:179. doi: 10.1186/s12967-020-02344-6

Peter, E. K., and Schug, A. (2020). The inhibitory effect of a Corona virus spike protein fragment with ACE2. *bioRxiv* [preprint] doi: 10.1101/2020.06.03.132506

Potdar, V., Cherian, S., Deshpande, G., Ullas, P., Yadav, P., Choudhary, M., et al. (2020). Genomic analysis of SARS-CoV-2 strains among Indians returning from Italy, Iran & China, & Italian tourists in India. *Indian J. Med. Res.* 151, 255–260. doi: 10.4103/ijmr.IJMR_1058_20

Rambaut, A. (2018). *FigTree 1.4.4 Software*. Edinburgh: Institute of Evolutionary Biology University.

Sagulenko, P., Puller, V., and Neher, R. A. (2018). TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol.* 4:vex042. doi: 10.1093/ve/vex042

Sarif Hassan, S., Pal Choudhury, P., Roy, B., and Sankar Jana, S. (2020). Missense mutations in SARS-CoV2 genomes from Indian patients. *Genomics* 112, 4622–4627.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Shastri, A., Wheat, J., Agrawal, S., Chaterjee, N., Pradhan, K., Goldfinger, M., et al. (2020). Delayed clearance of SARS-CoV2 in male compared to female patients: high ACE2 expression in testes suggests possible existence of gender-specific viral reservoirs. *medRxiv* [preprint] doi: 10.1101/2020.04.16.20060566

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., and Siddique, R. (2020). COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. *J. Adv. Res.* 24, 91–98. doi: 10.1016/j.jare.2020.03.005

Strzelecki, A. (2020). The second worldwide wave of interest in coronavirus since the COVID-19 outbreaks in South Korea, Italy and Iran: a google trends study. *Brain. Behav. Immun.* 88, 950–951. doi: 10.1016/j.bbi.2020.04.042

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* 83:104351. doi: 10.1016/j.meegid.2020.104351

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9. doi: 10.1038/nprot.2015.123

Velazquez-Salinas, L., Zarate, S., Eberl, S., Novella, I., and Borca, M. V. (2020). Positive selection of ORF3a and ORF8 genes drives the evolution of SARS-CoV-2 during the 2020 COVID-19 pandemic. *bioRxiv* [preprint] doi: 10.1101/2020.04.10.035964

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., et al. (2020). Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* 27, 325–328. doi: 10.1016/j.chom.2020.02.001

Xu, S., and Li, Y. (2020). Beware of the second wave of COVID-19. *Lancet* 395, 1321–1322. doi: 10.1016/S0140-6736(20)30845-X

Yadav, P., Potdar, V., Choudhary, M., Nyayanit, D., Agrawal, M., Jadhav, S., et al. (2020). Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J. Med. Res.* 151, 200–209. doi: 10.4103/ijmr.IJMR_663_20

Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics* 112, 3588–3596. doi: 10.1016/j.ygeno.2020.04.016

Zhang, D., Gao, F., Jakovliæ, I., Zou, H., Zhang, J., Li, W. X., et al. (2020). PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* 20, 348–355. doi: 10.1111/1755-0998.1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 April 2021
doi: 10.3389/fgene.2021.644516

G2S: A New Deep Learning Tool for Predicting Stool Microbiome Structure From Oral Microbiome Data

Simone Rampelli[1], Marco Fabbrini[1,2], Marco Candela[1], Elena Biagi[1], Patrizia Brigidi[2] and Silvia Turroni[1]*

[1] *Unit of Microbiome Science and Biotechnology, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy,* [2] *Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy*

Deep learning methodologies have revolutionized prediction in many fields and show the potential to do the same in microbial metagenomics. However, deep learning is still unexplored in the field of microbiology, with only a few software designed to work with microbiome data. Within the meta-community theory, we foresee new perspectives for the development and application of deep learning algorithms in the field of the human microbiome. In this context, we developed G2S, a bioinformatic tool for taxonomic prediction of the human fecal microbiome directly from the oral microbiome data of the same individual. The tool uses a deep convolutional neural network trained on paired oral and fecal samples from populations across the globe, which allows inferring the stool microbiome at the family level more accurately than other available approaches. The tool can be used in retrospective studies, where fecal sampling was not performed, and especially in the field of paleomicrobiology, as a unique opportunity to recover data related to ancient gut microbiome configurations. G2S was validated on already characterized oral and fecal sample pairs, and then applied to ancient microbiome data from dental calculi, to derive putative intestinal components in medieval subjects.

Keywords: gut microbiome, oral microbiome, deep learning, microbiome, paleomicrobiology

OPEN ACCESS

Edited by:
Saumya Patel,
Gujarat University, India

Reviewed by:
Francesco Asnicar,
University of Trento, Italy
Khanh N. Q. Le,
Taipei Medical University, Taiwan

****Correspondence:***
Simone Rampelli
simone.rampelli@unibo.it

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *21 December 2020*
Accepted: *09 March 2021*
Published: *09 April 2021*

Citation:
Rampelli S, Fabbrini M, Candela M, Biagi E, Brigidi P and Turroni S (2021) G2S: A New Deep Learning Tool for Predicting Stool Microbiome Structure From Oral Microbiome Data. Front. Genet. 12:644516. doi: 10.3389/fgene.2021.644516

INTRODUCTION

Deep learning is increasingly being used to make inference on large and complex data. Unlike traditional algorithms, in which the expertise and rules are already coded, deep learning algorithms are built to automatically detect patterns in data (Murphy, 2012; Bishop, 2016), also embedding the computation of variables into the models themselves to yield end-to-end models (Goodfellow et al., 2016). In particular, the construction and training of deep learning algorithms have been enabled by the increasing availability of big data and the rapid growth in the number and size of public available databases. So far, deep neural networks have been key to advances in modern

artificial intelligence, with applications such as facial recognition, speech recognition and self-driving vehicles. More recently, new applications have been pioneered in the fields of molecular biology and metagenomics. Indeed, the same deep learning approaches are beginning to be applied to genetics, agriculture and medicine (Alipanahi et al., 2015; Leung et al., 2016; Ching et al., 2018; Demirci et al., 2018; Wainberg et al., 2018; Webb, 2018; Le, 2019; Le and Huynh, 2019; Le et al., 2019; Quang and Xie, 2019). However, deep learning is still unexplored in the field of microbial metagenomics, with only a few approaches suitable for microbiome data (Geman et al., 2016; Reiman et al., 2017; Galkin et al., 2020), and a huge untapped potential yet unexplored.

The human microbiome, i.e., the sum of the different microbial ecosystems that colonize the niches of the human body, plays an important role in human physiology and its dysbiotic variations can severely impact our health (Kau et al., 2011). For example, shifts in the composition of microbial communities inhabiting the oral cavity and gastrointestinal tract have been associated with the onset and/or progression of various conditions, such as periodontitis (Griffen et al., 2012) and other modern chronic disorders, including inflammatory bowel disease (Glassner et al., 2020), obesity (Rampelli et al., 2018), cardiovascular disease (Pietiäinen et al., 2018) and some forms of cancer (Helmink et al., 2019; Karpiński, 2019; Wong and Yu, 2019). The importance of the human microbiome in health and disease makes it imperative to understand the drivers of its variation. In this context, a new frontier is represented by the meta-community theory, according to which human symbiont microbial ecosystems are in intimate connection, showing reciprocal influences and exchanges (Koskella et al., 2017; Miller et al., 2018). Supporting a meta-community view of human microbial ecology, a close link between oral and intestinal microbiomes has recently been hypothesized, with the former reflecting changes in the latter, in both healthy and diseased individuals (Bajaj et al., 2015; Iwauchi et al., 2019; Prodan et al., 2019; Schmidt et al., 2019). Another scale of human microbiome variation is represented by its change across the evolutionary timeline. In particular, a large body of literature indicates that the current human gut microbiome has evolved toward at least two different configurations, rural and urban, both associated with the corresponding subsistence strategy. Compared to the first, generally considered as the pristine human gut microbiome, the urban configuration is characterized by an overall compression of microbial biodiversity, a wholescale loss of commensal microbial groups, and an increased presence of genes related to antibiotic resistance and xenobiotics metabolism (Yatsunenko et al., 2012; Schnorr et al., 2014; Obregon-Tito et al., 2015; Rampelli et al., 2015; Ayeni et al., 2018; Jha et al., 2018). These changes, collectively referred to as "microbiota insufficiency syndrome" (Sonnenburg and Sonnenburg, 2019), have been identified as contributing factors to the rise in chronic inflammatory non-communicable diseases. However, mainly due to the paucity of ancient stool samples, the truly ancestral human gut microbiome is still unknown and the evolutionary trajectories and drivers leading to its contemporary configurations have yet to be described, leaving important gaps in knowledge of the gut microbiome-human host co-evolutionary trajectories. Contrary to ancient fecal samples, dental ones are more common and well preserved, allowing for the extraction of the ancient oral microbiome from ancient DNA preserved in dental tartar. Consistent with the meta-community vision, the ancient configuration of the oral microbiome can somehow mirror the structural features of the intestinal one due to the intrinsic connections between the two ecosystems. In this scenario, here we developed a new deep learning-based tool, G2S, which infers the gut microbiome configuration from the oral microbiome data of a given individual. G2S is based on a model trained and tested on a total of 305 and 79 paired samples of oral and stool microbiome, respectively, retrieved from multiple studies with individuals of various geographical origins, including United States, Fiji, United Kingdom, and European countries (The Human Microbiome Project Consortium, 2012; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Our approach may be relevant for predicting the gut microbiome configuration when fecal data are not available, and particularly suitable for human archeological records, where coprolites and fecal sediments are indeed rare compared to dental calculi and other human remains.

MATERIALS AND METHODS

G2S software is built in an R environment, using the R packages "base," "stats," and "keras," containing "tensorflow." The G2S source code is available on the website https://github.com/simonerampelli/g2s and it can be run using a command line interface on computer with Windows, Linux and OS X as the operating system.

The G2S tool was trained and tested on a total of 768 paired samples (i.e., oral and stool samples from the same 384 individuals), including samples from 171 healthy adults from United States, 7 from Italy, 29 from Sweden, 37 from United Kingdom, and 140 from Fiji (The Human Microbiome Project Consortium, 2012; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Eighty% of the subjects were used for the training dataset and 20% for the test dataset, without overlapping to avoid overfitting. Both 16S rRNA gene reads and shotgun metagenomics sequences were used, analyzed by the QIIME 2 pipeline (Bolyen et al., 2019) or the MetaPhlAn2 software (Truong et al., 2015), respectively.

The performance of G2S in predicting fecal microbiome configuration from the same individual's oral microbiome sample was compared with that of other available approaches, including Random Forest (Breiman, 2001) and a stochastic algorithm, i.e., a customized method that generates mock profiles of the stool microbiome by randomly imputing the abundances of bacterial families in the range of the training dataset (see **Supplementary File 1** for script source).

Microbiome data from dental calculi of 4 adult human skeletons (G12, B17, B61, and B78), characterized by sequencing the V5 and V6 regions of the 16S rRNA gene (8 samples in total) (Warinner et al., 2014), were used to illustrate the potential and

results of G2S. No ethics committee approval was required to perform the analysis included in this study.

RESULTS

Implementation of the G2S Software

G2S adapted a deep convolutional neural network (ConvNet) to predict gut microbiome configurations from oral microbiome data. Several model architectures were tested in order to find the best performing algorithm, either by testing hidden layers with different number of units, and/or by adding a weight regularization step or a dropout procedure (data not shown). The final ConvNet was structured with two hidden layers, each with 50 units, and a final linear layer with 13 units and no activation function. We selected mean square error as the loss function, and mean absolute error as the metric to evaluate the differences between predictions and targets during training. In order to minimize overfitting problems due to the small number of samples within the dataset, we also included a weight regularization step, by adding to the loss function a cost associated with having high weights. The cost was proportional to the square of the weight coefficient value (L2 regularization or weight decay). Finally, to further prevent overfitting, dropout was applied to the first two layers, obtaining a better prediction and a significant reduction in losses and minimum absolute errors with a rate value of 0.5.

For ConvNet training and testing, we downloaded all available paired samples (i.e., gingival and stool samples from the same individual) from the HMP project (The Human Microbiome Project Consortium, 2012). In order to increase the generalization capability of our ConvNets, while minimizing geography-related bias (He et al., 2018), we integrated our dataset with all available paired samples (i.e., oral and fecal samples) from healthy adults from other literature studies (Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018), selecting both 16S rRNA gene and shotgun metagenomic datasets (see also **Supplementary Table 1**). Our final dataset included paired samples of 171 individuals from United States, 7 from Italy, 29 from Sweden, 37 from United Kingdom, and 140 from Fiji, for a total of 384 oral and 384 stool samples, divided into 528 16S rRNA gene and 240 shotgun fastq files. Specifically, 16S rRNA gene sequences were analyzed using the QIIME 2 pipeline (Bolyen et al., 2019) and the Greengenes database (DeSantis et al., 2006) in order to obtain the microbiome classification at different taxonomic levels. On the other hand, the shotgun metagenomic samples were analyzed by MetaPhlAn2 (Truong et al., 2015) using the default parameters. The genus-level abundance table of 384 oral microbiome samples was normalized feature-wise prior to its usage for deep learning. In particular, the data were centered on the mean of each specific genus and scaled according to their standard deviation. Only 50 genera present in more than 4 samples with relative abundance greater than 0.1% were retained for the analysis. The 12 bacterial families of the stool microbiome dataset with the highest contribution in terms of median relative abundance, including *Bacteroidaceae, Porphyromonadaceae, Lachnospiraceae, Ruminococcaceae, Veillonellaceae, Rikenellaceae, Alcaligenaceae, Streptococcaceae, Bifidobacteriaceae, Clostridiaceae, Prevotellaceae*, and *Erysipelotrichaceae*, were selected as features to be predicted by ConvNet analysis. An additional variable, called "Other" (i.e., the percentage remaining to reach 100%), was also considered a feature to be inferred. The training and test datasets were separated to contain 80 and 20% of all profiles, i.e., 305 and 79 paired oral and fecal samples, respectively. In order to better evaluate the model, we used a k-fold cross-validation approach with 4 partitions and 500 epochs. We got the best performance after the 151st epoch, with a mean absolute error of 4.1%. To increase the predictive performance of ConvNet, the results were then transformed as follows: (i) negative predictions were set to 0, and (ii) the sum of the value for each sample was rescaled to 100%. Finally, based on the results of the training dataset, we also built a confusion matrix to adjust the predictions of those families with recurring over- or underestimation. G2S includes all of these steps in a single R script, and requires only a relative abundance table of the oral microbiome (between 0 and 1) at the genus level with samples in the columns and

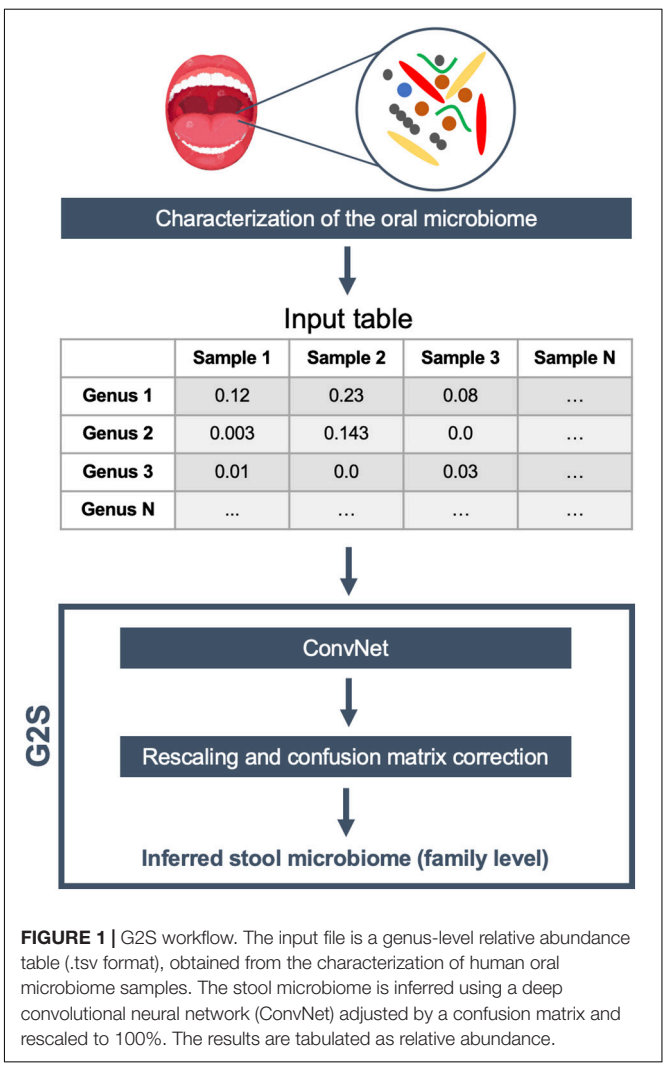

FIGURE 1 | G2S workflow. The input file is a genus-level relative abundance table (.tsv format), obtained from the characterization of human oral microbiome samples. The stool microbiome is inferred using a deep convolutional neural network (ConvNet) adjusted by a confusion matrix and rescaled to 100%. The results are tabulated as relative abundance.

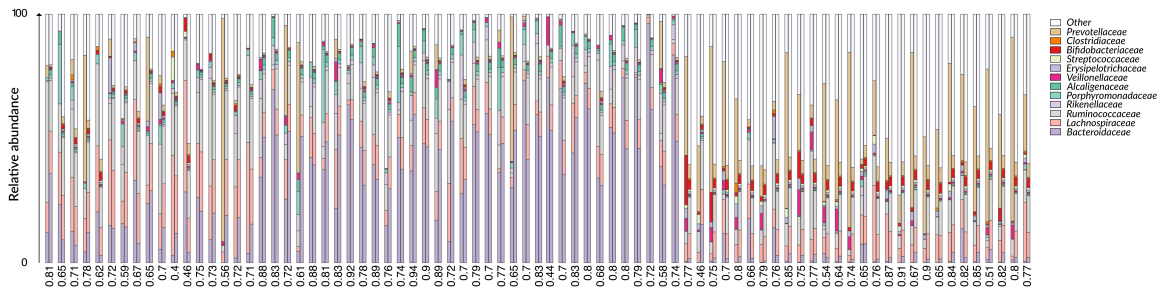

FIGURE 2 | Comparison between G2S predictions and real data from the test dataset. The family level bar plots of the 79 stool samples of the test dataset are visualized next to their inferred configurations obtained by G2S. Spearman correlation coefficients (r) are provided below each pair of bar plots. Samples are derived from the following studies: The Human Microbiome Project Consortium (2012), Zaura et al. (2015); Brito et al. (2016), Russo et al. (2018).

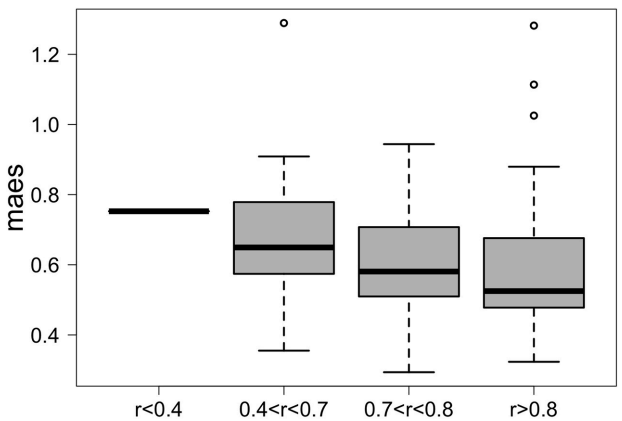

FIGURE 3 | G2S predictions are more accurate when the configurations to be inferred fall within the plane of variation of the training dataset. Box plots of the mean absolute error scaled to one standard deviation (maes) between the real stool microbiome configuration of the samples in the test dataset and the median configuration of the training dataset. Samples were divided into four groups based on the quality of the G2S predictions (i.e., the Spearman correlation coefficients between the real values and the inferred configurations).

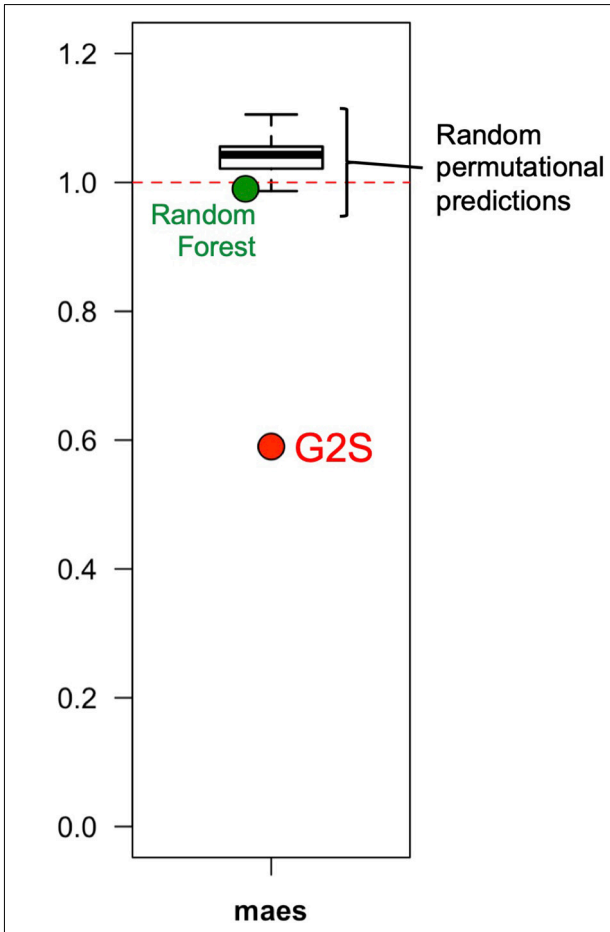

FIGURE 4 | G2S predicts the stool microbiome configuration with better performance than other methods. The mean absolute errors scaled to one standard deviation (maes) between the real data of the samples from the test dataset and the configurations inferred by G2S, Random Forest and a stochastic permutational method (100 predictions), are reported in the dot plot.

the full taxonomy following the Greengenes_05_2013 style in the rows as input file. For each sample analyzed, the predicted microbiome is summarized in a table as the relative abundance of the most abundant bacterial families. Additionally, histograms of the same families are provided, using the "graphics" and "base" R packages. The schematic overview of the G2S framework is provided in **Figure 1**.

Ascertaining the Performance of G2S on the Test Dataset

We first applied G2S to the test dataset to evaluate its cross-validated predictions. In particular, mean absolute errors for each family scaled to one standard deviation of real data (maes) < 1 were considered as reference parameters for a good quality of the prediction. As expected, G2S predicts relative abundances with an average maes of 0.59, ranging from the best score for *Bacteroidaceae* and *Erysipelotrichaceae* (maes = 0.46) to the worst case for *Ruminococcaceae* (maes = 0.77). To gain more insights into the predictive performance of G2S, we globally compared, sample by sample, the inferred microbiome configurations with

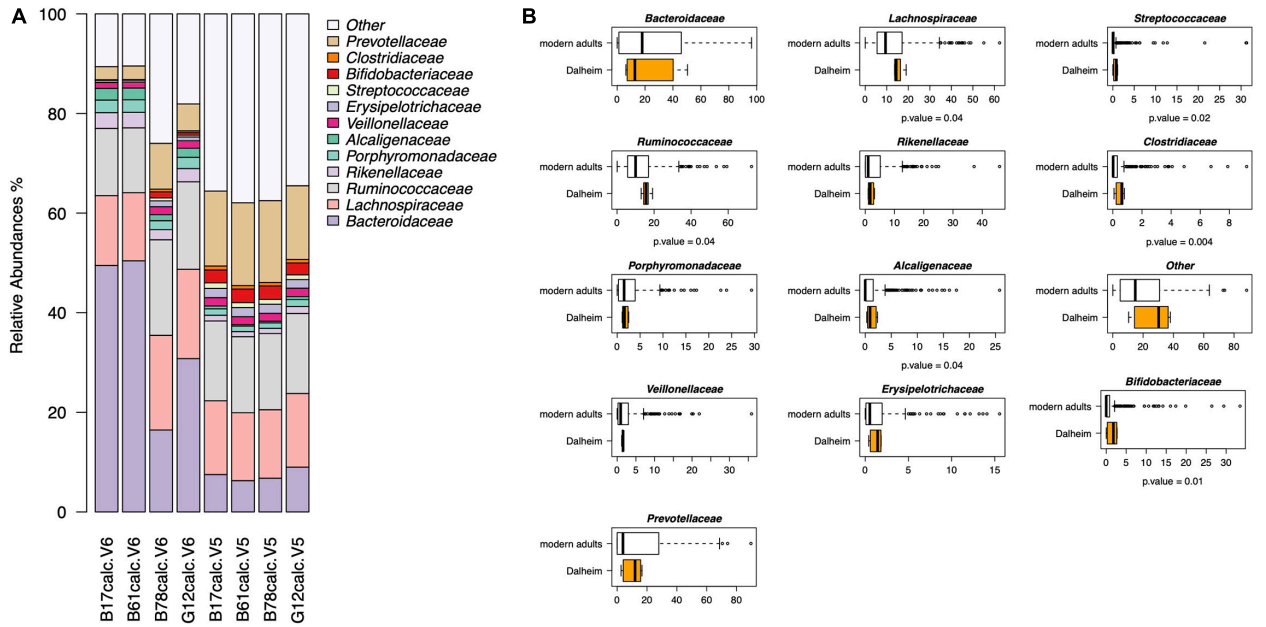

FIGURE 5 | Reconstructing the ancient stool microbiome of adult medieval individuals. **(A)** Bar plots of stool microbiome configurations inferred from 16S rRNA gene (V5 and V6 regions) sequencing data of ancient microbiomes (i.e., dental calculi from the medieval monastic site of Dalheim, Germany [ca. 950–1,200 CE]) (Warinner et al., 2014). **(B)** Comparison between the predicted ancient microbiome configurations and the modern stool microbiome of subjects from the dataset used to implement G2S (The Human Microbiome Project Consortium, 2012; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). *P*-values were determined by Wilcoxon test.

real data by means of bar plots (**Figure 2**). Spearman correlations between predicted and actual microbiome profiles were used to evaluate predictions for each subject. In particular, we considered as excellent those predictions with r > 0.8 (52% of predictions), good those with r between 0.71 and 0.8 (29% of predictions), discrete with *r* between 0.41 and 0.7 (18% of predictions), and incorrect with $r \leq 0.4$ (1% of predictions). When we analyzed the single case in which G2S inferred an incorrect prediction, we found that the stool microbiome configuration was very peculiar, with the relative abundances of the two keystone bacterial families *Bacteroidaceae* and *Lachnospiraceae* not reaching 5% of relative abundance together (while generally dominant in the ecosystem). It is important to note that G2S worked correctly even when the stool microbiome configurations to be predicted were not so close to the median configuration of the training dataset (maes < 1 even when $r < 0.7$) (**Figure 3**). This was likely due to the large variation captured by the pool of microbiome configurations of the samples in the training dataset.

G2S showed a better mimicry of the relative abundance of microbiomes in the test dataset than other methods, including Random Forest and a stochastic method developed specifically for this comparison, which generates mock profiles of the stool microbiome in the range of the training dataset (**Figure 4**). Random Forest under- or overestimated bacterial families with a global maes of 0.99, ranging from 0.77 for *Bacteroidaceae* to 1.74 for *Streptococcaceae*. The performance of our custom predictor was even more inaccurate, with a total of 100 permutational predictions showing maes between 0.98 and 1.11 (mean = 1.05). The best performance of G2S in predicting the stool microbiome structure is probably due to the predictive power of deep learning that automatically detects patterns in the data, by also embedding the computation of variables into the models themselves to yield end-to-end models.

Case Study: Using G2S in Paleomicrobiology to Predict the Stool Microbiome Profile From Ancient Dental Calculi

In the second part of our analysis, we used G2S to infer the stool microbiome from oral microbiome data of four adult human skeletons with evidence of mild to severe periodontal disease, from the medieval monastic site of Dalheim, Germany (ca. 950–1,200 CE) (Warinner et al., 2014). G2S inferred the stool microbiome structure at the family level, estimating the abundance of the 13 features, i.e., the 12 bacterial families and the category "Other" including all other families (**Figure 5A**). Interestingly, *Bacteroidaceae, Lachnospiraceae, Ruminococcaceae*, and *Prevotellaceae* were the predicted dominant components in the feces of the four subjects, using both V5 and V6 regions as targets of the 16S rRNA gene (together their relative abundance ranged from 52 to 80%). On the other hand, the family *Clostridiaceae* showed the lowest relative abundance (<1%) in all eight samples. Significant differences in taxon relative abundance were found with respect to the stool microbiome of modern subjects from the dataset used to implement G2S, including higher relative abundance of *Ruminococcaceae, Lachnospiraceae, Streptococcaceae, Alcaligenaceae, Clostridiaceae*,

and *Bifidobacteriaceae* in the predicted ancient microbiome configurations (p-value < 0.05, Wilcoxon test) (**Figure 5B**). This is not unexpected given the profoundly different lifestyles of ancient individuals of the Middle Ages and modern people, in terms of diet, contact with the environment and sanitization practices (The Human Microbiome Project Consortium, 2012; Warinner et al., 2014; Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Future studies in larger worldwide cohorts, including paired samples of oral and intestinal microbiome, are needed to refine the accuracy of the G2S software and predict a higher number of bacterial families as well as possibly taxa at different phylogenetic levels, possibly including genera and species.

DISCUSSION

G2S is specifically designed to predict the structure of the human stool microbiome from oral microbiome data. In particular, it uses relative abundance tables of the oral microbiome generated by next-generation sequencing, and a deep learning approach that allows high-speed prediction of the stool microbiome without any downstream process. It could be used with both modern and ancient samples, providing a good prediction of the fecal microbiome with a net saving of time and costs. This is particularly relevant in the context of paleomicrobiology, where human coprolites and fecal sediments are very rare compared to dental calculi. However, as G2S appears to work best when the input oral microbial composition is close to the average used during training, caution must still be taken in interpreting the prediction data. Furthermore, G2S was implemented using both 16S rRNA gene and shotgun metagenomics data from different populations across the globe (from United States, Italy, Sweden, United Kingdom, and Fiji), with a good generalization of the results as evidenced by the findings on the test dataset. This provides an opportunity for users who can apply the tool on data obtained through different sequencing techniques simply by formatting their abundance tables with a taxonomy congruent with the Greengenes database. It should also be noted that G2S was built and validated using the 768 paired samples currently available in the literature. This stresses the importance of collecting paired samples (i.e., oral and fecal) in future studies from cohorts from different geographic locations, in order to further extend the range of the training dataset and thus the applicability of G2S. Finally, other future implementations could include predictions at different taxonomic levels, as well as functional predictions thanks to the recent expansion of shotgun metagenomics.

In summary, G2S opens up new possibilities in bioinformatics approaches related to metagenomics, extending *in silico* procedures to predict the human stool microbiome from oral microbiome data. Starting from either modern or ancient oral microbiome samples, the tool infers the stool microbiome with family level resolution. Its main field of application is probably paleomicrobiology, as a tool that can help understand how the gut microbiome of the past was structured, and its implications for human evolution. An update of the G2S tool will be periodically performed to incorporate newly released microbiome studies.

DATA AVAILABILITY STATEMENT

The datasets used for setting up G2S are available at the Human Microbiome Project website https://www.hmpdacc.org/HMQCP/ and NCBI SRA as SRP057504 (Zaura et al., 2015), PRJNA217052 (Brito et al., 2016) and PRJNA356414 (Russo et al., 2018). Microbiome data from ancient samples were taken from the study conducted by Warinner and colleagues (Warinner et al., 2014).

AUTHOR CONTRIBUTIONS

SR: conceptualization and software. SR and MF: formal analysis. SR, MC, and ST: writing—original draft preparation. MF, EB, and PB: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.644516/full#supplementary-material

Supplementary File 1 | R script containing the stochastic method that generates mock profiles of the stool microbiome in the range of the training dataset.

Supplementary Table 1 | List of paired fecal and oral samples from the HMP study as well as from other literature studies dealing with healthy adults (Zaura et al., 2015; Brito et al., 2016; Russo et al., 2018). Both 16S rRNA gene sequencing and shotgun metagenomics studies were considered. For each sample, the following data are reported: sample ID, subject ID (and visit when available), geographical origin, reference, sequencing method and body site.

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300

Ayeni, F. A., Biagi, E., Rampelli, S., Fiori, J., Soverini, M., Audu, H. J., et al. (2018). Infant and adult gut microbiome and metabolome in rural Bassa and urban settlers from Nigeria. *Cell Rep.* 23, 3056–3067. doi: 10.1016/j.celrep.2018.05.018

Bajaj, J. S., Betrapally, N. S., Hylemon, P. B., Heuman, D. M., Daita, K., White, M. B., et al. (2015). Salivary microbiota reflects changes in gut microbiota in cirrhosis with hepatic encephalopathy. *Hepatology* 62, 1260–1271. doi: 10.1002/hep.27819

Bishop, C. M. (2016). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., et al. (2016). Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535, 435–439. doi: 10.1038/nature18927

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15:20170387. doi: 10.1098/rsif.2017.0387

Demirci, S., Peters, S. A., de Ridder, D., and Van Dijk, A. D. J. (2018). DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J.* 95, 13979. doi: 10.1111/tpj.13979

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Galkin, F., Mamoshina, P., Aliper, A., Putin, E., Moskalev, V., Gladyshev, V. N., et al. (2020). Human gut microbiome aging clock based on taxonomic profiling and deep learning. *Iscience* 23:101199. doi: 10.1016/j.isci.2020.101199

Geman, O., Chiuchisan, I., Covasa, M., Doloc, C., Milici, M. R., and Milici, L. D. (2016). "Deep learning tools for human microbiome big data," in *Proceedings of the 7th International Workshop Soft Computing Applications SOFA 2016. Advances in Intelligent Systems and Computing*, Vol. 633, eds V. Balas, L. Jain, and M. Balas (Cham: Springer), 265–275.

Glassner, K. L., Abraham, B. P., and Quigley, E. M. M. (2020). The microbiome and inflammatory bowel disease. *J. Allergy Clin. Immunol.* 145, 16–27. doi: 10.1016/j.jaci.2019.11.003

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Griffen, A. L., Beall, C. J., Campbell, J. H., Firestone, N. D., Kumar, P. S., Yang, Z. K., et al. (2012). Distinct and complex bacterial profiles in human periodontitis and health revealed by 16S pyrosequencing. *ISME J.* 6, 1176–1185. doi: 10.1038/ismej.2011.191

He, Y., Wu, W., Zheng, H. M., Li, P., McDonald, D., Sheng, H. F., et al. (2018). Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* 24, 1532–1535. doi: 10.1038/s41591-018-0164-x

Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V., and Wargo, J. A. (2019). The microbiome, cancer, and cancer therapy. *Nat. Med.* 25, 377–388. doi: 10.1038/s41591-019-0377-7

Iwauchi, M., Horigome, A., Ishikawa, K., Mikuni, A., Nakano, M., Xiao, J. Z., et al. (2019). Relationship between oral and gut microbiota in elderly people. *Immun. Inflamm. Dis.* 7, 229–236. doi: 10.1002/iid3.266

Jha, A. R., Davenport, E. R., Gautam, Y., Bhandari, D., Tandukar, S., Ng, K. M., et al. (2018). Gut microbiome transition across a lifestyle gradient in Himalaya. *PLoS Biol.* 16:e2005396. doi: 10.1371/journal.pbio.2005396

Karpiński, P. M. (2019). Role of oral microbiota in cancer development. *Microorganisms* 7, 20. doi: 10.3390/microorganisms7010020

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336. doi: 10.1038/nature10213

Koskella, B., Hall, L. J., and Metcalf, C. J. E. (2017). The microbiome beyond the horizon of ecological and evolutionary theory. *Nat. Ecol. Evol.* 1, 1606–1615. doi: 10.1038/s41559-017-0340-2

Le, N. Q. K. (2019). Fertility-gru: identifying fertility-related proteins by incorporating deep-gated recurrent units and original position-specific scoring matrix profiles. *J. Proteome Res.* 18, 3503–3511. doi: 10.1021/acs.jproteome.9b00411

Le, N. Q. K., and Huynh, T. T. (2019). Identifying SNAREs by incorporating deep learning architecture and amino acid embedding representation. *Front Physiol.* 10:1501. doi: 10.3389/fphys.2019.01501

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., and Yeh, H. Y. (2019). Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext n-grams. *Front Bioeng Biotechnol.* 7:305. doi: 10.3389/fbioe.2019.00305

Leung, M. K. K., Delong, A., Alipanahi, B., and Frey, B. J. (2016). Machine learning in genomic medicine: a review of computational problems and data sets. *Proc. IEEE* 104, 176–197.

Miller, E. T., Svanbäck, R., and Bohannan, B. J. M. (2018). Microbiomes as metacommunities: understanding host-associated microbes through metacommunity ecology. *Trends Ecol. Evol.* 33, 926–935. doi: 10.1016/j.tree.2018.09.002

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., et al. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* 6:6505. doi: 10.1038/ncomms7505

Pietiäinen, M., Liljestrand, J. M., Kopra, E., and Pussinen, P. J. (2018). Mediators between oral dysbiosis and cardiovascular diseases. *Eur. J. Oral Sci.* 126, 26–36. doi: 10.1111/eos.12423

Prodan, A., Levin, E., and Nieuwdorp, M. (2019). Does disease start in the mouth, the gut or both? *Elife* 8:e45931. doi: 10.7554/eLife.45931

Quang, D., and Xie, X. (2019). FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* 166, 40–47. doi: 10.1016/j.ymeth.2019.03.020

Rampelli, S., Guenther, K., Turroni, S., Wolters, M., Veidebaum, T., Kourides, Y., et al. (2018). Pre-obese children's dysbiotic gut microbiome and unhealthy diets may predict the development of obesity. *Commun. Biol.* 1:222. doi: 10.1038/s42003-018-0221-5

Rampelli, S., Schnorr, S. L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., et al. (2015). Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. *Curr. Biol.* 25, 1682–1693. doi: 10.1016/j.cub.2015.04.055

Reiman, D., Metwally, A., and Dai, Y. (2017). Using convolutional neural networks to explore the microbiome. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2017, 4269–4272. doi: 10.1109/EMBC.2017.8037799

Russo, E., Bacci, G., Chiellini, C., Fagorzi, C., Niccolai, E., Taddei, A., et al. (2018). Preliminary comparison of oral and intestinal human microbiota in patients with colorectal cancer: a pilot study. *Front. Microbiol.* 8:2699. doi: 10.3389/fmicb.2017.02699

Schmidt, T. S. B., Hayward, M. R., Coelho, L. P., Li, S. S., Costea, P. I., Voigt, A. Y., et al. (2019). Extensive transmission of microbes along the gastrointestinal tract. *Elife* 8:e42693. doi: 10.7554/eLife.42693

Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., et al. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* 5:3654. doi: 10.1038/ncomms4654

Sonnenburg, E. D., and Sonnenburg, J. L. (2019). The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.* 17, 383–390. doi: 10.1038/s41579-019-0191-8

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Wainberg, M., Merico, D., Delong, A., and Frey, B. J. (2018). Deep learning in biomedicine. *Nat. Biotechnol.* 36, 829–838. doi: 10.1038/nbt.4233

Warinner, C., Rodrigues, J. F., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., et al. (2014). Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* 46, 336–344. doi: 10.1038/ng.2906

Webb, S. (2018). Deep learning for biology. *Nature* 554, 555–557. doi: 10.1038/d41586-018-02174-z

Wong, S. H., and Yu, J. (2019). Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat. Rev. Gastroenterol. Hepatol.* 16, 690–704. doi: 10.1038/s41575-019-0209-8

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Zaura, E., Brandt, B. W., Teixeira de Mattos, M. J., Buijs, M. J., Caspers, M. P. M., Rashid, M. U., et al. (2015). Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. *mBio* 6, e01693–e01695. doi: 10.1128/mBio.01693-15

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 May 2021
doi: 10.3389/fgene.2021.669495

DRAGoM: Classification and Quantification of Noncoding RNA in Metagenomic Data

*Ben Liu[1], Sirisha Thippabhotla[1], Jun Zhang[2,3] and Cuncong Zhong[1,4,5]**

[1] Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS, United States, [2] Division of Medical Oncology, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, United States, [3] Department of Cancer Biology, University of Kansas Medical Center, Kansas City, KS, United States, [4] Bioengineering Program, The University of Kansas, Lawrence, KS, United States, [5] Center for Computational Biology, The University of Kansas, Lawrence, KS, United States

Noncoding RNAs (ncRNAs) play important regulatory and functional roles in microorganisms, such as regulation of gene expression, signaling, protein synthesis, and RNA processing. Hence, their classification and quantification are central tasks toward the understanding of the function of the microbial community. However, the majority of the current metagenomic sequencing technologies generate short reads, which may contain only a partial secondary structure that complicates ncRNA homology detection. Meanwhile, *de novo* assembly of the metagenomic sequencing data remains challenging for complex communities. To tackle these challenges, we developed a novel algorithm called DRAGoM (Detection of RNA using Assembly Graph from Metagenomic data). DRAGoM first constructs a hybrid graph by merging an assembly string graph and an assembly de Bruijn graph. Then, it classifies paths in the hybrid graph and their constituent readsinto differentncRNA families based on both sequence and structural homology. Our benchmark experiments show that DRAGoMcan improve the performance and robustness over traditional approaches on the classification and quantification of a wide class of ncRNA families.

Keywords: metagenomics, noncoding RNA, covariance model, homology search, genome assembly

OPEN ACCESS

Edited by:
Mohammed Kuddus, University of Hail, Saudi Arabia

Reviewed by:
Khurshid Ahmad, Yeungnam University, South Korea
Sudhir P. Singh, Center of Innovative and Applied Bioprocessing (CIAB), India

****Correspondence:***
Cuncong Zhong
cczhong@ku.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *18 February 2021*
Accepted: *23 March 2021*
Published: *05 May 2021*

Citation:
Liu B, Thippabhotla S, Zhang J and Zhong C (2021) DRAGoM: Classification and Quantification of Noncoding RNA in Metagenomic Data. Front. Genet. 12:669495. doi: 10.3389/fgene.2021.669495

INTRODUCTION

Noncoding RNAs (ncRNAs) can perform versatile functional roles and their importance in cellular physiology is being increasingly recognized. For example, riboswitch is a class of cis-regulator that locates in the 5′-UTR of its target gene and can alter the expression efficiency of the target through alternating its fold structure upon the binding with molecules such as small metabolites or ion ligands (Tucker and Breaker, 2005; Garst et al., 2011; Breaker, 2018). A different trans-regulatory mechanism was found to be exerted by bacterial small RNAs (sRNAs), which attenuate (in rare cases promote) their target mRNA expressions through sequence complementarity-based binding (in a similar way as eukaryote microRNAs) (Gottesman and Storz, 2011; Storz et al., 2011; Nitzan et al., 2017; Waters et al., 2017). ncRNAs can also catalyze biochemical reactions (ribozymes) (Doherty and Doudna, 2001), as exemplified by the well-known ribosomal RNAs (which catalyze protein synthesis) and group I and II introns (which catalyze the excision of themselves from the transcript) (Adams et al., 2004a,b). With the prevalence of metagenomics (Virgin and Todd, 2011; Huttenhower et al., 2012; Shokralla et al., 2012; Williamson and Yooseph, 2012;

Davison et al., 2015; Quince et al., 2017), more microbial genomic sequences, including the previously uncharacterized ones, have been accumulated into public databases. The amazing richness of microbial genomic data renders a great opportunity to study ncRNA. Indeed, the diversity and richness of microbial ncRNA function revealed from analyzing metagenomic data are beyond our existing knowledge (Weinberg et al., 2010; Nawrocki and Eddy, 2013a; Tobar-Tosse et al., 2013; Stav et al., 2019), including many long ncRNA classes such as OLE, GOLLD, and HEARO (Harris and Breaker, 2018). The discoveries underpin the importance of ncRNA functions in bacterial physiology, ecology, and interaction with the environment.

Despite the importance of functional ncRNA, reliable classification and quantification of ncRNA elements from metagenomic sequencing dataremain challenging. Because the function of ncRNA is more determined by its structural fold rather than its primary sequence (except few ncRNA classes such as microRNA, Bartel, 2009; Davis and Hata, 2009; Winter et al., 2009), the homology search of ncRNA often relies on both primary sequence and secondary structure conservation (Klein and Eddy, 2003; Zhang et al., 2006). Both types of information of a given ncRNA family can be complied using stochastic context-free grammar (SCFG) into a covariance model (CM) to facilitate family-level homology detection (Eddy and Durbin, 1994), in a similar idea of using the profile hidden Markov model (HMM) for protein family characterization (Sonnhammer et al., 1997). In the context of metagenomic sequencing data, the short reads (∼100–150 bp) may only contain partial secondary structure information, leading to inferior ncRNA homology search performance. The issue has been partially addressed via the development of the truncated Cocke–Younger–Kasami (trCYK) algorithm for parsing reads with an incomplete secondary structure (Kolbe and Eddy, 2009), but its performance remained lower compared to a homology search with a complete secondary structure. On the other hand, while a natural way to resolve this issue is to reconstruct complete secondary structure information via *de novo* genome assembly, the assembly itself remained fragmentary and incomplete for metagenomic data generated from a complex microbial community (Ghurye et al., 2016; Sczyrba et al., 2017; Breitwieser et al., 2019; Olson et al., 2019). Many ncRNA reads, especially the low-abundance ones, may not be assembled into contigs and cannot be detected in the subsequent homology search stage.

To tackle the challenge of ncRNA homology search from metagenomic sequencing data, we have developed DRAGoM (Detection of RNA using Assembly Graph from Metagenomic data). DRAGoM aligns CM against paths in an assembly graph and classifies the paths and their constituent reads into different ncRNA families based on the alignment. Note that a path in an assembly graph corresponds to a set of overlapping reads, which is more likely to contain complete secondary structure information that facilitates homology detection. Hence, we can expect DRAGoM to outperform the strategy of performing a homology search directly on unassembled reads (subsequently referred to as the "read-based" strategy). On the other hand, using the complete set of paths in the assembly graph without topological simplification (e.g., bubble removal and tip trimming, Bankevich et al., 2012; Simpson and Durbin, 2012; Nurk et al., 2017) and traversal (e.g., as Eulerian paths, Pevzner et al., 2001) is more likely to retain the original metagenome information (such as polymorphism and stain-level sequence variation). As a result, DRAGoM is also expected to rescue many ncRNA reads that cannot be assembled into contigs and to outperform the strategy of performing a homology search on assembled contigs (thereafter referred to as the "assembly-based" strategy).

We have benchmarked DRAGoM with a representative of the read-based strategy (i.e., CMSearch, Nawrocki and Eddy, 2013b), which includes the trCYK algorithm (Kolbe and Eddy, 2009) for detecting incomplete secondary structures, and representatives of the assembly-based strategy (i.e., assembling the metagenomic reads using a string graph assembler SGA, Simpson and Durbin, 2012, or a de Bruijn graph assembler SPAdes, Bankevich et al., 2012; Nurk et al., 2017, followed by searching the resulting contigs using CMSearch). Our benchmark experiment has considered both simulated and real datasets and includes16S rRNA and a large collection of CMs for different ncRNA families registered in Rfam (Nawrocki et al., 2015). We show that DRAGoM has a higher performance compared to the read-based or assembly-based method and demonstrates the most robust performance on ncRNA families with different lengths and conservation levels. Thus, DRAGoM will have potential applications in future metagenomic data analyses, as well as in the functional studies of microbial ncRNAs.

MATERIALS AND METHODS

DRAGoM Algorithm

The DRAGoM algorithm contains two main stages: (1) the construction of a hybrid assembly graph and (2) the identification of homologous ncRNA paths and reads from the resulting hybrid assembly graph. By hybrid assembly graph, we mean the assembly graph resulting from merging a string graph (Myers, 2005) and a de Bruijn graph (Idury and Waterman, 1995), the two main computational models used in sequence assembly. A string graph is constructed based on a suffix–prefix overlap between the reads, while a de Bruijn graph is constructed based on the shared *k*-mers between reads. Either of the model has its own advantages and limitations, with the string graph being more accurate but fragmentary. Both models have been integrated to improve sequence assembly (Huang and Liao, 2016). To illustrate the idea, we present a toy example in **Figure 1A**. The top-left panel shows an artificial genome sequence and the corresponding short reads. The bottom-left panel shows the string graph constructed from the reads with a minimum overlap length of 4 bp. Because of the uneven (and lower) coverage at the middle of the artificial genome, only four reads out of six can be overlapped. A missing link (the blue dashed line) exists between the two subgraphs, leading to a subsequent fragmentary assembly. For de Bruijn graph construction shown in the top-right panel, all reads can be connected using 3-mers as the vertices. While the de Bruijn graph completely recovers all reads, its graph topology is complex, and it can be traversed in more than one way (with or without going into the loop). However, note that the sequence of one of the

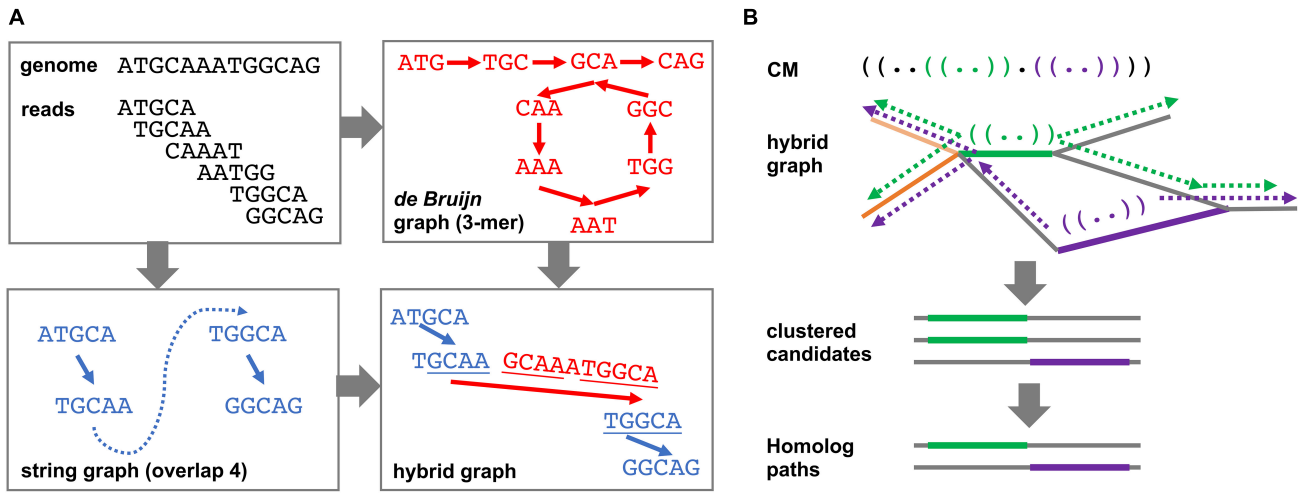

FIGURE 1 | Schematic illustration of the DRAGoM algorithm. **(A)** The construction of the hybrid assembly graph. The hybrid graph, resulting from the merging of a de Bruijn graph and a string graph, can perfectly represent the original genome used in this toy example. **(B)** The search of ncRNA homologs against a hybrid graph. The green and purple parentheses in the querying ncRNACM (covariant model) represent local secondary structural components. The thick green and purple lines in the hybrid graph indicate anchors for path extension. Arrows indicate path extensions of the corresponding anchors.

traversals (i.e., with the loop) can be aligned to some terminal sequences in the string graph (bottom-right panel, underlined sequences), indicating that the corresponding subgraphs can be reconnected using de Bruijn graph paths. The resulting hybrid assembly graph perfectly represents the original genome.

The hybrid graph construction stage of DRAGoM implements the above intuition. Specifically, SGA (version 0.10.15) (Simpson and Durbin, 2012) was used to generate the string graph, and SPAdes (version 3.13.0) (Nurk et al., 2017) was used to generate the de Bruijn graph. When running SPAdes, the "–meta" tag was enabled to indicate metagenomic input (also known as "metaSPAdes"). Both programs were run in the paired-end mode. Detailed command lines for running both assemblers are available from the **Supplementary Methods**. The intermediate output of SGA (i.e., the.asqg file) was further simplified (using in-house scripts) to condense unbranched paths into single edges. Terminal edges (i.e., edges with an in-degree or out-degree of 0) of the resulting string graph were then aligned to the set of verified SPAdes contig sequences (no coverage hole, see more in **Supplementary Methods**) using BWA. Only alignments with a minimum score of 45 (per BWA manual, +1 for a match, −4 for a mismatch, and −6 for a gap), a minimum alignment length of 100, and no clipping at the open end (i.e., the end with a degree of 0 in the string graph) were considered. Then, for each SPAdes contig, if it had recruited more than one alignment, the corresponding terminals in the string graph defined by any pair of alignments were connected using the corresponding interval sequence of the SPAdes contig. If a SPAdes contig had recruited only one alignment, the corresponding string graph terminal was extended using the corresponding prefix or suffix sequence of the contig. SPAdes contigs with no recruited alignment were also retained as isolated vertices in the hybrid graph. In a CAMI (Sczyrba et al., 2017) dataset (DS5 as defined in the **Benchmark Datasets and Metrics** section) that contained ∼15M vertices in its string graph, ∼0.7M connections were made. DRAGoM allows the output of the hybrid graph as its intermediate result, which can be traversed by other assemblers for metagenomere construction.

The second main stage of the DRAGoM algorithm is to identify homologous paths and reads with respect to a given querying CM from the resulting hybrid assembly graph. Intuitively, one can exhaustively enumerate all paths of the hybrid graph and align them against the querying CM. However, this naïve approach would be practically infeasible because the number of paths grows exponentially with the number of reads in the dataset. To address this issue, we designed a filter-based heuristic for the speedup (**Figure 1B**). To begin with, the querying CM was aligned to each edge of the hybrid graph (note that an edge corresponds to a condensed path without branching, or unitig). The edges bearing significant similarity to the querying CM were recorded as anchors. This stage allowed the detection of conserved short structural components (e.g., the green and purple stem-loops in the CM and the bolded paths in the hybrid graph of **Figure 1B**). The anchors were then extended toward both directions, aiming to reconstruct complete sequences of the candidate ncRNA homologs (the broken arrows in **Figure 1B**). The extension lengths for each anchor were determined by length of the unaligned prefix and suffix of the CM (with a further extension of 10% of the prefix or suffix length to account for potential gaps). Because some edges of the hybrid graph might represent similar sequences (e.g., the heavy and light orange edges in **Figure 1B**), all paths resulting from extending the anchors were subject to sequence redundancy removal using CD-Hit (Li and Godzik, 2006). Finally, the set of nonredundant paths were realigned to the querying CM, and the paths passing the gathering score threshold were selected as homologs of the corresponding ncRNA family. Note that the homologous paths are only being used as templates for classifying

individual reads but should not be taken as individual ncRNA genes. This is because many of the homologous paths are derived from the exhaustive traversal of all paths of the graph and could be chimeric and redundant (see more in section **Discussion**). Finally, individual reads were further mapped to the homologous paths for their annotation and to quantify the corresponding ncRNA family in the datasets. More details regarding this stage can be found in **Supplementary Methods**.

The above algorithm was implemented as the DRAGoM software package. DRAGoM accepts a set of querying CM and a given metagenomic sequencing dataset and assigns a subset of the reads to the corresponding ncRNA families. DRAGoM was implemented using GNU C++ and Python and has been tested under several major Linux distributions (RedHat, Fedora, and Ubuntu). It is freely available under the Creative Commons BY-NC-ND 4.0 License Agreement[1].

Benchmark Datasets

We constructed five datasets to benchmark the performance of DRAGoM, as summarized in **Table 1**. Two datasets were simulated in-house, one was generated by an independent research group for a similar benchmark purpose (Yuan et al., 2015), one was from the open metagenomic data analysis challenge CAMI (Sczyrba et al., 2017), and the last one was from a real human gut microbiome (SRR341583). Detailed information regarding the reference genomes included their respective relative abundances, and the *in silico* simulation parametersare available from **Supplementary Table 1**. All datasets are also available for download from https://cbb.ittc.ku.edu/DRAGoM.html. These five benchmark datasets include the following:

- DS1 (the REAGO dataset): This simulated dataset represented a low-diversity metagenomic dataset that contains microbes from different clades with staggered abundances. The dataset was used in the benchmark experiment of REAGO (Yuan et al., 2015). It was simulated *in silico* with an average read length of 100nt and an expected error rate of 1%, containing 4,653,918 paired-end reads.
- DS2 (the streptococci dataset): This simulated dataset represented a community with highly related microbial genomes from the same genus (e.g., streptococcus). The dataset was simulated *in silico* using eight streptococcus genomes, with an average read length of 100nt and an expected error rate of 1%. This dataset contained 600,000 paired-end reads.
- DS3 (the marine dataset): This dataset represented a subset of microbial metagenome that was often observed from the marine environment. It was simulated from 28 marine genomes with an average read length of 100nt and an expected error rate of 1% and contained 3,700,000 paired-end reads.
- DS4 (the subsampled gut dataset): This dataset represented a real human gut microbiome community (SRR341583). To facilitate the generation of ground-truth homology for the benchmark purpose, we subsampled the dataset via read mapping to a collection of microbial genomes often found in the human gut environment. Only reads that were mapped to the selected reference genomes were retained, leaving 11,228,362 paired-end reads for this dataset.
- DS5 (the subsampled CAMI dataset): This dataset was downloaded from CAMI (Sczyrba et al., 2017), a comprehensive simulated dataset. To focus on the more challenging cases of metagenomics analysis, only reads representing low-coverage genomes (<10X) were selected (via read mapping). This dataset contained 31,311,294 paired-end reads.

[1]https://creativecommons.org/licenses/by-nc-nd/4.0/

TABLE 1 | Summary of experimental datasets.

Dataset	Description	No. of genomes	Abundance	No. of reads	Read length	Error rate
DS1	REAGO	14	Staggered	4.6M	100	1%
DS2	Streptococcus	8	5x	0.6M	100	1%
DS3	Marine	28	5x	3.7M	100	1%
DS4	Human gut	3,499	Staggered	11.2M	74	–
DS5	CAMI	4,679	Staggered	31.3M	100	–

Benchmark Experiment Setup

Given a querying ncRNA family, we define its ground-truth homologs as the reads that were generated or mapped (>60% of their total lengths) to the genomic intervals that were annotated as the ncRNA family by CMSearch (Nawrocki and Eddy, 2013b) (under its default stringency cutoff). The command lines used for ground-truth generation are available from the **Supplementary Methods**.

Given the ground-truth definition, we defined true positives (TPs) as the homologous reads that were identified by a given method. We defined false positives (FPs) as the nonhomologous reads that were identified and false negatives (FN) as the homologous reads that were not identified. We further defined recall and precision as

$$\text{Recall} = \frac{TP}{TP + FN}, \ \text{Precision} = \frac{TP}{TP + FP}$$

and subsequently F-score as

$$F - \text{score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

All methods were tested under a series of different stringency cutoffs to generate the receiver operating characteristic (ROC) curve. The ROC curves were extrapolated to the points (recall: 0, precision: 1) and (recall: 1, precision: 0) to calculate the area under the curve (AUC).

We benchmarked our graph-based ncRNA homology search strategy DRAGoM (homology search against assembly graph) with the read-based strategy (homology search against unassembled reads) and the assembly-based strategy (homology search against assembled contigs). For read-based strategy, we chose CMSearch as the representative and refer to it as "CMSearch" thereafter. For assembly-based strategy, we chose SGA (as the representative of string graph assemblers) and SPAdes (as the representative of de Bruijn graph assemblers) and

refer to them as "SGA+CMSearch" and "SPAdes+CMSearch," respectively. Command lines for executing the programs are available in the **Supplementary Methods**. Each method was benchmarked using different sets of querying ncRNA families (determined based on their presence in the selected reference genomes, details available from **Supplementary Table 2**). The reported performance corresponded to the unweighted arithmetic mean performance among the sets of querying ncRNA families. Note that the search performances for 16S rRNA were reported individually, given its importance in metagenome taxonomic profiling.

RESULTS

The performances of all tested methods on DS1 (the REAGO dataset, 42 ncRNA families searched) are shown in **Figure 2**. For non-16S rRNA queries (**Figure 2A**), DRAGoM was able to achieve the highest recall, representing a gain of 7.3% recall rate as compared to the second-best performer SPAdes+CMSearch (**Table 2**). CMSearch alone performed significantly worse than DRAGoM and SPAdes+CMSearch, potentially due to the lack of complete secondary structure information in unassembled reads. SGA+CMSearch seemed to be adversely impacted by the low coverage of this dataset and showed the lowest recall but also showed the highest precision rate. The observation was in line with our current understanding of the characteristics of the string graph and de Bruijn graph assembly approaches. In terms of the peak *F*-score, DRAGoM achieved 93.6%, followed by SPAdes+CMSearch with 92.2%. In terms of AUC, DRAGoM was also the best performer with 96.8%, as compared to 93.9% of the second-best method SPAdes+CMSearch. For 16S rRNA, all methods performed well (**Figure 2B**). DRAGoM remained the best method with a marginal improvement (99.5% *F*-score and 99.6% AUC, followed by 97.6% *F*-score and 98.8% AUC of the second-best method CMSearch, see **Table 3**). Surprisingly, SPAdes+CMSearch showed the lowest sensitivity, potentially due to the polymorphism information lost during the graph simplification and traversal stages of SPAdes. Overall, DRAGoM showed a higher performance than any tested method and was robust for both non-16S and 16S rRNA searches.

For DS2 (the streptococcus dataset, 27 ncRNA families searched), the performance of the methods on non-16S rRNAs was similar to that of DS1 (**Figure 3A**). DRAGoM again performed the best on this dataset (91.4% *F*-score and 93.0% AUC), followed by SPAdes+CMSearch (90.2% *F*-score and 90.7% AUC, see **Table 2**). The lower performances of CMSearch and SPAdes+CMSearch were also observed as in DS1 and may be due to similar reasons as discussed previously. For 16S rRNA (**Figure 3B**), SGA+CMSearch performed the best (99.2% *F*-score and 99.8% AUC), with DRAGoM as the second-best method in *F*-score (98.1%) and CMSearch in AUC (99.4%, see **Table 3**). The performance of SGA seemed to benefit from its preservation of polymorphism information in 16S rRNA via a more conservative graph simplification strategy. On the other hand, DRAGoM remained the most sensitive method (with the highest recall rate of 99.9%), but its overall performance appeared to be compromised by the lower precision rate due to exhaustive path traversal (96.2%, see **Table 3**).

For DS3 (simulated marine, 93 ncRNA families searched; see **Figure 4**) and DS4 (subsampled human gut, 60 ncRNA families searched; see **Figure 5**), the performance of the methods also followed the same trend as that observed in DS1 and DS2. DRAGoM outperformed the other methods in non-16S rRNA queries (for DS3 shown in **Figure 4A**, DRAGoM had 89.9% *F*-score and 94.9% AUC; for DS4 shown in **Figure 5A**, it had 74.4% *F*-score and 77.4% AUC). Note that the lower performance

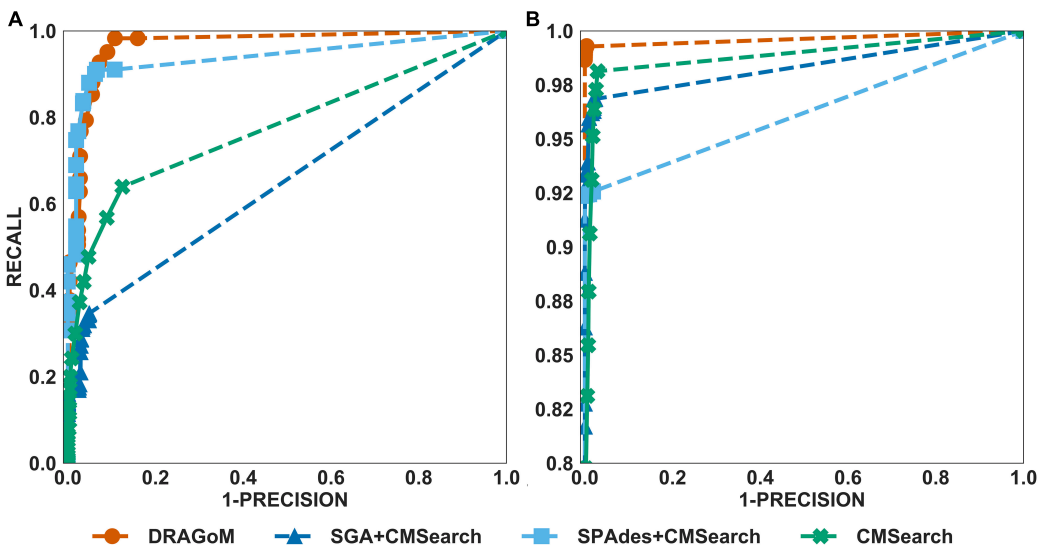

FIGURE 2 | The ROC curves for searching **(A)** 42 non-16S rRNAncRNA families and **(B)**16S rRNAs using the corresponding programs against DS1 (the REAGO dataset).

TABLE 2 | Performance summary of the tested methods on DS1–DS4 (for non-16S rRNA queries).

Dataset	Matrices	DRAGoM	SGA+CMSearch	SPAdes+CMSearch	CMSearch
DS1	Precision	89.2%	**95.1%**	93.4%	87.6%
	Recall	**98.3%**	34.7%	91.0%	63.9%
	F1	**93.6%**	50.8%	92.2%	73.9%
	AUC	**96.8%**	65.2%	93.9%	77.7%
DS2	Precision	88.7%	**94.0%**	92.0%	91.9%
	Recall	**94.3%**	8.4%	88.4%	49.4%
	F1	**91.4%**	15.5%	90.2%	64.2%
	AUC	**93.0%**	51.3%	90.7%	70.0%
DS3	Precision	87.4%	**93.6%**	91.7%	87.5%
	Recall	**92.4%**	4.6%	87.6%	55.7%
	F1	**89.9%**	8.7%	89.6%	68.0%
	AUC	**94.9%**	49.2%	92.9%	72.9%
DS4	Precision	86.4%	**95.7%**	85.8%	77.6%
	Recall	**65.2%**	23.5%	58.4%	36.9%
	F1	**74.4%**	37.8%	69.5%	50.0%
	AUC	**77.4%**	60.1%	73.3%	58.2%

The highest performance of each category is bolded.

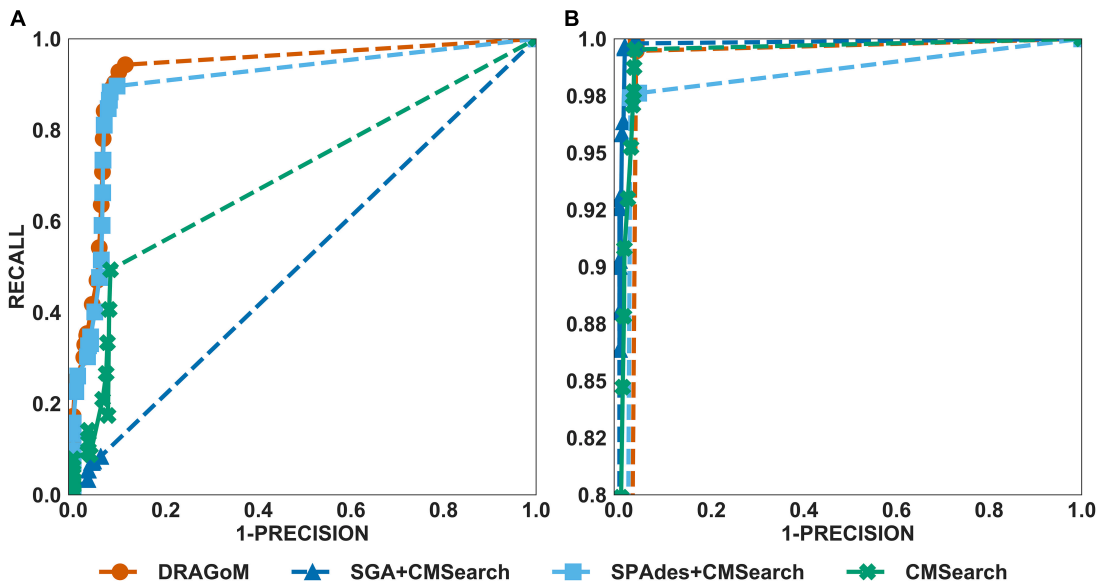

FIGURE 3 | The ROC curves for searching **(A)** 27 non-16S rRNAncRNA families and **(B)**16S rRNAs using the corresponding programs against DS2 (the simulated streptococcus dataset).

on DS4 for all methods was due to the fact that DS4 was generated by subsampling a real dataset, which contains more experimental noises than the simulated ones. SPAdes+CMSearch also remained as the second-best method on both DS3 and DS4. For 16S rRNA, DRAGoM performed the best on DS3 (99.1% *F*-score and 99.3% AUC; see **Figure 4B** and **Table 3**). On DS4, SGA+CMSearch performed the best (96.1% *F*-score and 96.4% AUC; see **Figure 5B** and **Table 3**), followed by DRAGoM (94.2% *F*-score and 94.4% AUC). These observations were also consistent with those made in DS1 and DS2.

DS5 (subsampled CAMI) was tested using the largest number of querying ncRNA families (276); hence, we categorize the performance of non-16S rRNA searches based on the ncRNA families' sequence identity and average length (**Figure 6** and **Table 4**). Although the performances differed in different categories of ncRNA families, DRAGoM consistently showed the best performance in all categories. The lowest performance gain made by DRAGoM was for the category with<50% sequence identity and 200–400 bp length, where the improvement was 0.6% in *F*-score and 2.4% in AUC compared to the second-best method SPAdes+CMSearch (**Figure 6B**). The largest gain made by DRAGoM was found in the category with 50–70% sequence identity and 200–400 bp length. Interestingly, the improvement was 11.4% in *F*-score (as compared to SPAdes+CMSearch) and

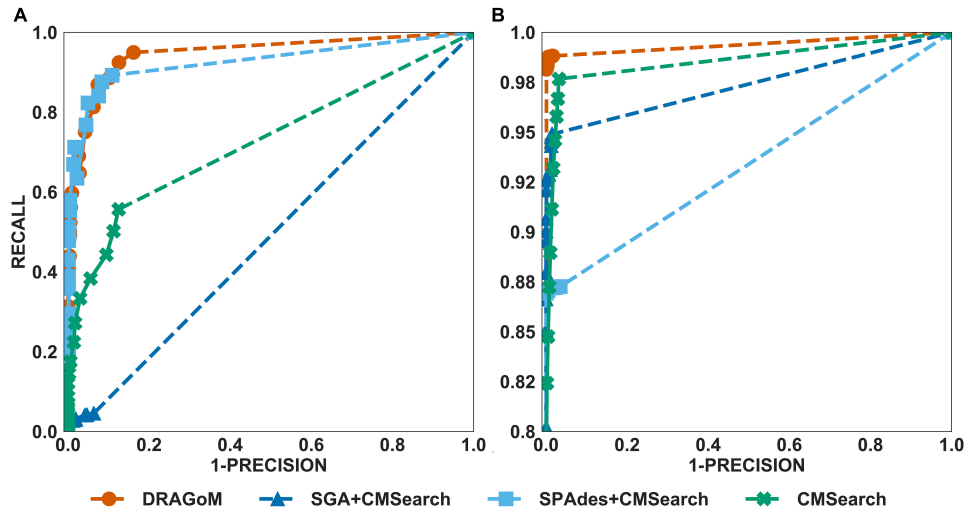

FIGURE 4 | The ROC curves for searching **(A)** 93 non-16S rRNAncRNA families and **(B)**16S rRNAs using the corresponding programs against DS3 (the simulated marine dataset).

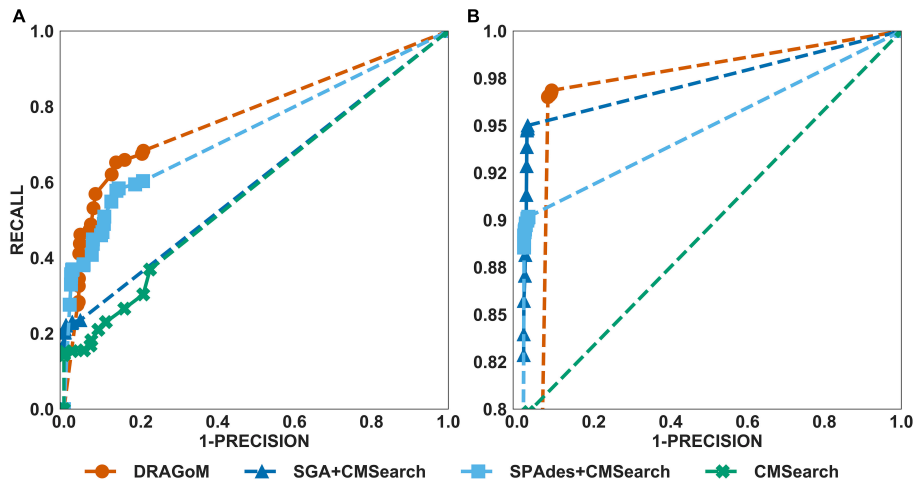

FIGURE 5 | The ROC curves for searching **(A)** 60 non-16S rRNAncRNA families and **(B)**16S rRNAs using the corresponding programs against DS4 (the subsampled human gut dataset).

10.1% in AUC (as compared to CMSearch). Our interpretations for the difference in performance gain in different categories of ncRNA families are present in the **Discussion** section. For 16S rRNA, DRAGoM had the best performance in *F*-score (96.4%, **Table 3**) but the second-best performance in AUC (96.8%, compared to the best performance of 97.6% made by CMSearch).

Taken together, DRAGoM consistently delivered superior search performance in nearly all datasets and all categories of querying ncRNA families. Specifically, DRAGoM produced the best ncRNA homology prediction for all non-16S rRNA in all datasets and two out five datasets (DS1 and DS3) for 16S rRNA searches (DRAGoM was the second-best method for the other three cases). The assembly-based approach SPAdes+CMSearch seemed to be the second-best choice overall. However, the read-based approach CMSearch appeared to be the second-best choice when analyzing ncRNA families with sequence identity between 70 and 90% and length between 200 and 400 bp (**Figure 6F**) and in the searches of 16S rRNAs on DS1, DS3, and DS5. Comparably, DRAGoM was the most robust method in addition to its superior performance.

DISCUSSION

We have demonstrated using benchmark data that DRAGoM can improve ncRNA homology search as compared to the traditional read-based and assembly-based strategies. In addition to the higher performance, another unique advantage of DRAGoM is its robustness. We observed from the benchmark results that the homology search performance

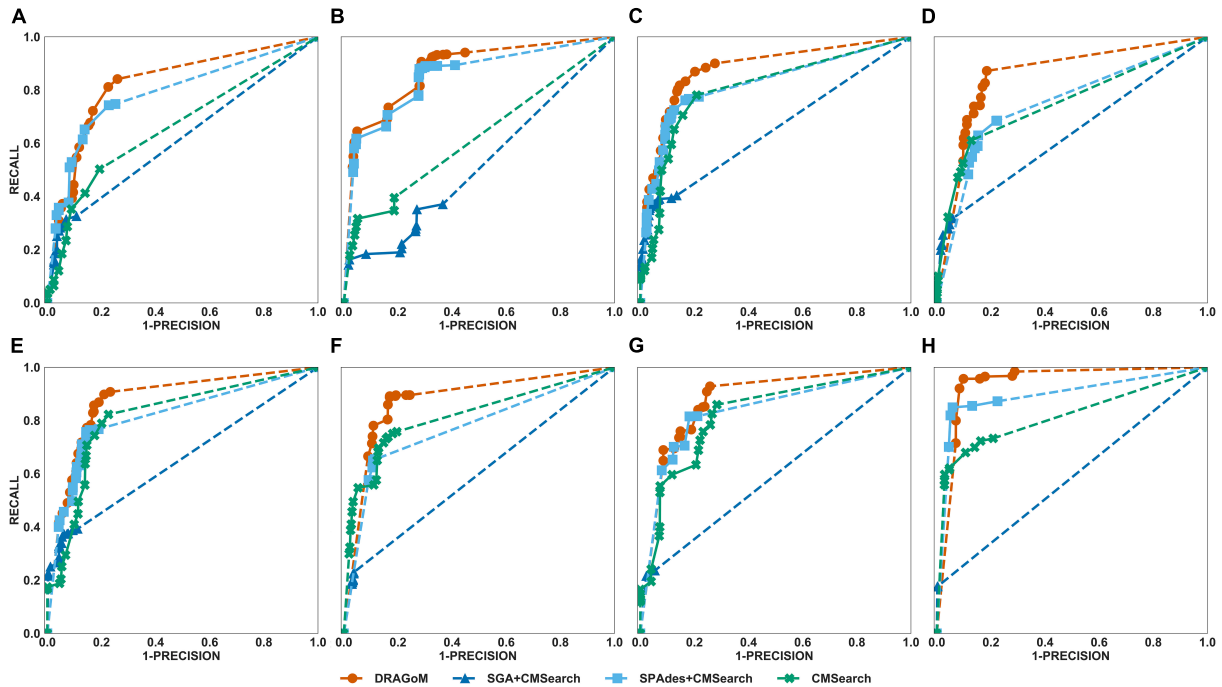

FIGURE 6 | The ROC curves for searching **(A)** 29 ncRNA families with a sequence identity of less than 50% and sequence length from 100 to 200nt, **(B)** eightnc RNA families with a sequence identity of less than 50% and sequence length from 200 to 400nt, **(C)** 78 ncRNA families with a sequence identity from 50 to 70% and sequence length from 100 to 200nt, **(D)** 19 ncRNA families with a sequence identity from 50 to 70% and sequence length from 200 to 400nt, **(E)** 86 ncRNA families with a sequence identity from 70 to 90% and sequence length from 100 to 200nt, **(F)** 19 ncRNA families with a sequence identity from 70 to 90% and sequence length from 200 to 400nt, **(G)** 25 ncRNA families with a sequence identity of more than 90% and sequence length from 100 to 200nt, and **(H)** 12 ncRNA families with a sequence identity of more than 90% and sequence length from 200 to 400nt using the corresponding programs against DS5 (the subsampled CAMI dataset).

is both querydependent and datasetdependent. For example, in DS5 (CAMI), SPAdes+CMSearch performed better than CMSearch when searching ncRNA families with an identity of <50% and between 100 and 200 bp long (**Figure 6A**) but performed worse than CMSearch for ncRNA families with an identity of 70–90% with the same length range (**Figure 6E**). We conjecture that some factors could contribute to such a difference. If the ncRNA families are highly divergent, sequence information alone may not be sufficient for its detection, and therefore, the complete secondary structure information needs to be reconstructed for its detection (shown by the higher performance of assembly-based methods for low-identity ncRNA families). On the other hand, for highly conserved families, their corresponding reads could be treated as repeats, with a significant amount of polymorphism information lost (for the lower performance of assembly-based methods for high-identity ncRNA families). Meanwhile, the performance of the existing methods also differs in searching the same ncRNA family against different datasets, as shown by the higher performance of CMSearch (as compared to SGA+CMSearch) in the 16S rRNA search against DS3 (**Figure 4B**) and its lower performance in the 16S rRNA search against DS4 (**Figure 5B**). The performance difference could be due to assembly quality. Datasets from less diverse community and sequenced with higher coverage are easier to assemble, leading to the higher performance of assembly-based methods. Given the above observation, the ideal case is that we choose an appropriate analysis strategy based on the query and the dataset. However, it is in many cases infeasible. The robustness of DRAGoM makes it an ideal solution to this issue, allowing consistent biological information to be extracted for diverse research objectives and from heterogenousmetagenomic datasets.

Because DRAGoM directly operates on the assembly graph, the quality of the assembly graph will likely affect the performance of DRAGoM. Currently, the string graph and de Bruijn graph dominate the modeling of sequence overlap information in *de novo* assembly. DRAGoM, which is based on the combination of the two graphical models, outperformed the use of either of them alone (i.e., SGA+CMSearch and SPAdes+CMSearch). The observation is consistent with our current understanding of the two models, where each of them has its unique advantages (where the string graph accurately represents the intact information and the de Bruijn graph generates more complete and longer assembly). We further observed that in most cases, SPAdes+CMSearch outperformed SGA+CMSearch in most cases, suggesting that the reconstruction of a complete secondary structure (facilitated by the longer assembly of SPAdes) is more important than the preservation of polymorphism information (as retained in the string graph). Of course, the conclusion is merely for generic cases, as we did observe examples where SGA+CMSearch outperformed SPAdes+CMSearch (e.g., **Figure 5B**).

TABLE 3 | Performance summary of the tested methods on DS1–DS5 (for 16S rRNA queries).

Dataset	Matrices	DRAGoM	SGA+ CMSearch	SPAdes+ CMSearch	CMSearch
DS1	Precision	99.9%	99.0%	**100.0%**	97.1%
	Recall	**99.2%**	96.0%	92.4%	98.1%
	F1	**99.5%**	97.5%	96.0%	97.6%
	AUC	**99.6%**	98.4%	96.2%	98.8%
DS2	Precision	96.2%	**98.7%**	97.5%	96.5%
	Recall	**99.9%**	99.7%	97.4%	99.5%
	F1	98.1%	**99.2%**	97.5%	98.0%
	AUC	98.1%	**99.8%**	97.5%	99.4%
DS3	Precision	99.7%	98.6%	**100.0%**	96.8%
	Recall	**98.6%**	94.9%	87.0%	97.7%
	F1	**99.1%**	96.7%	93.0%	97.2%
	AUC	**99.3%**	97.4%	93.4%	98.6%
DS4	Precision	91.9%	97.2%	**97.7%**	97.0%
	Recall	**96.5%**	95.0%	89.8%	79.8%
	F1	94.2%	**96.1%**	93.6%	87.6%
	AUC	94.4%	**96.4%**	94.1%	88.3%
DS5	Precision	**95.1%**	94.7%	94.6%	94.2%
	Recall	**97.7%**	96.9%	92.2%	97.6%
	F1	**96.4%**	95.8%	93.4%	95.9%
	AUC	96.8%	96.6%	94.1%	**97.6%**

The highest performance of each category is bolded.

We expect to further improve the speed of DRAGoM in the future. Specifically, the efficiency bottleneck of DRAGoM comes from the fact that it needs to exhaustively align the querying CM with all paths generated from anchors. We envision two potential ways to improve the efficiency, i.e., via more intelligent path filtering criteria and graph simplification techniques. We plan to incorporate additional information, such as the GC content, coverage, and covariant mutation compatibility, to filter out paths that are unlikely to be from the same genome before CM alignment. We also expect to reduce the complexity of the assembly graph through incorporating additional information, such as paired end, long read, or Hi-C data, if applicable (Ghurye and Pop, 2019). In general, we observed that DRAGoM was slower when searching long ncRNA families, because the time for CM alignment and the number of candidate paths to align both grow with the length. As a result, for a long querying ncRNA family, we plan to break it down into a set of smaller components by temporarily removing long-range interactions, aligning each small component individually, and checking if the removed long-range interactions can be recovered given the alignments. This heuristic has been proven effective in speeding up the alignment of RNA structural motifs with pseudo knots while retaining satisfying alignment quality (Zhong et al., 2010). We believe the running time of DRAGoM can be significantly reduced with the above optimization techniques.

In addition to the ncRNA family abundance profile, DRAGoM may also be used to improve taxonomic analysis of metagenomic datasets in two ways. First, DRAGoM can improve the traditional 16S rRNA-based taxonomic analysis. The existing methods for this purpose first identify a set of 16S rRNA-related reads from the metagenomic datasets using read-based homology search, perform local assembly on the identified reads, and then infer the taxonomy (Yuan et al., 2015). DRAGoM can improve this strategy in the 16S rRNA homology search step, as it has demonstrated advantage over the traditional read-based homology search approaches. A more accurate and comprehensive set of 16S rRNA reads to start with before assembly will likely lead to a more complete and finer-grained view of the taxonomic profile, as well as potential insight into the previously unidentified species. A second potential way that DRAGoM can improve taxonomic analysis is through facilitating the use of ncRNA families as taxonomic biomarkers, in a similar way as the protein taxonomic biomarkers (Brocchieri, 2001; Wu and Scott, 2012; Klingenberg et al., 2013). However, we note that in the current implementation, DRAGoM only outputs unassembled homologous reads rather than the assembled ncRNA gene sequences. The reason is that many homologous paths arisen from branchy regions of the assembly graph appear to be artificial and redundant. We plan to incorporate a more sophisticated algorithm into DRAGoM to untangle the homologous paths and to output assembled ncRNA gene sequences, via either finding the minimum set of paths that covers the entire homolog-read assembly graph (as in REAGO, Yuan et al., 2015; and Xander, Wang et al., 2015) or using statistical inference methods that find the most probable subset of paths that explain the observed abundances for each edge (as isoform abundance inference for RNA-seq data, Pertea et al., 2016). We believe that by integrating both protein and ncRNA taxonomic biomarkers, we will be able to obtain unbiased and comprehensive taxonomic profiles.

The current version of DRAGoM only included CMSearch as its core homology search engine, requiring only family-level CMs as queryrather than specific ncRNA sequences. The design is due to the lack of complete reference genomes and concrete gene sequences in many metagenomic studies (Kyrpides et al., 2014). In the future, we plan to further extend DRAGoM to allow for single-sequence ncRNAs as query through providing interfaces for other ncRNA homology search tools. Specifically, we will provide interfaces for RSEARCH (Klein and Eddy, 2003) and FastR (Zhang et al., 2005) if both the ncRNA sequence and secondary structure are available. We will provide interfaces for Dynalign (Mathews and Turner, 2002), FoldAlign (Havgaard et al., 2005), PMcomp (Hofacker et al., 2004), LocARNA (Will et al., 2007), and SPARSE (Will et al., 2015) when only the ncRNA sequence is available. These tools implement variants of the simultaneous alignment and folding (SAF) algorithm (Sankoff, 1985) and do not require an annotated secondary structure for the query. We expect that the incorporation of these software into DRAGoM's framework will improve the performances by themselves, as DRAGoM provides the hybrid assembly graph and longer candidate paths to characterize the features of different ncRNA genes.

In summary, in this article, we present DRAGoM, a novel algorithm for family-based ncRNA homology search against

TABLE 4 | Performance summary of the tested methods on DS5 (for non-16S rRNA queries).

Dataset	Matrices	DRAGoM	SGA+CMSearch	SPAdes+CMSearch	CMSearch
Identity:<50%, length: 100–200	Precision	77.5%	**89.2%**	77.3%	80.7%
	Recall	**81.2%**	32.6%	74.3%	50.3%
	F1	**79.3%**	47.8%	75.8%	61.9%
	AUC	**82.4%**	61.7%	78.9%	66.4%
Identity: < 50%, length: 200–400	Precision	71.3%	73.0%	72.1%	**81.4%**
	Recall	**90.7%**	35.2%	87.9%	39.5%
	F1	**79.8%**	47.5%	79.2%	53.2%
	AUC	**87.6%**	52.0%	85.2%	62.2%
Identity: 50–70%, length: 100–200	Precision	85.4%	**86.6%**	83.4%	79.2%
	Recall	**81.8%**	40.4%	76.1%	78.1%
	F1	**83.6%**	55.1%	79.6%	78.6%
	AUC	**87.8%**	65.4%	82.3%	80.7%
Identity: 50–70%, length: 200–400	Precision	81.5%	**94.1%**	78.0%	87.3%
	Recall	**87.2%**	31.8%	68.4%	61.1%
	F1	**84.3%**	47.5%	72.9%	71.9%
	AUC	**85.3%**	63.4%	75.0%	75.2%
Identity: 70–90%, length: 100–200	Precision	82.8%	**88.8%**	85.7%	77.4%
	Recall	**85.7%**	39.3%	75.7%	82.3%
	F1	**84.2%**	54.5%	80.4%	79.8%
	AUC	**87.5%**	65.5%	81.6%	81.6%
Identity: 70–90%, length: 200–400	Precision	83.0%	**96.3%**	88.5%	83.6%
	Recall	**89.1%**	23.0%	65.5%	73.8%
	F1	**86.0%**	37.1%	75.3%	78.4%
	AUC	**87.5%**	59.6%	77.3%	81.7%
Identity:>90%, length: 100–200	Precision	74.2%	**94.6%**	81.8%	71.7%
	Recall	**92.8%**	23.6%	81.5%	85.9%
	F1	**82.5%**	37.8%	81.7%	78.2%
	AUC	**87.7%**	59.4%	83.5%	82.3%
Identity:>90%, length: 200–400	Precision	90.1%	**99.7%**	94.2%	83.8%
	Recall	**95.7%**	17.7%	84.9%	72.3%
	F1	**92.8%**	30.0%	89.3%	77.6%
	AUC	**93.9%**	58.7%	89.5%	81.6%

The highest performance of each category is bolded.

metagenomic sequencing data. We have demonstrated the advantages of DRAGoM as compared to the traditional read-based and assembly-based approaches.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CZ initially conceived the project. CZ and BL designed the algorithm. BL and ST implemented the algorithm. BL performed the benchmark experiments. CZ, BL, and JZ analyzed the results. All authors wrote the manuscript.

FUNDING

This work is funded by the National Science Foundation EPSCoR First Awards in Microbiome Research and the National Science Foundation CAREER award DBI-1943291.

ACKNOWLEDGMENTS

The authors would like to thank Mr. HaoXuan and Mr. Adam Podgorny for their contributions in software testing and manuscript draft proofreading.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.669495/full#supplementary-material

REFERENCES

Adams, P. L., Stahley, M. R., Gill, M. L., Kosek, A. B., Wang, J., and Strobel, S. A. (2004a). Crystal structure of a group I intron splicing intermediate. *RNA*10, 1867–1887. doi: 10.1261/rna.7140504

Adams, P. L., Stahley, M. R., Kosek, A. B., Wang, J., and Strobel, S. A. (2004b). Crystal structure of a self-splicing group I intron with both exons. *Nature*430, 45–50. doi: 10.1038/nature02642

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*19, 455–477. doi: 10.1089/cmb.2012.0021

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*136, 215–233. doi: 10.1016/j.cell.2009.01.002

Breaker, R. R. (2018). Riboswitches and translation control. *Cold Spring Harb. Perspect. Biol.*10:a032797. doi: 10.1101/cshperspect.a032797

Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.*20, 1125–1136. doi: 10.1093/bib/bbx120

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: review and critique. *Theor. Popul. Biol.*59, 27–40. doi: 10.1006/tpbi.2000.1485

Davis, B. N., and Hata, A. (2009). Regulation of MicroRNA biogenesis: a miRiad of mechanisms. *Cell Commun. Signal.*7:18.

Davison, M., Hall, E., Zare, R., and Bhaya, D. (2015). Challenges of metagenomics and single-cell genomics approaches for exploring cyanobacterial diversity. *Photosynth. Res.*126, 135–146. doi: 10.1007/s11120-014-0066-9

Doherty, E. A., and Doudna, J. A. (2001). Ribozyme structures and mechanisms. *Annu. Rev. Biophys. Biomol. Struct.*30, 457–475. doi: 10.1146/annurev.biophys.30.1.457

Eddy, S. R., and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Res.*22, 2079–2088. doi: 10.1093/nar/22.11.2079

Garst, A. D., Edwards, A. L., and Batey, R. T. (2011). Riboswitches: structures and mechanisms. *Cold Spring Harb. Perspect. Biol.*3:a003533. doi: 10.1101/cshperspect.a003533

Ghurye, J. S., Cepeda-Espinoza, V., and Pop, M. (2016). Focus: microbiome: metagenomic assembly: overview, challenges and applications. *Yale J. Biol. Med.*89:353.

Ghurye, J., and Pop, M. (2019). Modern technologies and algorithms for scaffolding assembled genomes. *PLoS Comput. Biol.*15:e1006994. doi: 10.1371/journal.pcbi.1006994

Gottesman, S., and Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.*3:a003798. doi: 10.1101/cshperspect.a003798

Harris, K. A., and Breaker, R. R. (2018). "Large noncoding RNAs in bacteria," in *Regulating with RNA in Bacteria and Archaea*, eds G.Storz and K.Papenfort (Hoboken, NJ: John Wiley & Sons), 515–526. doi: 10.1128/microbiolspec.rwr-0005-2017

Havgaard, J. H., Lyngsø, R. B., Stormo, G. D., and Gorodkin, J. (2005). Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*21, 1815–1824. doi: 10.1093/bioinformatics/bti279

Hofacker, I. L., Bernhart, S. H., and Stadler, P. F. (2004). Alignment of RNA base pairing probability matrices. *Bioinformatics*20, 2222–2227. doi: 10.1093/bioinformatics/bth229

Huang, Y.-T., and Liao, C.-F. (2016). Integration of string and de Bruijn graphs for genome assembly. *Bioinformatics*32, 1301–1307. doi: 10.1093/bioinformatics/btw011

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*486:207. doi: 10.1038/nature11234

Idury, R. M., and Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *J. Comput. Biol.*2, 291–306. doi: 10.1089/cmb.1995.2.291

Klein, R. J., and Eddy, S. R. (2003). RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinform.*4:44. doi: 10.1186/1471-2105-4-44

Klingenberg, H., Aßhauer, K. P., Lingner, T., and Meinicke, P. (2013). Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics*29, 973–980. doi: 10.1093/bioinformatics/btt077

Kolbe, D. L., and Eddy, S. R. (2009). Local RNA structure alignment with incomplete sequence. *Bioinformatics*25, 1236–1243. doi: 10.1093/bioinformatics/btp154

Kyrpides, N. C., Hugenholtz, P., Eisen, J. A., Woyke, T., Goker, M., Parker, C. T., et al. (2014). Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol.*12:e1001920. doi: 10.1371/journal.pbio.1001920

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Mathews, D. H., and Turner, D. H. (2002). Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*317, 191–203. doi: 10.1006/jmbi.2001.5351

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*21, ii79–ii85.

Nawrocki, E. P., and Eddy, S. R. (2013a). Computational identification of functional RNA homologs in metagenomic data. *RNA Biol.*10, 1170–1179. doi: 10.4161/rna.25038

Nawrocki, E. P., and Eddy, S. R. (2013b). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., et al. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*43, D130–D137.

Nitzan, M., Rehani, R., and Margalit, H. (2017). Integration of bacterial small RNAs in regulatory networks. *Annu. Rev. Biophys.*46, 131–148. doi: 10.1146/annurev-biophys-070816-034058

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*27, 824–834. doi: 10.1101/gr.213959.116

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., et al. (2019). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.*20, 1140–1150. doi: 10.1093/bib/bbx098

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.*11:1650. doi: 10.1038/nprot.2016.095

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci.*98, 9748–9753. doi: 10.1073/pnas.171285098

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.*35, 833–844. doi: 10.1038/nbt.3935

Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*45, 810–825. doi: 10.1137/0145048

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., et al. (2017). Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat. Methods*14, 1063–1071.

Shokralla, S., Spall, J. L., Gibson, J. F., and Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.*21, 1794–1805. doi: 10.1111/j.1365-294x.2012.05538.x

Simpson, J. T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.*22, 549–556. doi: 10.1101/gr.126953.111

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*28, 405–420. doi: 10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l

Stav, S., Atilho, R. M., Arachchilage, G. M., Nguyen, G., Higgs, G., and Breaker, R. R. (2019). Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiol.*19:66. doi: 10.1186/s12866-019-1433-7

Storz, G., Vogel, J., and Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*43, 880–891. doi: 10.1016/j.molcel.2011.08.022

Tobar-Tosse, F., Rodríguez, A. C., Vélez, P. E., Zambrano, M. M., and Moreno, P. A. (2013). Exploration of noncoding sequences in metagenomes. *PLoS One*8:e59488. doi: 10.1371/journal.pone.0059488

Tucker, B. J., and Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*15, 342–348. doi: 10.1016/j.sbi.2005.05.003

Virgin, H. W., and Todd, J. A. (2011). Metagenomics and personalized medicine. *Cell*147, 44–56. doi: 10.1016/j.cell.2011.09.009

Wang, Q., Fish, J. A., Gilman, M., Sun, Y., Brown, C. T., Tiedje, J. M., et al. (2015). Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome*3:32.

Waters, S. A., Mcateer, S. P., Kudla, G., Pang, I., Deshpande, N. P., Amos, T. G., et al. (2017). Small RNA interactome of pathogenic E. coli revealed through crosslinking of RN ase E. *EMBO J.*36, 374–387. doi: 10.15252/embj.201694639

Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., et al. (2010). Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*11:R31.

Will, S., Otto, C., Miladi, M., Möhl, M., and Backofen, R. (2015). SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*31, 2489–2496. doi: 10.1093/bioinformatics/btv185

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*3:e65. doi: 10.1371/journal.pcbi.0030065

Williamson, S. J., and Yooseph, S. (2012). From bacterial to microbial ecosystems (metagenomics). *Methods Mol. Biol.*804, 35–55. doi: 10.1007/978-1-61779-361-5_3

Winter, J., Jung, S., Keller, S., Gregory, R. I., and Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat. Cell Biol.*11, 228–234. doi: 10.1038/ncb0309-228

Wu, M., and Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*28, 1033–1034. doi: 10.1093/bioinformatics/bts079

Yuan, C., Lei, J., Cole, J., and Sun, Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics*31, i35–i43.

Zhang, S., Borovok, I., Aharonowitz, Y., Sharan, R., and Bafna, V. (2006). A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements. *Bioinformatics*22, e557–e565.

Zhang, S., Haas, B., Eskin, E., and Bafna, V. (2005). Searching genomes for noncoding RNA using FastR. *IEEE/ACM Trans. Comput. Biol. Bioinform.*2, 366–379. doi: 10.1109/tcbb.2005.57

Zhong, C., Tang, H., and Zhang, S. (2010). RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res.*38:e176. doi: 10.1093/nar/gkq672

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 May 2021
doi: 10.3389/fgene.2021.648229

A Benchmark of Genetic Variant Calling Pipelines Using Metagenomic Short-Read Sequencing

*Sergio Andreu-Sánchez[1,2], Lianmin Chen[1,2†], Daoming Wang[1†], Hannah E. Augustijn[1], Alexandra Zhernakova[1] and Jingyuan Fu[1,2]**

[1] Department of Genetics, University of Groningen and University Medical Center Groningen, Groningen, Netherlands,
[2] Department of Pediatrics, University of Groningen and University Medical Center Groningen, Groningen, Netherlands

Microbes live in complex communities that are of major importance for environmental ecology, public health, and animal physiology and pathology. Short-read metagenomic shotgun sequencing is currently the state-of-the-art technique for exploring these communities. With the aid of metagenomics, our understanding of the microbiome is moving from composition toward functionality, even down to the genetic variant level. While the exploration of single-nucleotide variation in a genome is a standard procedure in genomics, and many sophisticated tools exist to perform this task, identification of genetic variation in metagenomes remains challenging. Major factors that hamper the widespread application of variant-calling analysis include low-depth sequencing of individual genomes (which is especially significant for the microorganisms present in low abundance), the existence of large genomic variation even within the same species, the absence of comprehensive reference genomes, and the noise introduced by next-generation sequencing errors. Some bioinformatics tools, such as metaSNV or InStrain, have been created to identify genetic variants in metagenomes, but the performance of these tools has not been systematically assessed or compared with the variant callers commonly used on single or pooled genomes. In this study, we benchmark seven bioinformatic tools for genetic variant calling in metagenomics data and assess their performance. To do so, we simulated metagenomic reads to mimic human microbial composition, sequencing errors, and genetic variability. We also simulated different conditions, including low and high depth of coverage and unique or multiple strains per species. Our analysis of the simulated data shows that probabilistic method-based tools such as HaplotypeCaller and Mutect2 from the GATK toolset show the best performance. By applying these tools to longitudinal gut microbiome data from the Human Microbiome Project, we show that the genetic similarity between longitudinal samples from the same individuals is significantly greater than the similarity between samples from different individuals. Our benchmark shows that probabilistic tools can be used to call metagenomes, and we recommend the use of GATK's tools as reliable variant callers for metagenomic samples.

Keywords: metagenomics, shotgun sequencing, short-reads, variant-calling, benchmark

OPEN ACCESS

Edited by:
Saumya Patel,
Gujarat University, India

Reviewed by:
Anna Heintz-Buschart,
German Centre for Integrative Biodiversity Research (iDiv), Germany
Dhaval K. Acharya,
B N Patel Institute of Paramedical, India

****Correspondence:***
Jingyuan Fu
j.fu@umcg.nl

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *31 December 2020*
Accepted: *22 March 2021*
Published: *10 May 2021*

Citation:
Andreu-Sánchez S, Chen L, Wang D, Augustijn HE, Zhernakova A and Fu J (2021) A Benchmark of Genetic Variant Calling Pipelines Using Metagenomic Short-Read Sequencing. Front. Genet. 12:648229. doi: 10.3389/fgene.2021.648229

INTRODUCTION

Short-read metagenomic sequencing is the technique most widely used to explore the natural habitat of millions of bacteria. In comparison with 16S rRNA sequencing, shotgun metagenomic sequencing (MGS) provides sequence information of the whole genomes, which can be used to identify different genes present in an individual bacterium and enables the examination of other genomic features such as gene synteny or genetic variation. In recent years, MGS datasets have been generated to explore the composition of the gut microbiome in a number of large human cohorts (Human Microbiome Project Consortium, 2012; Zhernakova et al., 2016; Lloyd-Price et al., 2017; Gacesa et al., 2020; Salosensaari et al., 2020). Large inter-individual variation in gut microbial composition has been widely observed, and variations in composition have been linked to lifestyle, host genetics, health, and disease. However, most of these associations reflect variations in microbial diversity and bacterial abundance, and our understanding of the genetic variations within gut bacteria is still limited.

Enthusiasm is now rising for techniques that can assess the genetic variation in the gut microbiome, which would allow us to pinpoint the putative causal bacterial genes underlying the observed associations and thereby generate testable hypotheses for mechanistic research. Single-nucleotide variation (SNV) refers to a one-nucleotide difference in a homologous region of at least two organisms. SNVs are of major importance for understanding the role of genetics in evolution, disease, phenotypes, or population genetics dynamics. The first major attempt to explore the bacterial genetic landscape revealed 10.3 million SNVs as well as many other types of genetic variants in 252 fecal samples (Schloissnig et al., 2013). However, there have been few efforts to assess the inter-individual differences in bacterial genetic profiles.

Despite its potential, SNV calling in a metagenome remains challenging. Many factors hamper the widespread application of variant-calling analysis, including the low-depth sequencing of individual genomes (which is especially significant for microorganisms present in lower abundance), large genomic variation (even within the same species), the absence of comprehensive reference genomes, and the noise introduced by next-generation sequencing errors. A plethora of different software have been produced to separate SNVs from sequencing errors after genomic mapping to a known reference. However, most tools require deeply sequenced single genomes with a known ploidy and, in all cases, mapping to a homologous region for proper function. Metagenomes also contain an unknown number of haploid organisms. Additionally, the identification of homologous regions is complicated by the presence of other bacteria that share the same evolutionary history and by possible horizontal gene-transfer events.

At present, there are several tools that have been developed specifically for metagenomic variant calling, such as MetaSNV (Costea et al., 2017) and InStrain (Olm et al., 2021). However, other variant callers have also been designed to be ploidy naïve or to address complications like an unknown number of pooled samples, including VarScan2 (Koboldt et al., 2012), freebayes (Garrison and Marth, 2012), and GATK's Mutect2 (DePristo et al., 2011). Other widely used variant-calling tools in the world of genomics include BCFtools and GATK's HaplotypeCaller (DePristo et al., 2011). All-in-all, these tools fall into two categories: *probabilistic tools* that calculate probabilities for a genotype given the read depth and quality of the base pairs (e.g., BCFtools, Mutect2, HaplotypeCaller, and freebayes) and *non-probabilistic* tools that call variants that pass specific thresholds such as minimal read depth or supporting reads (**Table 1**). While variant-calling benchmarks have been carried out in the context of bacterial variation (Yoshimura et al., 2019; Bush et al., 2020), currently, there is no benchmark on the metagenomic realm, where more complex issues exist.

We therefore aimed to benchmark different variant-calling tools in the context of metagenomes. We simulated complex metagenomic communities based on the 45 most abundant and prevalent gut microbial species across populations and disease groups (Gupta et al., 2020), which correspond, on average, to 74% of the human gut metagenome composition. We then applied seven tools to this simulated data and compared their performance under different scenarios. We further applied the tools that showed best performance on the simulated data, Mutect2 and HaplotypeCaller, to longitudinal, metagenomic-sequenced data from the Human Microbiome Project (HMP) (Schloissnig et al., 2013). This revealed the high individual specificity of microbial genetic variants, which allows them to be used to distinguish samples from the same individual taken at different times, with more power than bacterial taxonomic abundance.

TABLE 1 | Summary of tools benchmarked and used for different analyses.

Tool name	Probabilistic	Pool population	Joint calling	Minimal coverage	ROC curve	Real data
BCFtools	Yes	No	Yes	No	Yes	No
freebayes	Yes	Yes	Yes	No	Yes	No
HaplotypeCaller	Yes	No	Yes	No	Yes	Yes
Mutect2	Yes	Yes	Yes	No	No	Yes
VarScan2	No	Yes	No	8	No	No
metaSNV	No	Yes	Yes	4	No	No
InStrain	No	Yes	No	5	No	No

MATERIALS AND METHODS

Bacterial Species Selection and Reference Genome Download

To determine which references would be used for variant calling, we selected the 48 most abundant (mean relative abundance > 0.5%) and prevalent (presence rate > 20%) bacterial species from an integrated dataset of 4,347 publicly available human stool metagenomes, which were pooled across multiple studies encompassing various disease states (Gupta et al., 2020). These 48 species accounted for a mean total abundance of 81% (**Supplementary Table 1**), indicating that they capture a substantial proportion of human gut microbial composition. From these, three unclassified species were removed because no clear reference genome could be selected. The remaining 45 species accounted for 74% of mean abundance. To reach a 100% composition, we included one extra species (*Streptococcus australis*) with a dummy high abundance of 26%. We then used InSilicoSeq's (Gourlé et al., 2019) Download_ncbi script to query GenBank for the assemblies of the selected species using Biopython's entrez (Cock et al., 2009) Python package. When multiple assemblies were found for a given bacterial taxon, a reference genome was randomly selected from among the available assemblies. The reference used and the quality statistics are presented in **Supplementary Table 2**. Quality statistics measured include number of contigs; total length of the genome; minimum and maximum contig length; N50, N75, and N90 (shortest contig length needed to cover 50, 75, and 90% of the genome, respectively); and auN (area under the curve of all possible Nx metrics).

Synthetic Read Generation

We considered two different scenarios: a *uni-strain scenario* in which only one dominant strain exists per species and a *multi-strain scenario* where two dominant strains exist per species. We generated two sets of synthetic variants, considered true positives (TP), by randomly changing 1% of the total nucleotides in each of the reference genomes (including the dummy taxa). The choice of this SNV rate was based on a previous estimate that found that the SNV diversity of most intestinal species was around 1% (range 0.018–3.9%) (Truong et al., 2017). The first dataset was used for the uni-strain scenario. The second dataset was then combined with the first, and the combined set used as the multi-strain scenario. Additionally, we repeated the simulation with 4% variation to reflect highly divergent strains and assess whether tools performance differed for highly divergent species.

Using the mutated reference genomes, we ran InSilicoSeq (iss generate) (Gourlé et al., 2019) on the known bacterial taxonomy table (–abundance_file) to generate a simulated set of ∼15 million Hiseq paired-end reads, a sequencing depth similar to other metagenomic datasets (Zhernakova et al., 2016; Byrd et al., 2021). InSilicoSeq simulates reads using an error model based on Illumina's Hiseq technology. We can estimate the expected genome coverage by adjusting the Lander–Waterman estimation method for computing coverage by the abundance of the taxon (Lander and Waterman, 1988). Using the reference genomic length, simulated abundance, read length (126 bp), and number of reads, we estimated the expected coverage for each of the microbial species using Equation (1).

$$Expected\ coverage_i = \frac{Reads \times 126 \times Abundance_i}{Genome\ length_i}$$

Equation 1. Expected coverage of a given species i. Reads is a constant per simulation indicating the number of simulated reads. Abundance indicates the relative abundance of a species (0–1). Genome length is the total number of base pairs in the reference genome of a species i.

Read Trimming

The simulated dataset was trimmed following a typical metagenomics pipeline. We removed low-quality reads from the raw metagenomic sequencing data using KneadData (version 0.7.4). KneadData can also remove host genome-contaminated reads, which should not exist in the simulated scenario, but is necessary in real-life human-derived microbiome projects. KneadData uses Bowtie2 (version 2.3.4.3) (Langmead and Salzberg, 2012) and Trimmomatic (version 0.39) (Bolger et al., 2014). In brief, the data-cleaning procedure includes two main steps: (1) filtering out of the human genome-contaminated reads by aligning raw reads to the human reference genome (GRCh37/hg19) and (2) removal of adaptor sequences using Trimmomatic (default trimming: SLIDINGWINDOW:4:20 MINLEN:70).

Genome Mapping

We used the standard setting (*–sensitive* mode) of Bowtie2 (version 2.3.4.3) (Langmead and Salzberg, 2012) to map the simulated metagenomic reads to the unmutated original reference genomes (the reference genomes were mapped one at a time), using default options. Reads were sorted using SAMtools (version 1.9) (Li et al., 2009), and duplicates were marked and removed by running the MarkDuplicates module (version 2.18.26-SNAPSHOT) (*REMOVE_DUPLICATES = True*) of Picard. We cleaned the resulting BAM files using the CleanSam module (version 2.18.26-SNAPSHOT) of Picard.

Redundancy of Genome Assessment

To compute the similarity between the 46 chosen genomes, we used Mash (Ondov et al., 2016) with a k-mer size of 17 and a sketch size of 10,000. In addition, we estimated the proportion of multi-mapping reads in a combined reference of all 46 species. With this, we aimed to characterize how much genome homology would impact read assignment. We therefore extracted the information regarding the number of concordant reads (i.e., both pairs mapping meaningfully), concordant reads with multiple equally good mapping positions, pairs that mapped non-concordantly, unpaired reads mapped uniquely, and unpaired reads mapping to multiple positions. We estimated the number of multi-mappers by summing both

paired mapped and unpaired mapped equally well mapping reads (Equation 2).

$$Multi\_mapping\ rate\ = \frac{2\ \times \left(Multi\_mapped\ pairs\right) + Multi\_mapped\ unpaired\ reads}{2 \times Paired\ read\ number}$$

Equation 2. Multi-mapping rate. The number of multi-mapped reads include two times the number of multi-mapped pairs and the unpaired reads that were mapped to multiple positions.

In addition, if we consider non-concordant pairs as reads mapping to an incorrect position (since there are no structural variations in the reference), we can get a second estimate of the reads mapping to positions other than their origin (Equation 3).

$$Incorrect\ or\ multi\ mapping\ rate\ = \frac{Non\_concordant\ pairs + \ Multi\_mapping\ reads}{2 \times Paired\ read\ number}$$

Equation 3. Incorrect or multi-mapping rate. The number of multi-mapped reads include two times the number of multi-mapped pairs and the unpaired reads that were mapped to multiple positions. Non-concordant pairs refer to pairs where both reads mapped to a single position but do not follow the expected read orientation.

Variant Calling

Using the cleaned BAM files and the reference (not-mutated) genomes, we performed variant calling using the following tools and specifications.

BCFtools

BCFtools variant calling is based on BCFtool's mpileup output. For each bacterial alignment, we used mpileup (default options) and BCFtools call. Ploidy was set to 1, and we used the multi-allelic calling algorithm (-m). The BCFtools algorithm does not consider a population of pooled samples, and as we run it on a sample-by-sample basis, it only assesses two possible genotypes: reference or alternative. It is also worth noting that in order to calculate likelihoods, BCFtools uses a prior based on the human effective population size (theta) of 0.001[1].

Freebayes

freebayes is a haplotype-based variant caller (Garrison and Marth, 2012). This means that instead of calling variants position-by-position based on an aligned read, it checks the whole haplotype of the read independently of the precise alignment positions. This solves the issue of multiple ambiguous alignment possibilities between the read and the homologous genomic region. We set ploidy to 1. As we have an unknown number of pooled samples (bacteria) that might align with a homologous region, we set the parameter –pooled-continuous. The joint-calling options –min-alternate-count and –min-alternate-fraction were set to 2 and 0, respectively.

[1]https://samtools.github.io/BCFtools/call-m.pdf

HaplotypeCaller

GATK's assembly-based variant caller HaplotypeCaller (DePristo et al., 2011) is able to handle non-diploid organisms as well as pooled experiment data. We therefore applied HaplotypeCaller with default settings, with the exception of setting ploidy to 1. Haplotypes are called by HaplotypeCaller via local re-assembly of regions of a potential variant site, from which a pair-HMM alignment of reads to haplotypes is generated. In the final step, the algorithm determines the likelihoods of the genotypes and reports the most likely genotype at each site[2].

Mutect2

GATK's Mutect2 (DePristo et al., 2011) uses a similar approach to HaplotypeCaller in calling variants, including active region-based identification, assembly-based haplotype reconstruction, and pair-HMM alignment of reads to haplotypes. However, whereas HaplotypeCaller is designed to call germline variants, Mutect2 is designed to call somatic variants. Mutect2 therefore includes somatic-specific genotyping and filtering steps. It is designed to have a high specificity but cannot calculate reference confidence and define ploidy. We employed Mutect2 in tumor-only mode with default settings but including "–af-of-alleles-not-in-resource 0.33," as recommended when using a non-human organism as input[3].

VarScan2

VarScan2 employs a heuristic approach to call variance that relies on parameter thresholds to determine variants (Koboldt et al., 2012). Given a SAMtools mpileup-formatted alignments file, VarScan2 first performs a read-filtering step that discards any reads that align to multiple locations or do not comply with the quality criteria. VarScan2 then screens the alignments on a per-read basis to detect sequence variance and merges variants detected in multiple reads into unique SNPs and indels. Only variants meeting user-defined parameter thresholds are reported. Here, we applied the VarScan2 mpileup2snp algorithm using default settings, including minimum read depth of 8, base quality of 15, supporting reads of 2, allele frequency of 0.01, and a Fisher's exact test p-value below 0.99.

MetaSNV

MetaSNV was specifically designed for metagenomic datasets and can handle large multi-species references (Costea et al., 2017). We applied the default parameters of metaSNV for SNV calling. metaSNV determines the existence of a candidate variant on a per-nucleotide basis, building upon the mpileup tool in the Samtools suite (Li et al., 2009). All reads from all samples that align to a given position are considered together. If at least four variant-containing reads cover a position (across all samples), it is considered a potential SNV. Variants are split into two classes: population and individual variants. Population variants are non-reference nucleotides observed in $>$ 1% of all reads combined across all samples. Individual variants are those that fall below the 1% population frequency threshold but are confidently observed

[2]https://gatk.broadinstitute.org/hc/en-us/articles/360036712151-HaplotypeCaller
[3]https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2

in at least one sample (at least four reads containing the variant). If multiple different non-reference nucleotides are observed, all are reported independently.

InStrain

InStrain (Olm et al., 2021) was devised to detect SNVs and profile intra-population genetic diversity based on metagenomic short-read alignment. InStrain first performs read filtering to remove read pairs that do not meet the quality criteria. Then, for each position with multiple aligned reads supported in the reference genome, both biallelic and multiallelic SNVs are identified by detecting bases that are different from the reference genome at the same position. The frequencies of SNVs are also counted. Additionally, if the gene annotation for the reference genome is provided, InStrain also classifies the identified SNVs as synonymous, non-synonymous, or intergenic SNVs. In our benchmark testing pipeline, we ran InStrain with default parameters: minimal coverage of a position of five reads, minimal frequency of an SNP of 0.05 and an FDR (based on *a priori* empirical tests) of 1×10^{-06}.

Joint Variant Calling

Most variant callers pull information from a population of samples to make their calling more accurate. The same options we described above were therefore used to perform variant calling on two BAM files at the same time to test for improved performance. Each of the BAM files represented a different scenario, namely, uni-strain or multi-strain. For BCFtools, we performed a joint mpileup call followed by a BCFtools call. For freebayes, we included both BAM files in the input. In HaplotypeCaller and Mutect2, we performed classic joint calling by calling variants simultaneously across BAM. In metaSNV, we profiled both BAM files together. To the best of our knowledge, variant calling in InStrain and VarScan2 does not benefit from joint variant calling and was not done.

Statistics Assessment

Variant calling outputs were reformatted to a homogenous format. From the VCF outputs (BCFtools, freebayes, HaplotypeCaller, Mutect2), we extracted information regarding chromosome, position, reference allele, and alternative allele. When variants of more than one nucleotide were reported, they were decoded into as many independent variants as polymorphisms found. If the Phred quality score was available, this information was included in the standardized file. Multiple alternative allele variants were encoded as independent variations. MetaSNV was similarly reformatted, and multivariate positions were decoded as independent variants. InStrain's positions were transformed to 1-based indexing, and each nucleotide that did not match the reference was decoded as an independent variation.

In joint variant calling, for VCF files, we used the predicted genotype in each of the input files.

The set of sequence-covered variants was determined by overlapping the list of simulated true variants with the coverage profile generated by Samtools (v1.9) mpileup.

Next, the list of mutations covered was overlapped with each of the tool's variant calls. True positives (TP) were when the variant was present in both the called profile and the covered variants. False negatives (FN) were when the covered variants were not present in the called profile. False positives (FP) were when the called variants were not present in the covered profile. True negatives (TN) were all the covered positions that remained after subtracting TPs, TNs, and FPs. Sensitivity was calculated using Equation (4) and precision using Equation (5).

$$Sensitivity = \frac{TP}{(TP + FN)}$$

Equation 4. Sensitivity. TP, true positives; FN, false negatives.

$$Precision = \frac{TP}{(TP + FP)}$$

Equation 5. Precision. TP, true positives; FP, false positives.

Receiver Operator Characteristic Curves

We generated sensitivity and precision statistics using the different Phred score thresholds provided in the probabilistic methods BCFtools, freebayes, and HaplotypeCaller. We generated 50 quality thresholds ranging from the 2% quantile to the 100% quantile of the Phred score distribution in each sample.

Variant Calling in Real Data

Metagenomic data from the HMP (Schloissnig et al., 2013) belonging to 43 participants from samples taken at two timepoints up to a year apart (86 samples total) were downloaded from the HMP public repositories[4]. We then selected IDs based on **Supplementary Table 1** from Schloissnig et al. (2013). Reads were pruned of human contamination, trimmed using KneadData, and mapped to the reference genomes of 10 representative species that we previously benchmarked on simulated data. These representative species were selected by calculating the mean sensitivity and precision statistics measured for each species in the simulation dataset across all tools. The genomes were selected to represent both good and poor SNV calling performance and the overall genetic diversity (**Supplementary Figure 2**). A Manhattan distance matrix was then calculated based on mean sensitivity and precision of each species. Finally, based on the calculated Manhattan distance matrix, we assigned all species into 10 clusters using the partitioning clustering pam() function in R. A representative species in each cluster was randomly selected. The representatives were merged in a unique reference and mapped against the 86 paired-end samples using Bowtie2 (Langmead and Salzberg, 2012). Picard's MarkDuplicates and CleanSam were used to clean the mapped reads. Variant calling on BAM files was performed with HaplotypeCaller and Mutect2, as described above.

Genetic Distance Calculation

We computed the genetic distance between pairs of HMP samples. For this, we first defined a set of variants (reference

[4]https://hmpdacc.org/hmp/

variants) by including all called variants in any taxon and sample that were present in at least two HMP samples (removing singletons). We further produced a site frequency spectrum plot by counting the number of individuals in which each variant was observed. We profiled each sample by creating a presence–absence matrix for each of the reference variants. We then computed the Manhattan distance [python scipy, spatial.distance.pdist (metric = "cityblock")] between the different samples. We also reproduced the same analysis using only the variants found in each of the species in the reference genome.

Clustering of Intra-Individual and Inter-Individual Samples

Genetic distances were clustered using hierarchical clustering with the nearest point algorithm. The number of samples clustering together at both baseline and follow-up were counted. In addition, we did a second clustering based on Bray–Curtis distances using the HMP abundance data generated with the previously described settings (Chen et al., 2020). The distances between the same individual at baseline and follow-up (intra-individual) and among independent samples (inter-individual) were compared using the Wilcoxon test.

Statistical Analysis

Rv3.5.1 and Pythonv3.7.0 were used for plotting and statistical calculations. To address the effect of bacterial abundance and genome coverage on the specificity and sensitivity of the different tools, we associated the precision and sensitivity of different tools with bacterial abundance and coverage using Spearman's correlation. The effect of genome quality on precision and sensitivity was assessed by linear modeling (ordinary least squares). We built a null model explaining either precision or sensitivity while including an interaction effect of tools and simulation design (uni-strain or multi-strain), which was found to be significant. We then built a second model including either the N50 or the number of contigs as a proxy of genome quality on top of the null model. We also built a third model that considered an interaction effect of the genome quality and tool and design. We assessed significance among the nested models using a likelihood ratio test.

To address the effect of simulation benchmark metrics and the number of variants called in a genome on real data clustering performance, we built a linear model using the percentage of samples where baseline and follow-up clustered together as the dependent variable and the tool, benchmarked sensitivity and accuracy, and number of called variants per bacteria as regressors.

We used Wilcoxon tests to estimate whether there were differences between the specificity and sensitivity statistics of joint-called samples or individually called samples. In addition, we compared sensitivity and accuracy metrics between the simulations with 1 and 4% of mutated positions using Pearson correlation.

We set a significant p-value threshold of 0.05.

Data Availability

The pipelines used for simulation and variant calling in simulated data and for variant calling in real data were written in Snakemake (v5.9.1) (Köster and Rahmann, 2012) and can be found, together with the plotting and statistical analyses scripts, in our Github repository[5].

RESULTS

Experimental Design and Simulations

We first simulated metagenomic datasets. We did so by including the 45 most common bacterial species, which accounted for an average 74% of bacterial abundance in a recent large multi-ethnic study (Gupta et al., 2020) and adding one genome with the remaining abundance to reach 100%. The reference genomes for the 46 species were randomly selected from the species genomes available in the NCBI. We then introduced known SNV variants in 1% of the genomic positions. On average, the number of contigs found in each reference genome was 112.17 and ranged between 1 and 1,541 (**Supplementary Figure 1**), and the average N50 was 1,297,250.67 base pairs (5,884–6,271,157) (**Supplementary Figure 1**). The N50 distribution of all reference genomes followed a bimodal distribution, showing the existence of both high- and low-quality reference genomes.

To address the presence of homology among species, which may bias read mapping, we measured the k-mer-based distance of the reference genomes using Mash (Ondov et al., 2016). We found a mean Mash distance of 0.35 (0.04–1, SD = 0.08) (dendrogram shown in **Supplementary Figure 2**). In addition, we counted the number of reads that mapped equally well in more than one position to a concatenated multi-reference genome and found 8% of reads to be multi-mappers. However, if we consider discordant pairs as incorrectly assigned reads due to homology or horizontal gene transfer events, this percentage grows to nearly 36% (if multi-mappers are also considered), which will influence the false positives identified by SNV-calling tools.

We processed the simulated reads (using KneadData for trimming and Bowtie2 for mapping) and ran the seven different variant-calling tools (**Figure 1**). Of these, four are probabilistic methods—BCFtools, Mutect2, HaplotypeCaller, and freebayes—meaning that they use the coverage, base call quality, and error rate expectations to infer genotype likelihoods and call non-reference nucleotides. We used options for haploid variant calling and, as a metagenome, can be considered as a pooled sample of an unknown number of multiple organisms; if an option for running an unknown number of pooled samples existed, we used it (e.g., in freebayes). The other three tools—VarScan2, metaSNV, and InStrain—are non-probabilistic methods and are mainly based on applying specific filters as the minimal coverage to consider a variant, which we set to the default value for each tool (**Table 1**).

We ran two different simulations, one assuming that one unique strain was present per bacterial taxa (uni-strain scenario)

[5]https://github.com/GRONINGEN-MICROBIOME-CENTRE/Groningen-Microbiome/tree/master/Projects/Metagenomics_SNVcalling_Benchmark

FIGURE 1 | Representation of the pipeline used. Mutations are introduced into each reference genome before simulation. Simulated reads are trimmed, and human contamination removed. Reference genomes are indexed and mapped using simulated cleaned reads. Alignments are further cleaned before variant calling. All variant outputs are converted to the same format. Introduced variants covered by simulated reads are used as a set of true positives. Variants are checked in the formatted variant calls, and receiver operator characteristic (ROC) curves are calculated by repeating this process at different quality thresholds. All statistics are combined in a single file.

and another assuming two different strains per bacterial taxa (multi-strain scenario).

Probabilistic Methods Show Better Sensitivity and Precision

We computed sensitivity and precision statistics from the variant-calling results (**Supplementary Table 3**). In the uni-strain scenario, all four probabilistic methods showed high sensitivity for most of the organisms (**Figure 2A**). Mutect2 and freebayes had the highest sensitivity, recalling nearly 100% of covered variants, followed by HaplotypeCaller and InStrain (the only non-probabilistic method with high sensitivity), which had low sensitivity for some taxa. BCFtools ranked 5th, and there was a large difference in performance between metaSNV, which showed the largest variability among taxa, and VarScan2, which had poor sensitivity in most samples. Both metaSNV and VarScan2 missed many covered variants (**Figures 2C,D**). Precision was also higher in the probabilistic tools (**Figure 2B**), where both BCFtools and HaplotypeCaller achieved a similar performance, with a low number of FP (**Figure 2E**) and little variation in precision among taxa compared with other tools. They were followed by Mutect2 and InStrain. MetaSNV showed a better average precision than freebayes, but also had a higher standard deviation. Once again, despite its low number of FN, VarScan2 was penalized by its low number of TP and showed the highest variation and average low precision.

Overall, probabilistic methods showed the best compromise between sensitivity and precision (**Figure 2F**). They showed a lower precision and sensitivity variability among taxa compared with non-probabilistic methods, which were penalized by unequal coverage in low-abundance species. This was especially true when comparing sensitivity. However, the most sensitive tools, freebayes and Mutect2, showed high variability in their

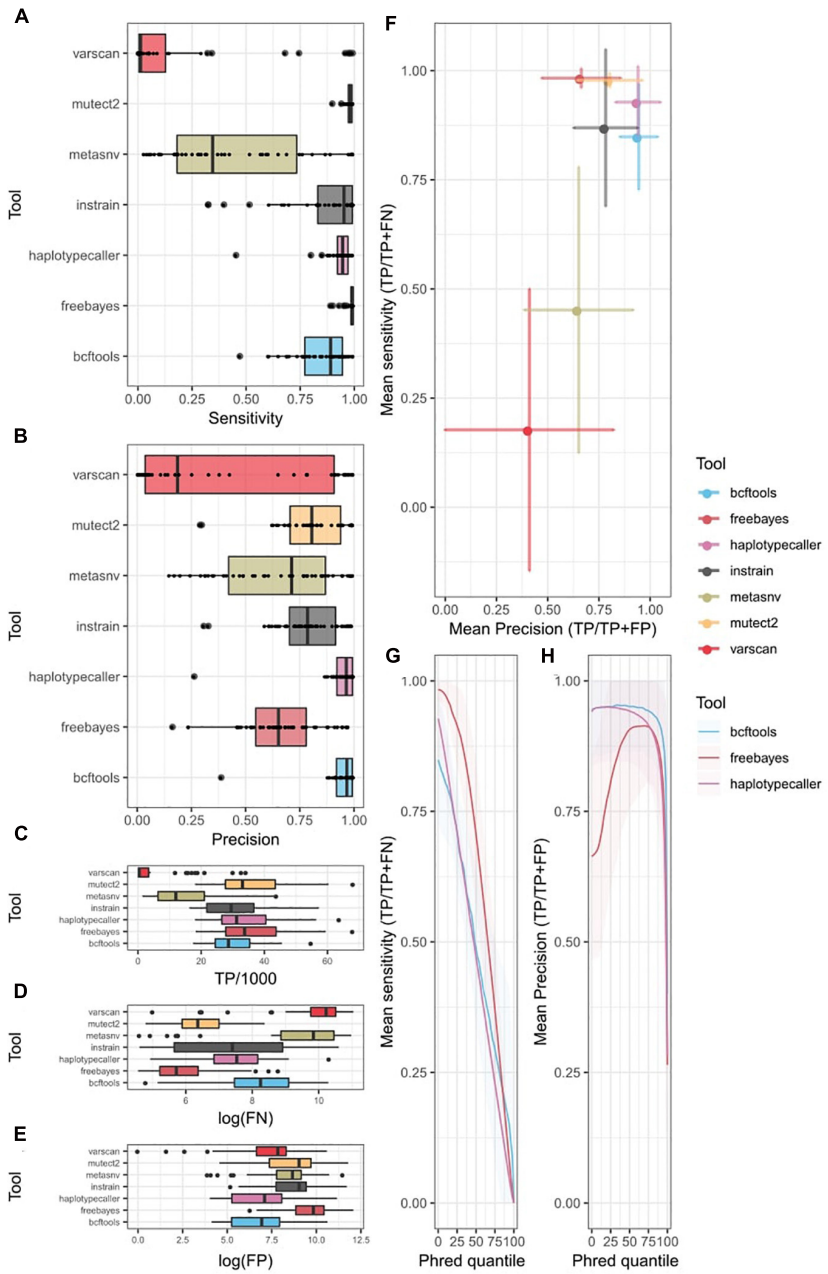

FIGURE 2 | Single strain variant-calling statistics of seven different tools. Colors indicate different tools. **(A)** Sensitivity (TP/TP + FN) of each tool. Tukey box plot presents the distribution of precision. Dots show precision per individual bacteria. **(B)** Precision (TP/TP + FP) of each tool. Tukey's box plot presents the distribution of precision. Dots show precision per individual bacteria. Distribution of **(C)** TP, **(D)** FN, and **(E)** FP per tool as Tukey box plots. Individual dots indicate bacteria over 1.5 times the interquartile distance. **(F)** Precision vs. sensitivity plot. Dots indicate mean values among all bacteria. Error bars represent the standard deviation from the mean. **(G,H)** ROC curve of probabilistic methods. *X*-axis represents the quantile Phred filter. **(G)** Mean sensitivity changes with changes in Phred score. Line shows the mean value among bacteria. Shading represents the standard deviation from the mean. **(H)** Mean precision changes with Phred score changes. Line represents mean value among bacteria. Shading represents the standard deviation from the mean. TP, true positives; FP, false positives; FN, false negatives.

precision. HaplotypeCaller, with the highest precision, seems a better option than BCFtools, which had a lower sensitivity despite having a similar precision.

In addition, we tested tool performance by including more divergent strains from the reference (4%) in a uni-strain scenario. These results replicated our observations from the 1% divergence scenario (**Supplementary Figure 3**) and showed an overall high correlation coefficient both in sensitivity and precision (**Supplementary Table 4**).

An additional advantage of probabilistic methods is the availability of a quality metric that can be easily tuned to recalculate sensitivity and precision values (**Table 1**). We

generated an ROC curve for different values of this quality in freebayes, HaplotypeCaller, and BCFtools (**Figures 2G,H**) and observed an almost linear decrease in sensitivity with higher-quality thresholds in all three tools. freebayes remained the most sensitive method at almost any quality threshold. BCFtools achieved a better sensitivity than the other two tools at higher-quality thresholds, which seems to indicate that the highest quality values of freebayes and HaplotypeCaller have, on average, more FP. The precision curve showed large variability among bacteria. Both BCFtools and freebayes, which showed the highest precision without any tuning, did not improve their precision with higher thresholds, on average. However, their precision did decrease at the highest thresholds, probably showing the existence of FP when using a high-quality threshold. On the other hand, freebayes' precision was improved substantially by increasing the quality threshold, but needs up to a 75% quantile, on average, to catch up to the other tools, which results in an acute decrease in sensitivity. However, given the large number of low-quality calls in freebayes (**Supplementary Figure 4**), a minimal quality filter may be required to substantially improve the performance of this tool.

We further explored the sensitivity and precision of tool performances in the multi-strain scenario (**Supplementary Table 5**). Since both strains only differed in the genomic locations where the 1% of variants were generated, no structural or copy number variations were included. The major difference compared with the uni-strain scenario was that there were twice as many variants. Therefore, there might be multiple variants in the same locus, and the number of reads covering a single variant from a specific strain were now reduced by half (both strains are assumed to have equal abundance, so each has half the abundance simulated in the uni-strain scenario).

Sensitivity showed an acute decrease (**Supplementary Figure 5**). While some tools achieved a mean sensitivity of ∼90% in the uni-strain case, in the multi-strain scenario, this value was only achieved for some species. Such species did not have a significantly higher abundance than the ones with lower sensitivity. Mutect2 and freebayes remained the two tools with the highest recall, with median sensitivity around 50%, which might be explained by the fact that the most-covered positions might not include the variant from one of the strains but rather the reference allele from the other strain. This could increase the FNs, despite the mutation not being covered. The precision results showed the same pattern and range as in the single-strain scenario since FNs are not used, showing high estimates mainly in HaplotypeCaller and BCFtools. ROC curves in BCFtools, freebayes, and HaplotypeCaller also showed similar results to the uni-strain scenario.

Joint Variant Calling Benefits Complex Scenarios With Multiple Strains

Most of the tools analyzed also enable joint variant calling of multiple samples, meaning that different samples (in our case simulations) are pooled together during the variant-calling process, in contrast to the independent variant calling in each individual sample that we had performed previously. This could improve overall performance, although it would make it more difficult to detect singletons. We combined our two simulated datasets in order to perform joint variant calling in the tools where we expected this would be beneficial: BCFtools, freebayes, Mutect2, HaplotypeCaller, and metaSNV (**Figure 3**). Our results show that joint variant calling increased the sensitivity in the multi-strain scenario ($p = 3.68 \times 10^{-18}$) (**Figure 3C**). Precision was also significantly improved in freebayes, while Mutect2, BCFtools, and HaplotypeCaller had a significant decrease in precision (**Figure 3D**). In the uni-strain scenario, joint calling only improved metaSNV's sensitivity (**Figure 3A**) and freebayes' precision (**Figure 3B**) and led to an overall decrease in both precision and sensitivity.

Factors Affecting Variant-Calling Performance: Species Abundance and Coverage Have a Tool-Specific Effect, Whereas the Effect of Genome Quality Is Constant

The precision and sensitivity of SNV calling is affected by both species' abundance and coverage, with the non-probabilistic tools VarScan2 and metaSNV especially affected (**Figure 4**). For species with low or medium abundance, the precision and sensitivity of metaSNV and VarScan2 were lower than for other tools, and the performance of these two tools improved linearly as species coverage increased. For species with a low abundance, the sensitivities of InStrain and BCFtools were significantly affected by the coverage. The performances of HaplotypeCaller and Mutect2 were not significantly affected by species abundance and coverage: their precision and sensitivity were high and stable even in low abundance species and in both uni- and multi-strain settings. The performance of freebayes was unstable and not linearly associated with species abundance and coverage (**Supplementary Table 6**).

In addition, we tested if the reference genome chosen has a significant effect on tool performance using two different proxies of genome quality. First, we tested the tools with the classic N50 metric and found no significant effect on either sensitivity or precision metrics. The number of contigs per reference did have an overall significant negative effect on sensitivity ($p = 8.42 \times 10^{-6}$) and precision ($p = 3.76 \times 10^{-9}$), but there was no significant interaction effect with specific tools.

Genetic Distances Are More Individual Specific Than Bacterial Abundance

Finally, we decided to apply the best performing tools to real HMP data from 43 individuals taken at two timepoints up to 1 year apart (Schloissnig et al., 2013). For this, we chose Mutect2, which showed the best sensitivity under all conditions tested, and HaplotypeCaller, which showed the best precision and a better sensitivity than BCFtools. The overall genetic distance between samples was estimated from the combined SNV profile from 10 selected bacteria. The number of variants identified with both methods included a large number of singletons (**Figure 5A**). We therefore considered only variants observed in at least two samples. The genetic distance between individuals in

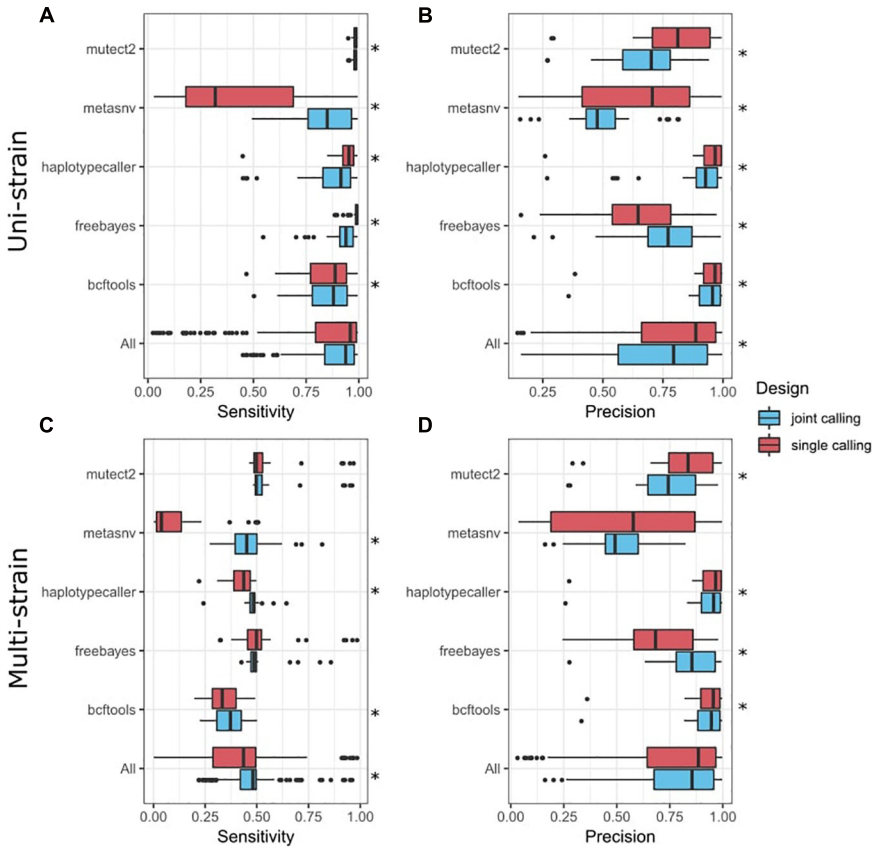

FIGURE 3 | Comparison of variant-calling performance between joint calls and single calls. *Y*-axis presents tools, including a combination of all tools (All). Tukey box plots are shown and colored according to the variant-calling mode. Single points represent samples >1.5 times the interquartile distance. Asterisks indicate statistically significant differences ($p < 0.05$) in a paired Wilcoxon test comparing both groups. **(A)** Sensitivity metrics in the uni-strain scenario. **(B)** Precision metrics in the uni-strain scenario. **(C)** Sensitivity metrics in the multi-strain scenario. **(D)** Precision metrics in the multi-strain scenario.

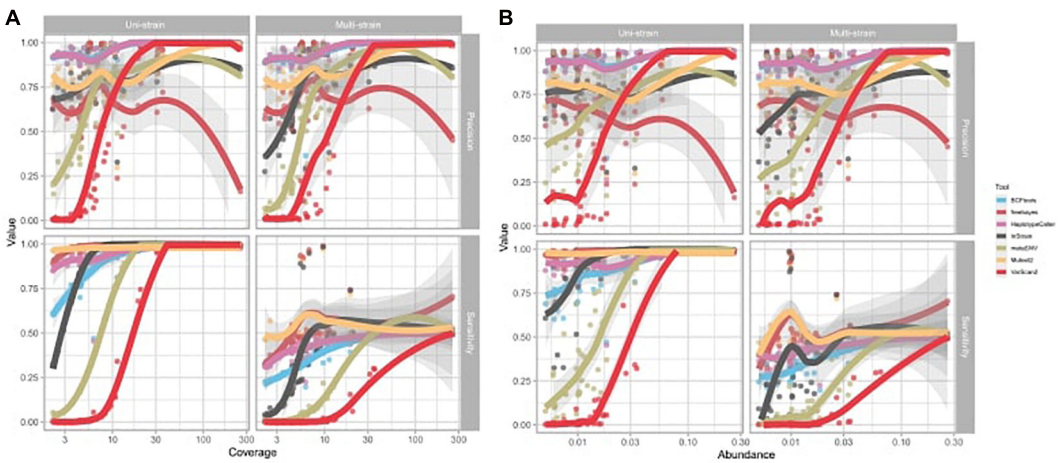

FIGURE 4 | Effect of average depth of genome coverage and species abundance on variant calling performance. Each dot represents a sample. **(A)** Effect of reference genome coverage on precision and sensitivity of each tool. **(B)** Effect of species abundance on precision and sensitivity of each tool. Trend lines were fitted with local polynomial regression (LOESS). Shading represents the 95% confidence interval of the trend line.

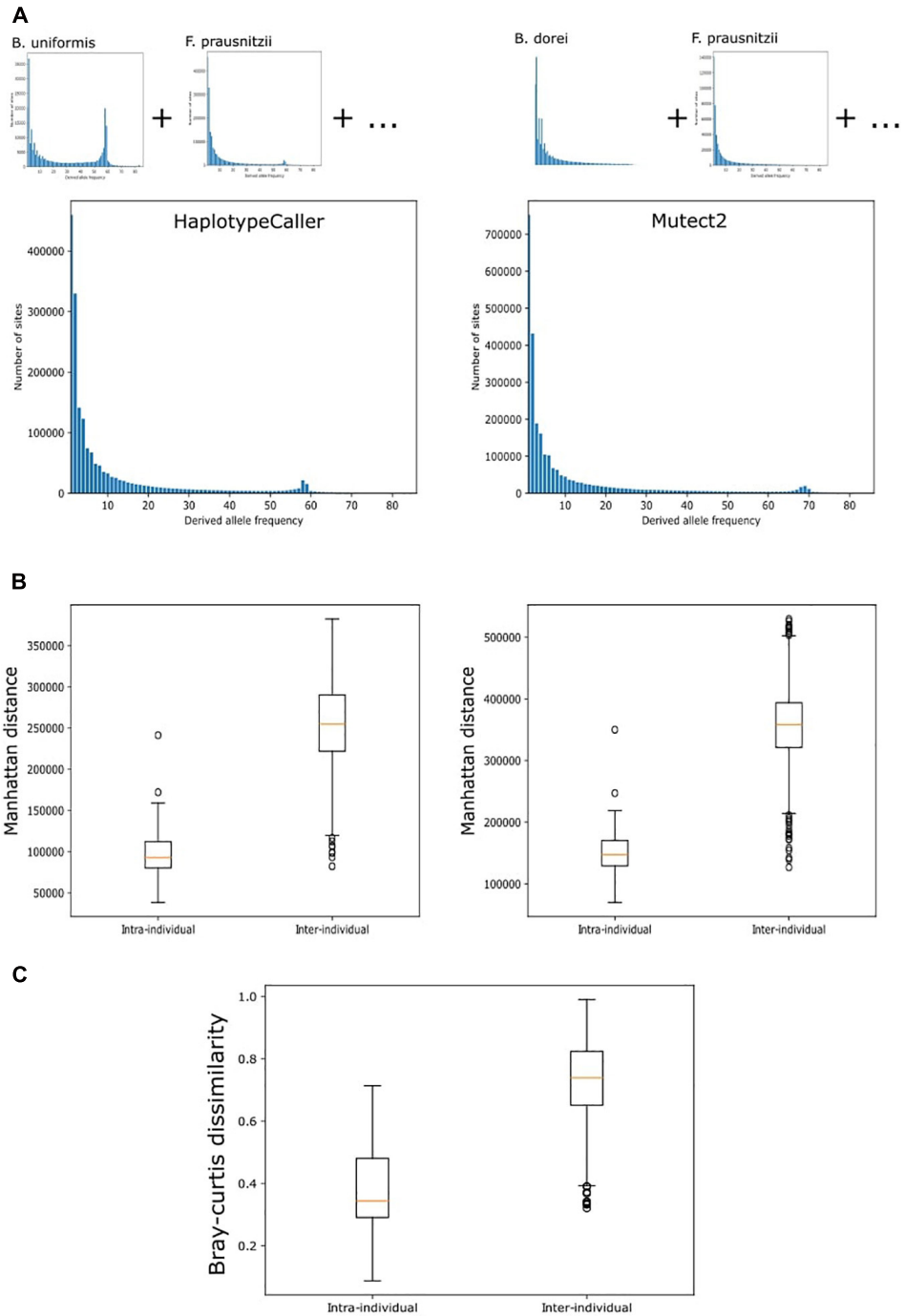

FIGURE 5 | Variant calling on real data. **(A)** Site frequency spectrum of variants in the whole metagenome called by HaplotypeCaller and MetaSNV. Total site frequency spectrum is the sum of the variants in the 10 chosen bacterial taxa. **(B)** Genetic distance between samples. Manhattan distances were calculated from the SNV profile of each sample. Tukey boxplot shows the distribution of distances between samples belonging to the same individual at two timepoints (intra-individual) and between samples from different people (inter-individual). These distances were used to cluster the data. **(C)** Tukey boxplot of the distribution of Bray-Curtis dissimilarities computed from the estimation of taxonomic abundance between samples from the same individual and from different people.

both Mutect2 and HaplotypeCaller, as measured by Manhattan distance, showed that samples from the same individual clustered together in 93% of cases (40 of 43 samples) (**Table 2**). Using individual bacteria instead of the combined genetic distance, the clustering values ranged from 4.6% (*Eubacterium hallii*, both tools) to 97% (*Bacteroides uniformis*, using Mutect2) (**Table 2**). Bacteria with higher resolution potential showed no correlation with simulation scores but had a positive

TABLE 2 | Single nucleotide variation (SNV) profiles in the Human Microbiome Project (HMP) data.

Taxa_profiled	Tool	Number of variants	% Clustered	Sensitivity	Precision
Akkermansia_muciniphila	M	132,878	20.9	0.955	0.992
Akkermansia_muciniphila	H	115,968	18.6	0.952	0.996
Alistipes_shahii	M	180,781	74.4	0.984	0.852
Alistipes_shahii	H	98,442	69.8	0.951	0.976
Bacteroides_dorei	M	264,480	93	0.968	0.290
Bacteroides_dorei	H	225,677	93	0.453	0.263
Bacteroides_uniformis	M	272,535	97.7	0.994	0.705
Bacteroides_uniformis	H	218,598	93	0.957	0.986
Dorea_formicigenerans	M	120,086	16.3	0.974	0.682
Dorea_formicigenerans	H	77,326	7	0.879	0.881
Eubacterium_hallii	M	75,369	4.7	0.989	0.802
Eubacterium_hallii	H	48,786	4.7	0.963	0.964
Eubacterium_rectale	M	234,000	79.1	0.953	0.830
Eubacterium_rectale	H	189,376	65.1	0.949	0.993
Faecalibacterium_prausnitzii	M	324,595	46.5	0.972	0.931
Faecalibacterium_prausnitzii	H	217,101	37.2	0.970	0.995
Ruminococcus_gnavus	M	86,590	23.3	0.983	0.738
Ruminococcus_gnavus	H	52,391	16.3	0.922	0.910
Ruminococcus_sp_5_1_39BFAA	M	182,558	34.9	0.964	0.799
Ruminococcus_sp_5_1_39BFAA	H	107,252	14	0.938	0.968
Total SNV profile	M	1,873,872	93	NA	NA
Total SNV profile	H	1,350,917	93	NA	NA

Taxa profiled indicates the name of each of the chosen bacteria profiled. Total SNV profile refers to the total profile, including all variants. Tool indicates the variant caller used: M, Mutect2; H, HaplotypeCaller. Number of variants indicates the total number of variants uncovered with presence in at least two samples. % Clustered indicates the percentage of samples that clustered together at both follow-up and baseline. Sensitivity and precision are the statistics estimated from the uni-ref simulation.

association with the number of called variants (lineal model, F-test, $p = 5 \times 10^{-4}$). Using the complete variant dataset, we found a highly significant difference in the distribution of distances between intra-individual and inter-individual samples (Wilcoxon test, HaplotypeCaller: $p = 1.31 \times 10^{-28}$, Mutect2: $p = 1.51 \times 10^{-28}$) (**Figure 5B**). These results did not improve when we only considered variants present in both methods at the same time. In addition, we performed the same analysis based on taxonomic abundance, where we could cluster together 63.7% of the samples (27 of 43) (**Figure 5C**), highlighting the stability of genetic variation in comparison with taxonomic abundance.

DISCUSSION

Microbiome genomic analyses are currently complicated by several factors, including low taxon-specific read depth, unequal taxonomic abundance, the existence of orthologs and paralogs, and horizontal transfer of genetic material. On top of these issues, single nucleotide variant calling suffers from the lack of high-quality reference genomes and the pooling of a population consisting of an unknown number of genomes. This benchmark study therefore assessed the performance of current variant callers in this complex scenario.

We used a homogenous pipeline that does not consider the complexity layer of read mapping since we used the default bowtie2 options. We used 45 microbial species that are highly abundant and prevalent in the human gut (Gupta et al., 2020) to create two simulation datasets that mimic HiSeq MGS experiments. Reference genomes for each of the species were randomly selected from GenBank and contained both high- and low-quality assemblies. Although the number of contigs present in the assembly, which might indicate genome fragmentation and poorer assemblies, did correlate with an overall decrease in sensitivity and precision, this effect was not tool specific and should not bias our comparison. This does, however, indicate that genome quality is an important factor to consider in the variant calling processing. In this line, it is important to highlight that previous benchmarks of bacterial variant calling have shown that reference selection is a crucial step (Bush et al., 2020): greater genetic distance between the sequenced strain and the reference leads to poorer variant-calling performance. One possible approach to improve the accuracy of genetic analyses of the microbiome is to use metagenomic assembled contigs from the studied metagenome as the reference. For example, Lou et al. (2021) recently used this approach coupled with InStrain variant calling, and it can be applied with any of the variant-calling methodologies we describe here. On the other hand, taxonomic abundance, which is related to the mean coverage of the genome, does influence variant-calling performance. This is especially true in the non-probabilistic methods that rely on hard cutoffs for the number of reads supporting a variant. In practice, this threshold might be optimized according to the bacterial abundance and number of reads, but we used default threshold parameters for the purposes of this work.

We chose to benchmark four commonly used probabilistic variant callers: BCFtools, Mutect2, HaplotypeCaller, and freebayes. We also included VarScan2 because it performs well for pooled samples and in circumstances where probabilistic methods do not work. We also chose to test InStrain and metaSNV as representatives of variant callers developed specifically for metagenomic datasets. Of these tools, freebayes, Mutect2, metaSNV, and InStrain are also able to identify variants from a population of samples, as is the case if several strains coexist, or homologous regions from different taxa align. Variant calling was performed independently in each simulation set and bacteria, as was mapping. However, this might not be ideal for some tools. InStrain, for instance, recommends mapping to a database containing all reference genomes so that multi-mapping reads will be penalized with lower mapping quality. VarScan2 also relies on mapping quality trimming, which penalizes multi-mapping reads.

Our simulations consisted of two scenarios. In the first, only one strain per species was simulated. This might correspond to the real gut metagenomic data, since one major strain dominates the environment in many cases (Truong et al., 2017). In the second scenario, we assumed the existence of two strains with equal abundance per species. Both strains were simulated as only containing SNVs and with no other structural variations, which is an important simplification to consider when looking at our results. Our performance estimates only considered positions covered by reads, and thus, if most variants were missed due to a lack of coverage, we could not consider them. This is a double-edged sword because, in the multi-strain scenario, positions might be covered by one strain that does not contain the variant, and we will thus overestimate the FN fraction compared with the uni-strain scenario. Consistent with this expectation, our sensitivity results here are about half of those achieved in the uni-strain scenario. Nonetheless, the tool comparisons in both scenarios are similar. Most tools achieved high precision, particularly BCFtools and HaplotypeCaller. Both these methods are probabilistic and do not consider population variants, which means that the calls are more restrictive (no multiple alleles are expected in a haploid genome) but have more information to successfully call true variants. On the other hand, freebayes and Mutect2 achieved higher sensitivity, consistent with their ability to detect multiple variants per locus. These results highlight that, while non-probabilistic methods have been developed to deal with the issues associated with MGS variant calling, probabilistic methods can still perform better or similarly, at least when analyzing very abundant bacteria. However, we also show that the performance of non-probabilistic methods declined drastically for lower abundance bacteria. This might highlight the necessity of fine tuning the default threshold values according to the genome size and the number of reads produced. This is especially true for VarScan2, where default values are not tuned for metagenomic calling and resulted in very restrictive cutoffs that reduced the number of calls.

In addition, we also tested the differences in performance of HaplotypeCaller, freebayes, and BCFtools, which all give Phred-score quality values for their variant callers. Our results highlighted that the highest-scoring variants tended not to be TP and might indicate homologous regions with other bacteria. At the same time, only freebayes benefited substantially from quality filtering, which improved its precision as most of the variants found were of very low quality.

Joint variant calling of the uni-strain and multi-strain scenarios improved sensitivity in relation to non-joint variant calling. However, joint variant calling negatively affected the uni-strain results. As it is difficult to assess which situation is most likely to occur in real data and, given the good performance of non-joint variant calling in our simulation, we advocate performing SNV calling per sample instead of joint calling.

Finally, we investigated real gut metagenomic data from the HMP where we did not have certainty about which variants are true or false. However, given the longitudinal sampling of these HMP samples, we could use our variant set to compare samples at baseline and follow-up, assuming that most genetic variants would be stable within 1 year. Here we chose only 10 species for variant calling so that representatives of the different performances in the simulated data were used. Variants were profiled with two tools, HaplotypeCaller, which had the best precision in our benchmark, and Mutect2, which had the best sensitivity. Both tools showed good performance even for low-abundance species. Our results show that both methods we used to call HMP variants produced variant profiles that were closer between samples taken from the same individual at different times than among different individuals. In fact, we were able to demonstrate that this individual specificity is even higher than abundance-based estimations.

Variant-calling errors are expected to arise with lower read depth [due to the relative abundance of a given taxon or systematic bias during sequencing protocols (Browne et al., 2020)], with lower sequencing quality in certain regions [due to inherent sequencing biases that are platform dependent (Ross et al., 2013)], and with wrongly mapped reads (possibly in low-complexity or homologous regions), which have a fundamental role in variant-calling performance. Of these potential sources of bias, we assessed the effect of relative abundance. However, all our simulations follow an Illumina error model, which does not account for genomic features prone to generate sequencing errors, except for errors related to read position. With respect to incorrectly assigned reads, we give an estimate of 36%, but further efforts are needed to assess to what extent these incorrectly assigned reads impact the variant calling results. Furthermore, our simulation assumed that all introduced variants were neutral and occurred by chance and did not take evolutionary forces into consideration. To verify the SNV calling from short-read MGS data, variants might be confirmed with whole-genome sequencing from single-strain isolates.

Overall, this benchmark highlights the efficacy of using probabilistic variant callers on metagenomic data. We recommend using GATK's HaplotypeCaller or Mutect2 depending on concerns about FP (use HaplotypeCaller) or FN (use Mutect2). Both tools seem to perform equally well in real data, where we found a similar power to cluster follow-up samples.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://hmpdacc.org/hmp/.

AUTHOR CONTRIBUTIONS

SA-S, LC, DW, and JF contributed to the project conception. SA-S, LC, DW, and HA ran the tools to complete the project. SA-S, LC, and DW conducted the formal analysis and visualization. SA-S, LC, DW, HA, and JF wrote the study, while AZ edited the manuscript. AZ and JF supervised the completion of the study. All authors contributed to the article and approved the submitted version.

FUNDING

The researchers participating in this project were supported by the Netherlands Heart Foundation (IN-CONTROL CVON grants 2012-03 and 2018-27 to AZ and JF), the Netherlands Organization for Scientific Research (NWO) (NWO-VICI VI.C.202.022 to JF, NWO-VIDI 016.178.056 to AZ, NWO Gravitation Exposome-NL 024.004.017 to JF and AZ, and NWO Gravitation Netherlands Organ-on-Chip Initiative 024.003.001 to JF), the European Research Council (ERC) (ERC Starting Grant 715772 to AZ and ERC Consolidator Grant 101001678 to JF), and the Foundation De Cock-Hadders (grant 20:20-13 to LC). LC also holds a joint fellowship from the University Medical Centre Groningen and China Scholarship Council (CSC201708320268). DW holds a fellowship from the China Scholarship Council (CSC201904910478).

ACKNOWLEDGMENTS

We thank K. Mc Intyre for English editing. We also want to acknowledge the Genomics Coordination Center (GCC) of the University Medical Center Groningen for providing and maintaining the computing infrastructure used in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.648229/full#supplementary-material

Supplementary Figure 1 | Distribution of genome quality. **(A)** Distribution of number of contigs. **(B)** Distribution of N50 score. **(C)** Distribution of auN. **(D)** Distribution of genome length. Vertical line represents mean value.

Supplementary Figure 2 | Dendrogram of Mash-based distance between the 46 reference genomes used in the benchmark. Highlighted taxa are the ones picked for the variant calling in the HMP data.

Supplementary Figure 3 | Uni-strain variant calling statistics of the seven tools on a 4% divergence set-up. Colors indicated the tools. **(A)** Sensitivity (TP/TP + FN) of each tool. Tukey's box plot shows the distribution of precision. Dots show precision per individual bacteria. **(B)** Precision (TP/TP + FP) of each tool. Tukey's box plot shows the distribution of precision. Dots show precision per individual bacteria. Distribution of **(C)** TP, **(D)** FN, and **(E)** FP per tool, shown as Tukey's box plots. Individual dots present bacteria >1.5 times the interquartile distance. **(F)** Precision vs. sensitivity plot. Dots present mean values among all bacteria. Error bars represent the standard deviation from the mean.

Supplementary Figure 4 | Variant Phred quality distribution in each of the 46 analyzed species in BCFtools, HaplotypeCaller and Mutect2.

Supplementary Figure 5 | Two-strain variant calling statistics of the seven tools. Colors indicated the tools. **(A)** Sensitivity (TP/TP + FN) of each tool. Tukey's box plot shows the distribution of precision. Dots show precision per individual bacteria. **(B)** Precision (TP/TP + FP) of each tool. Tukey's box plot shows the distribution of precision. Dots show precision per individual bacteria. **(C)** Precision vs. sensitivity plot. Dots present mean values among all bacteria. Error bars represent the standard deviation from the mean. Distribution of **(D)** TP, **(E)** FN, and **(F)**. FP per tool, shown as Tukey's box plots. Individual dots present bacteria >1.5 times the interquartile distance.

Supplementary Table 1 | Table of species selected from Gupta et al., 2020.

Supplementary Table 2 | Table with species, abundance, reference and genome quality metrics.

Supplementary Table 3 | Variant calling statistics in the uni-strain scenario.

Supplementary Table 4 | Variant calling statistics in the uni-strain scenario including 4% variation from each taxon to the reference. Includes Pearson correlation coefficients between accuracy and sensitivity from the uni-strain scenario between a simulation of 1 and 4% of mutations.

Supplementary Table 4 | Variant calling statistics in the multi-strain scenario.

Supplementary Table 5 | Summary statistics of the effect of genome abundance and coverage on variant calling.

REFERENCES

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Browne, P. D., Nielsen, T. K., Kot, W., Aggerholm, A., Gilbert, M. T. P., Puetz, L., et al. (2020). GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *Gigascience* 9:giaa008. doi: 10.1093/gigascience/giaa008

Bush, S. J., Foster, D., Eyre, D. W., Clark, E. L., De Maio, N., Shaw, L. P., et al. (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *Gigascience* 9:giaa007. doi: 10.1093/gigascience/giaa007

Byrd, A. L., Liu, M., Fujimura, K. E., Lyalina, S., Nagarkar, D. R., Charbit, B., et al. (2021). Gut microbiome stability and dynamics in healthy donors and patients with non-gastrointestinal cancers. *J. Exp. Med.* 218:20200606. doi: 10.1084/jem.20200606

Chen, L., Collij, V., Jaeger, M., van den Munckhof, I. C. L., Vich Vila, A., Kurilshikov, A., et al. (2020). Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat. Commun.* 11:4018.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Costea, P. I., Munch, R., Coelho, L. P., Paoli, L., Sunagawa, S., and Bork, P. (2017). metaSNV: A tool for metagenomic strain level analysis. *PLoS One* 12:e0182392. doi: 10.1371/journal.pone.0182392

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.

Gacesa, R., Kurilshikov, A., Vila, A. V., and Sinha, T. (2020). The Dutch Microbiome Project defines factors that shape the healthy gut microbiome. *bioRxiv*. [preprint].

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*. [preprint].

Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., and Bongcam-Rudloff, E. (2019). Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics* 35, 521–522. doi: 10.1093/bioinformatics/bty630

Gupta, V. K., Kim, M., Bakshi, U., Cunningham, K. Y., Davis, J. M. III, Lazaridis, K. N., et al. (2020). A predictive index for health status using species-level gut microbiome profiling. *Nat. Commun.* 11:4635.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi: 10.1093/bioinformatics/bts480

Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2, 231–239. doi: 10.1016/0888-7543(88)90007-9

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550, 61–66. doi: 10.1038/nature23889

Lou, Y. C., Olm, M. R., Diamond, S., Crits-Christoph, A., Firek, B. A., Baker, R., et al. (2021). Infant gut strain persistence is associated with maternal origin, phylogeny, and functional potential including surface adhesion and iron acquisition. *bioRxiv* 428340. [preprint]. doi: 10.1101/2021.01.26.428340

Olm, M. R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B. A., Morowitz, M. J., and Banfield, J. F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* [preprint]. doi: 10.1038/s41587-020-00797-0

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., et al. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.* 14:R51.

Salosensaari, A., Laitinen, V., Havulinna, A., Meric, G., Cheng, S., Perola, M., et al. (2020). Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota. *medRxiv* [preprint]. doi: 10.1101/2019.12.30.19015842

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., et al. (2013). Genomic variation landscape of the human gut microbiome. *Nature* 493, 45–50. doi: 10.1038/nature11711

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27, 626–638. doi: 10.1101/gr.216242.116

Yoshimura, D., Kajitani, R., Gotoh, Y., Katahira, K., Okuno, M., Ogura, Y., et al. (2019). Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. *Microbial Genomics* 5:261. doi: 10.1099/mgen.0.000261

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 352, 565–569. doi: 10.1126/science.aad3369

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

DATA REPORT
published: 06 August 2021
doi: 10.3389/fgene.2021.704165

Draft Genome Sequence of *Bacillus amyloliquefaciens* Strain CB, a Biological Control Agent and Plant Growth-Promoting Bacterium Isolated From Cotton (*Gossypium* L.) Rhizosphere in Coimbatore, Tamil Nadu, India

Nakkeeran Sevugapperumal[1†], Vimalkumar S. Prajapati[2*†], Vanthana Murugavel[1] and Renukadevi Perumal[1]*

[1] Department of Plant Pathology, Center for Plant Protection Studies, Tamil Nadu Agricultural University, Coimbatore, India, [2] Division of Microbiology and Environmental Biotechnology, Aspee Shakilam Biotechnology Institute, Navsari Agricultural University, Surat, India

Keywords: ***Bacillus amyloliquefaciens*, illumina hi seq, whole genome shotgun sequencing, biocontrol, PGPR, NGS**

OPEN ACCESS

Edited by:
Dhaval K. Acharya, B N Patel Institute of Paramedical, India

Reviewed by:
Nicolas Plaza, Autonomous University of Chile, Chile
Maria Carolina Quecine, University of São Paulo, Brazil
Ramesh K. Kothari, Saurashtra University, India

***Correspondence:**
Nakkeeran Sevugapperumal
nakkeeranayya@tnau.ac.in
Vimalkumar S. Prajapati
vimalprajapati100@gmail.com
orcid.org/0000-0003-4257-1728

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *01 May 2021*
Accepted: *08 July 2021*
Published: *06 August 2021*

Citation:
Sevugapperumal N, Prajapati VS, Murugavel V and Perumal R (2021) Draft Genome Sequence of Bacillus amyloliquefaciens Strain CB, a Biological Control Agent and Plant Growth-Promoting Bacterium Isolated From Cotton (Gossypium L.) Rhizosphere in Coimbatore, Tamil Nadu, India. Front. Genet. 12:704165. doi: 10.3389/fgene.2021.704165

INTRODUCTION

Although rhizobacteria have been widely explored for their plant growth-promoting capabilities and to manage various fungal and bacterial diseases in plants, viral diseases are an ongoing challenge in the agricultural sector (Vinodkumar et al., 2017). As most plant viral diseases are transmitted through vectors, researchers around the globe are utilizing biotechnological approaches to generate resistant lines. Various antagonistic bio-agents contribute to host defense, and various *Bacillus* species have been shown to produce these agents to protect against a wide range of pathogens. Several reports have demonstrated the antiviral efficacy of various *Bacillus* species against the cotton leaf curl virus (Ramzan et al., 2016), the cucumber mosaic virus in tomato (Zehnder et al., 2000), the tomato mottle virus in tomato (Murphy et al., 2000), and the tobacco mosaic virus in tobacco (Wang et al., 2009).

This bacterium is well known for the production of antibacterial, antiviral, and antifungal substances like *Bacillomycin* D, *Surfactin*, and *Bacillaene*, which protect the plant from pathogenic organisms (Chen et al., 2009). Additionally, the proteases and amylases produced by *Bacillus amyloliquefaciens* are used in industrial applications (Prajapati et al., 2015, 2017). *Bacillus* species belonging to this group are reported to have 24 diverse antimicrobial peptide (AMP) genes, which lead to the production of numerous compounds such as iturin, bacilysin, bacillomycin, fengycin, surfactin, mersacidin, ericin, subtilin, subtilosin, and mycosubtilin (Chung et al., 2008; Mora et al., 2011). Moreover, *Bacillus* species synthesize various volatile and non-volatile compounds that synergistically restrict plant diseases (Fernando et al., 2005; Mora et al., 2011). *B. amyloliquefaciens* CB has been used to prevent stem rot of carnations, and it was observed that minimum percentage disease incidence and maximum plant growth promotion occurred in plant treated with isolate CB. Further detailed experimentation will be carried out to evaluate the in-depth potential of the *B. amyloliquefaciens* CB.

Here, we report a draft genome sequence of *B. amyloliquefaciens* strain CB, which was isolated from rhizospheric soil of the cotton plant, collected from a cotton farm on the Tamil Nadu Agricultural University (TNAU) campus in Coimbatore, Tamil Nadu, India. This bacterium is gram positive with long rod-shaped, aerobic motile rods arranged singly or in chains. *B. amyloliquefaciens* belongs to the group of free-living soil bacteria, which aid to suppress plant pathogens and assist in promoting plant growth.

VALUE OF THE DATA

The *B. amyloliquefaciens* CB draft genome can be used as a base/reference sequence to explore and map specific genes related to AMPs and other important enzymes. It could be a valuable resource to conduct comparative analyses among different species related to *B. amyloliquefaciens*, which may have similar biocontrol properties.

METHODS AND DATA ANALYSIS

Bacterial DNA from the CB strain was extracted using phenol-chloroform methodology, and purification was performed using a Genomic DNA Clean and Concentrator (Zymo Research, Irvine, CA, USA). One nanogram of highly purified and good-quality DNA was used for the DNA fragment libraries prepared using a Nextera XT DNA sample preparation kit. Sequencing was performed on (2 × 150 paired-end reads with the Illumina v2 reagent kit) (Illumina, San Diego, CA, USA) an Illumina HiSeq system using the standard protocols described by the manufacturer. In total, 4,623,289 reads were obtained, and quality-based read trimming was done using the Trimmomatic software (version 0.30) (Bolger et al., 2014) followed by quality checking with FastQC (version 3.0) (Andrews, 2010). The genome was assembled using GS De Novo Assembler v. 2.6, ABySS v. 1.5.1, Celera Assembler v. 8.3rc2, Edena v. 3.131028, Megahit v. 1.1.2, SOAPdenovo v. 2.04, Velvet v. 1.2.10, SPAdes v. 3.1.1, and SPAdes v. 3.11.0; and the final assembly was merged using CISA v. 1.3 and submitted to the National Center for Biotechnology Information (NCBI) GenBank having accession no. WODE00000000. The total sequence length has been counted up to be 4,113,229 bp consisting of 11 scaffolds/contigs, a contig N50 of 675,513 with L50 of 3, with the largest contig size of the submitted assembly being 1,050,139. The genome size was estimated to be 4.11 MB with a guanine–cytosine (GC) content of 46.30%. Gene annotation was performed using NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova et al., 2016), which identified 3,847 protein-coding sequences (**Table 1**).

To infer the phylogenetic relationship, all the 119 assemblies (**Supplementary File 1**) of *B. amyloliquefaciens* accessible in the NCBI database were considered for the Bacsort analysis (https://github.com/rrwick/Bacsort) including the *B. amyloliquefaciens* CB. A total of 61 clusters of the considered assemblies were generated (cluster accession provided in **Supplementary File 2**), and FastANI was employed to generate the matrix of all pairwise distance between the clusters. FastANI algorithm

TABLE 1 | Genomic features of *Bacillus amyloliquefaciens* strain CB annotated using National Center for Biotechnology Information—Prokaryotic Genome Annotation Pipeline (NCBI-PGAP) v. 4.10.

Items	Counts
Total genes	4,012
Total CDS	3,911
Coding genes	3,847
Coding CDS	3,847
Genes (RNA)	101
rRNAs	8, 8, 7 (5S, 16S, 23S)
tRNAs	73
ncRNAs	5
Total pseudo genes	64

(https://github.com/ParBLiSS/FastANI) generates pairwise Average Nucleotide Identity (ANI) measurements using the only sequence shared by two assemblies (**Supplementary File 3**), which makes it less swayed due to the accessory genome and produce more accurate trees. The phylogeny tree was created by BIONJ algorithm with bootstrap value of 1,000 to form the generated data and was drawn precisely using Interactive Tree Of Life (iTOL) v5, which is an online tool for the display, annotation, and management of phylogenetic trees (Letunic and Bork, 2021) (**Figure 1**). Out of 61 clusters, two distinct nodes were generated, in which 51 leaves and 10 leaves form a separate group. Fifty-one leaves split into another group having 17 leaves and 34 leaves, which generate other groups consequently as shown in **Figure 1**. The *B. amyloliquefaciens* strain CB forms a separate cluster (44) having branch length 0.00582, while its nearby cluster (40) includes two strains, *B. amyloliquefaciens* X030 and *B. amyloliquefaciens N11* (branch length 0.00456), while the cluster (61) comprises the *B. amyloliquefaciens* strain Jxnu-18 (branch length 0.00520). Clusters 44, 40 and 61 originated from a common node having branch length 0.00141 (**Figure 1**).

The genome of *B. amyloliquefaciens* CB was also mapped to the seed subsystem to obtain the high-quality genome annotation through Rapid Annotation using the Subsystem Technology (RAST; version 2.0) (http://rast.nmpdr.org) (Overbeek et al., 2014). The total 325 subsystem with 29% subsystem coverage resulted for *B. amyloliquefaciens* strain CB through RAST server (**Figure 2**). The present investigation revealed that highest number of the genes was allocated to the subsystem category of amino acids and derivatives (303 genes) followed by carbohydrates (214 genes); protein metabolism (205 genes); cofactors, vitamins, prosthetic groups, and pigments (146 genes); nucleosides and nucleotides (97 genes); dormancy and sporulation (97 genes); cell wall and capsule (82 genes); RNA metabolism (67 genes); DNA metabolism (64 genes); fatty acids, lipids, and isoprenoids (54 genes); stress response (47 genes); motility and chemotaxis (42 genes); membrane transport (42 genes); respiration (41 genes); and virulence, disease, and defense (37 genes). A total of 24 genes were found to be associated with iron acquisition and metabolism as well 24 genes for some other miscellaneous applications. More precisely in the

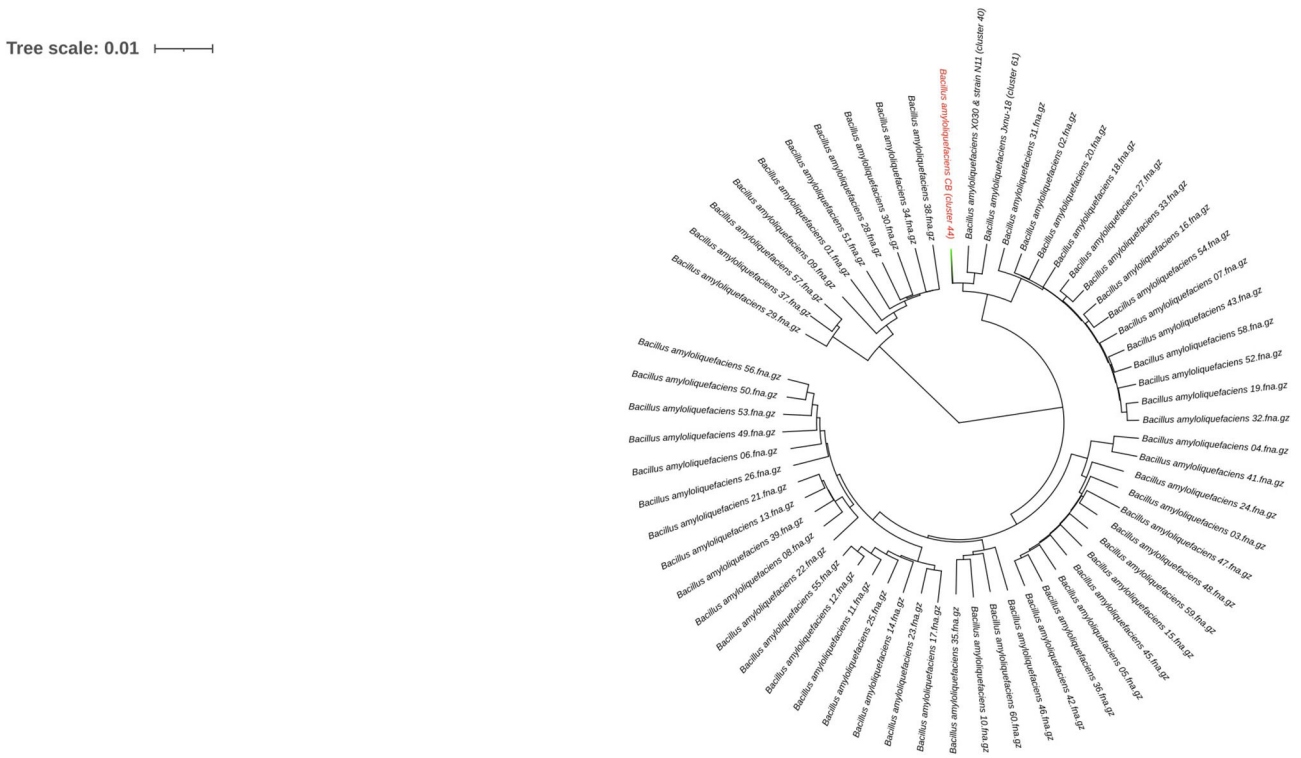

FIGURE 1 | Genome lineage tree of *Bacillus amyloliquefaciens* strain CB. BIONJ algorithm was used to construct the Newick format tree from the distance matrix of all the clustered assemblies (119 different strains of *B. amyloliquefaciens*). The generated Newick tree file was analyzed using the Interactive Tree Of Life (iTOL) v. 5 (strain CB highlighted in red color text with light green color branch).

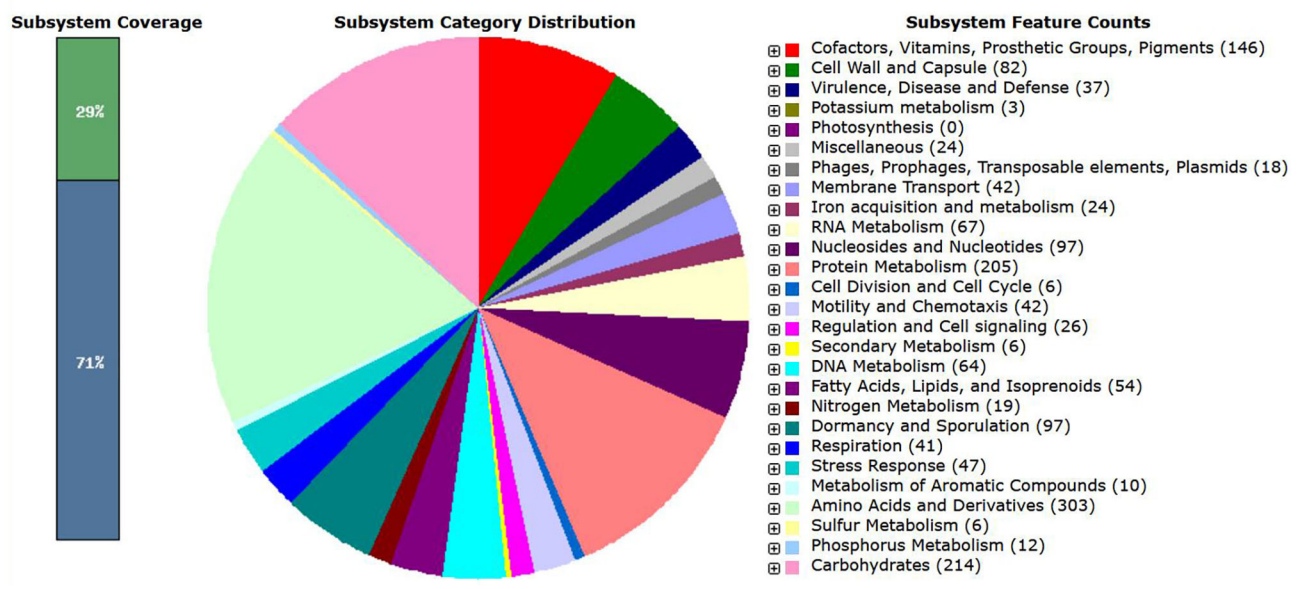

FIGURE 2 | Schematic overview of subsystem coverage, distribution of subsystem category, and subsystem feature counts anticipated in *Bacillus amyloliquefaciens* strain CB using RAST server.

category of miscellaneous application, 10 genes were specifically associated to iron–sulfur cluster assembly, five genes for niacin-choline transport and metabolism, and one gene for single-rhodanese-domain proteins. A total of 26 genes were found to be associated with regulation and cell signaling, while 18 genes were collectively specified for phages, prophages, transposable elements, plasmids, and 12 genes for phosphorus metabolism.

Most of the *Bacillus* spp. belonging to this group of genera have been reported to have antifungal potential and have been utilized for the management of the various fungal diseases; however, their efficacy against viral diseases is still not known (Vinodkumar et al., 2018). The PGAP annotation confirms that the *B. amyloliquefaciens* strain CB genome has gene locus srfAA, srfAD, srfAB, and srfAC, which produce various peptides like surfactin non-ribosomal peptide synthetase and surfactin biosynthesis thioesterase. It has been well documented that lipopeptides like surfactin have acquired more attention due to their high surface activity and antibiotic potential. Moreover, surfactin also possesses antiviral, antitumor, and hemolytic activities (Wang et al., 2010), which required further intensive experimentation for characterization to understand its exact mechanism for such action.

The whole-genome shotgun sequence of *B. amyloliquefaciens* strain CB and its annotation report presented here provide a resource for comparative analysis with other genera of *Bacillus* and can be used for engineering purposes where characteristics of the strain CB are desired. The genome representation of *B. amyloliquefaciens* strain CB showed antagonistic potential due to various AMPs imparting various properties like antifungal, antibacterial, and antiviral as well plant growth promotion, leading to strong future prospects for uplifting the sustainable agriculture.

DATA AVAILABILITY STATEMENT

Bacillus mayloliquefaciens strain CB, whole genome shotgun sequencing project data have been deposited at DDBJ/ENA/GenBank under the accession number WODE00000000. The version described in this data report is the first version having accession number WODE00000000.1. The assembled contigs and its annotation files (CDS, gff, and proteins) are available in https://www.ncbi.nlm.nih.gov/assembly/GCA_011754125.1#/st repository with all the annotations details in Readme file.

AUTHOR CONTRIBUTIONS

NS: funding and modeling the study. VP: genome assembly, annotations, and analysis. VP and NS: manuscript preparation. VM and RP: sampling and sequencing and other miscellaneous stuff. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We all authors are acknowledging Dr. Anthony E. Zamora, Assistant Professor [Medicine (Hematology and Oncology) and Microbiology; Immunology], Medical College of Wisconsin, USA for proof reading and correcting the manuscript for its language competence.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.704165/full#supplementary-material

Supplementary File 1 | List of the all the genomes/assemblies of *B. amyloliquefacines* considered in the presented study (strain, biosample and bioproject infromation, assembly level, GC%, Scafold, CDS).

Supplementary File 2 | Cluster assemblies: Cluster of assemblies were generated with removing the redundancy of assemblies. Similar assemblies are forming one cluster and have only a single representative chosen based on assembly N50 (Cluster_accessions file lists the cluster name, followed by a tab, followed by a comma-delimited list of the assemblies in that cluster, with the representative assembly marked with a *).

Supplementary File 3 | Distance matrix: FastANI produces pairwise ANI measurements of all the generated assemblies's cluster using only the sequence shared by two assemblies.

REFERENCES

Andrews, S. (2010). *Fastqc- a Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bolger, Anthony M., Marc, Lohse., and Bjoern, Usadel. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–212. doi: 10.1093/bioinformatics/btu170

Chen, X.H., Koumoutsi, A., Scholz, R., Schneider, K., Vater, J., Süssmuth, R., et al. (2009). Genome analysis of *Bacillus amyloliquefaciens* FZB42 reveals its potential for biocontrol of plant pathogens. *J. Biotechnol.* 140, 27–37. doi: 10.1016/j.jbiotec.2008.10.011

Chung, S., Kong, H., Buyer, J., Lakshman, D.K., Lydon, J., Kim, S.D. et al. (2008). Isolation and partial characterization of *Bacillus* subtilis ME488 for suppression of soil borne pathogens of cucumber and pepper. *Appl. Microbiol. Biotechnol.* 80, 115–123. doi: 10.1007/s00253-008-1520-4

Fernando, W. G. D., Ramarathnam, R., and Krishnamoorthy, A. S. (2005). Identification and use of potential bacterial organic antifungal volatiles in biocontrol. *Soil Biol. Biochem*. 37, 955–964. doi: 10.1016/j.soilbio.2004.10.021

Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, 293–296. doi: 10.1093/nar/gkab301

Mora, I., Cabrefiga, J., and Montesinos, E. (2011). Antimicrobial peptide genes in *Bacillus* strains from plant environments. *Int. Microbiol*. 14, 213–223. doi: 10.2436/20.1501.01.151

Murphy, J.F., Zehnder, G.W., Schuster, D.J., Sikora, E.J., Polston, J.E., and Kloepper, J.W. (2000). Plant growth-promoting rhizobacterial mediated protection in tomato against tomato mottle virus. *Plant Dis*. 84, 779–784. doi: 10.1094/PDIS.2000.84.7.779

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., et al. (2014). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res*. 42, 206–214. doi: 10.1093/nar/gkt1226

Prajapati, V.S., Ray, S., Narayan, J., Joshi, C.G., Patel, K.C., Trivedi, U.B., et al. (2017). Draft genome sequence of a thermostable, alkaliphilic α-amylase and

protease producing *Bacillus* amyloliquefaciens strain KCP2. 3. *Biotech* 7:372. doi: 10.1007/s13205-017-1005-1

Prajapati, V.S., Trivedi, U.B., and Patel, K.C. (2015). A statistical approach for the production of thermostable and alklophilic alpha-amylase from *Bacillus amyloliquefaciens* KCP2 under solid-state fermentation. *Biotech* 5, 211–220. doi: 10.1007/s13205-014-0213-1

Ramzan, M., Tabassum, B., Nasir, I.A., Khan, A., Tariq, M., Awan, M.F., et al. (2016). Identification and application of biocontrol agents against 147 cotton leaf curl virus disease in *Gossypium hirsutum* under greenhouse conditions. *Biotechnol. Equip.* 30, 469–478. doi: 10.1080/13102818.2016.1148634

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 44, 6614–6624. doi: 10.1093/nar/gkw569

Vinodkumar, S., Nakkeeran, S., Renukadevi, P., and Malathi, V.G. (2017). Biocontrol potentials of antimicrobial peptide producing *Bacillus* species: multifaceted antagonists for the management of stem rot of carnation caused by Sclerotinia sclerotiorum. *Front. Microbiol.* 8:446. doi: 10.3389/fmicb.2017.00446

Vinodkumar, S., Nakkeeran, S., Renukadevi, P., and Mohankumar, S. (2018). Diversity and antiviral potential of rhizospheric and endophytic *Bacillus* species and phyto-antiviral principles against tobacco streak virus in cotton. *Agric. Ecosyst. Environ.* 267, 42–51. doi: 10.1016/j.agee.2018.08.008

Wang, S., Huijun, W., Junqing, Q., Lingli, M., Jun, L., Yanfei, X., et al. (2009). Molecular mechanism of plant growth promotion and induced systemic resistance to tobacco mosaic virus by *Bacillus* spp. *J. Microbiol. Biotechnol.* 19, 1250–1258. doi: 10.4014/jmb.0901.008

Wang, Yu., Zhaoxin, Lu., Xiaomei, Bie., and Fengxia, L.V. (2010). Separation and extraction of antimicrobial lipopeptides produced by *Bacillus amyloliquefaciens* ES-2 with macroporous resin. *Eur. Food Res. Technol.* 231, 189–196. doi: 10.1007/s00217-010-1271-1

Zehnder, G.W., Yao, C., Murphy, J.F., Sikora, E.R., and Kloepper, J.W. (2000). Induction of resistance in tomato against cucumber mosaic cucumovirus by plant growth-promoting rhizobacteria. *Biocontrol* 45 (1), 127–137. doi: 10.1023/A:1009923702103

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 September 2021
doi: 10.3389/fgene.2021.724217

Comparative Genome Analysis of *Bacillus amyloliquefaciens* Focusing on Phylogenomics, Functional Traits, and Prevalence of Antimicrobial and Virulence Genes

Hualin Liu[1†], Vimalkumar Prajapati[2†], Shobha Prajapati[3], Harsh Bais[4] and Jianguo Lu[1,5]*

[1] School of Marine Sciences, Sun Yat-sen University, Zhuhai, China, [2] Division of Microbiology and Environmental, Biotechnology, Aspee Shakilam Biotechnology Institute, Navsari Agricultural University, Surat, India, [3] SVP-A School of Sardar Vallabhbhai National Institute of Technology, Surat, India, [4] Delaware Biotechnology Institute, University of Delaware, Newark, DE, United States, [5] Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, China

OPEN ACCESS

Edited by:
Saumya Patel,
Gujarat University, India

Reviewed by:
Archna Suman,
Indian Agricultural Research Institute (ICAR), India
Ravi Shah,
Medical College of Wisconsin, United States
Liming Wu,
Nanjing Agricultural University, China

***Correspondence:**
Vimalkumar Prajapati
vimalprajapati@nau.in
orcid.org/0000-0003-4257-1728

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *12 June 2021*
Accepted: *26 August 2021*
Published: *30 September 2021*

Citation:
Liu H, Prajapati V, Prajapati S, Bais H and Lu J (2021) Comparative Genome Analysis of Bacillus amyloliquefaciens Focusing on Phylogenomics, Functional Traits, and Prevalence of Antimicrobial and Virulence Genes. Front. Genet. 12:724217. doi: 10.3389/fgene.2021.724217

Bacillus amyloliquefaciens is a gram-positive, nonpathogenic, endospore-forming, member of a group of free-living soil bacteria with a variety of traits including plant growth promotion, production of antifungal and antibacterial metabolites, and production of industrially important enzymes. We have attempted to reconstruct the biogeographical structure according to functional traits and the evolutionary lineage of *B. amyloliquefaciens* using comparative genomics analysis. All the available 96 genomes of *B. amyloliquefaciens* strains were curated from the NCBI genome database, having a variety of important functionalities in all sectors keeping a high focus on agricultural aspects. In-depth analysis was carried out to deduce the orthologous gene groups and whole-genome similarity. Pan genome analysis revealed that shell genes, soft core genes, core genes, and cloud genes comprise 17.09, 5.48, 8.96, and 68.47%, respectively, which demonstrates that genomes are very different in the gene content. It also indicates that the strains may have flexible environmental adaptability or versatile functions. Phylogenetic analysis showed that *B. amyloliquefaciens* is divided into two clades, and clade 2 is further dived into two different clusters. This reflects the difference in the sequence similarity and diversification that happened in the *B. amyloliquefaciens* genome. The majority of plant-associated strains of *B. amyloliquefaciens* were grouped in clade 2 (73 strains), while food-associated strains were in clade 1 (23 strains). Genome mining has been adopted to deduce antimicrobial resistance and virulence genes and their prevalence among all strains. The genes *tmrB* and *yuaB* codes for tunicamycin resistance protein and hydrophobic coat forming protein only exist in clade 2, while *clpP*, which codes for serine proteases, is only in clade 1. Genome plasticity of all strains of *B. amyloliquefaciens* reflects their adaption to different niches.

Keywords: *B. amyloliquefaciens*, phylogenomics, genome evaluation, comparative genomics, functional traits, antimicrobial resistance and virulence genes

INTRODUCTION

Since the 19th century, Bacilli is one of the most well-documented and preeminently characterized bacterial genera comprising classical microbiology, biochemistry, and advanced genomic and proteomic approaches (Alcaraz et al., 2010). Among the various species of bacilli, *Bacillus amyloliquefaciens* gains lots of research interest and has wide application in agriculture, pharmaceuticals, food industry, environmental industry, etc. (Sharma and Satyanarayana, 2013). Various strains of *B. amyloliquefaciens* are common habitants and frequently screened from various ecological niches, including soil, animal feces, human food, aquatic environments, and many more, reflecting its versatile metabolic capabilities (Earl et al., 2008). During evolution, the bacterial population acclimatized to their respective ecological niches, which lead to the differentiation as evidenced by various genomic and physiological characteristics (De Wit et al., 2012).

Versatility of nature and metabolic competencies of different strains of *B. amyloliquefaciens* provoke to expedite the comparative genomic analysis to address more in detail the life style of bacteria, their adaptation to various niches and how they overcome contenders, as well as to catch clear revelation on their biochemistry, physiology, and genetics (Sharma and Satyanarayana, 2013; Owusu-Darko et al., 2020). *B. amyloliquefaciens* have been known to promote plant growth via a variety of mechanisms (Baghaee Ravari and Heidarzadeh, 2014; Shao et al., 2015; Liu et al., 2016), act as biocontrol against numerous plant diseases caused by soil-borne microorganisms (Tan et al., 2016), be widely used as biofertilizers and biopesticides (Wu et al., 2015), antagonize plant pathogens by competing essential nutrient (Wu et al., 2016), produce antibiotic compounds (Srivastava et al., 2016), as well induce systemic acquired resistance (Ng et al., 2016). Moreover, it is well documented that *B. amyloliquefaciens* can be tailored for numerous environmental and industrial applications such as degradation of crude oil from oil-contaminated sites (Zhang J. et al., 2016) and feather degradation (Yang et al., 2016); can produce various enzymes like proteases (Wang et al., 2016), feruloyl esterase (Wang et al., 2017), phytase (Verma et al., 2016), and amylases (Prajapati et al., 2015); and can be employed as a biosorbent for the removal of pollutants (Sun et al., 2016) and their degradation (Zühlke et al., 2016), production of biosurfactant and AMPs, probiotics, etc. (Perez et al., 2017).

The number of bacterial genome sequences has almost doubled over the decades due to the decreasing cost of the sequencing with advancement in high-throughput sequencing technology. The generated sequences data are available freely in the public domain, which ultimately stimulate researchers to do more on genomic analysis. Comparative genome analysis always sharpens our understanding of the bacterial genome structure and its diversity at a particular niche. Moreover, the pan-genome of species includes analysis of all core genes, dispensable genes, and strain-specific genes, which need to be comprehensively investigated as they reveal the essential functions for the species or laterally transferred functions in specific strains (Vernikos et al., 2015). *Bacillus* is one of the most extensively studied species with prevalent sets of genome sequences to date; however, very few reports are available on core genes and strain-specific genes in the *Bacillus* species (Alcaraz et al., 2010). Kim et al. (2017) have reported the core gene data of multiple *Bacillus* species through pan-genome analysis to explore the *Bacillus* species in food microbiome.

In the present investigation, we have curated all the 96 genome sequences of *B. amyloliquefaciens* available in the NCBI database to carry out comparative genomic analysis. Based on contextual information, we were trying to understand the distribution of all strains of *B. amyloliquefaciens* with respect to their ecological niches and their source of isolation and location to get better insights into their phylogenetic position using the core genome. PAN genome analysis of all strains of *B. amyloliquefaciens* was conducted to acquire better impression on their functional difference, which affects their dynamic evolutionary processes. We were also interested in understanding the comparative account of various antimicrobial and virulence genes presented among all *B. amyloliquefaciens* strains. The consensus information and conclusion drawn from this presented comparative genomic study can be used as a benchmark for designing wet-lab experimentation and validation as well as to formulate new hypothesis.

MATERIALS AND METHODS

In total, 96 genome sequences of *B. amyloliquefaciens* having an N50 size greater than 50 k were downloaded from the NCBI database (detailed in **Supplementary Table 1**). Pan-genome analysis was conducted by Roary (Page et al., 2015) embedded in the "Pan" module of PGCGAP v1.0.21 (Liu et al., 2020). Single-copy core proteins calling, alignment of sequences, sequences concatenating, best model chosen, and phylogenetic tree constructing were performed with the "CoreTree" module of PGCGAP v1.0.21. The pairwise similarity of genomes was calculated by Mash (Ondov et al., 2016) embedded in the module "MASH" of PGCGAP v1.0.21. COG annotation was performed with the module "pCOG" of PGCGAP v1.0.21 (Liu et al., 2020). The antimicrobial resistance and virulence genes were mined against the databases of argannot (Gupta et al., 2014), card (Jia et al., 2017), NCBI (Feldgarden et al., 2019), resfinder (Zankari et al., 2012), vfdb (Chen et al., 2016), and EcOH (Ingle et al., 2016) by the module "AntiRes" of PGCGAP v1.0.21 (Liu et al., 2020).

RESULTS

A total of 16,198 gene clusters were found by pan-genome analysis, of which 1,448 (8.95%) are single-copy and code for core proteins. Shell genes, soft-core genes, core genes, and cloud genes comprise 17.09, 5.48, 8.96, and 68.47%, respectively, which demonstrates that the genomes are very different in the gene content (**Supplementary Table 2**). The pan-genome curve shows that the number of total genes increased with the increase in the genome number; this indicates that *B. amyloliquefaciens* has an open pan-genome (**Figure 1**).

The evolutionary relationship between the 96 *B. amyloliquefaciens* strains was investigated by the construction

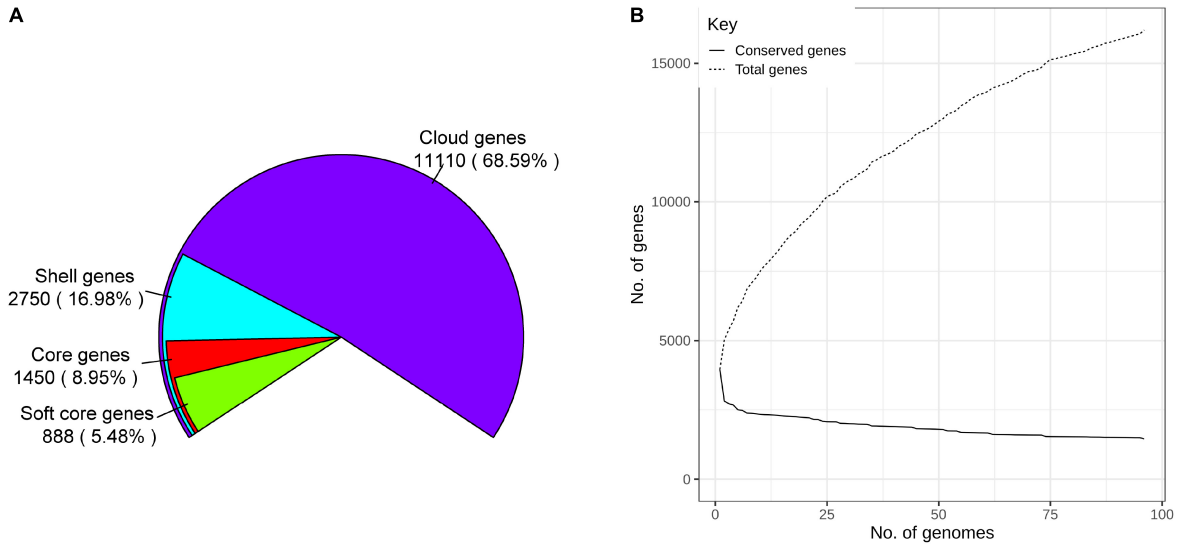

FIGURE 1 | Pan-genome of 96 *B. amyloliquefaciens*. **(A)** A pie chart of the breakdown of genes (core: 99% ≤ strains ≤ 100%, soft-core: 95% ≤ strains < 99%, shell: 15% ≤ strains < 95%, and cloud: 0% ≤ strains < 15%). **(B)** The pan-genome curve of the 96 genomes.

of a phylogenetic tree based on the alignment sequences of 1,154 concatenation core proteins (**Figure 2**). *Bacillus pumilus* SAFR-032 (Gioia et al., 2007) was used as the outgroup. The strains are divided into two clades, and clade 2 consists of two clusters. The location where the strain was isolated was mapped outside of the tree as the color strip. Strains from America are mainly located in cluster 2 of clade 2, while strains from Asia and Europe are scattered in all clades. The isolation source of the strain was also marked on the tree. According to known information, almost all the plant-associated strains are located in clade 2, and strains isolated from food are mainly located in clade 1. The above result implies that *B. amyloliquefaciens* has differentiated mainly into plant-associated and food-associated, as it clearly showed in the clades. However, some species of *B. amyloliquefaciens* isolated from water, soil, etc. are scattered in clade 2.

The similarity of genome pairs has been compared within and between clades and clusters (**Figure 3**). Genomes in clade 1 are found to be more similar than those that are observed in clade 2 ($p < 0.001$), while the similarity between genomes of the two clades is found to be very low, which indicates that strains in clade 2 undergone more differentiation, which may be related to their adaption to specific plants and other associated niches. When focusing on clade 2, genomes in cluster 1 are more similar than genomes in cluster 2 ($p < 0.001$), and the genome similarity between the two clusters is seen to be relatively low. Comparison of the genome size between both clades and its associated cluster has been carried out and depicted in **Figure 3B**. It has been observed that the genome size of clade 2 is slightly greater than that of clade 1, while the GC% content of clade 2 is significantly greater than that of clade 1 ($p < 0.001$; **Figure 3C**).

Compared with the genomes of clade 2, the genomes of clade 1 have a unique gene composition (**Figure 4A**). It was observed that all the species in clade I have lost 335 genes (**Supplementary Table 2** lines 2,592–2,926), which exists in all the genomes of clade 2 and have 490 unique core genes (**Supplementary Table 2** lines 3,969–4,458). To reveal the difference of gene contents between the two clades, the gene family analysis has been performed with module "OrthoF" of PGCGAP v1.0.21. A total of 9,245 orthogroups are found, out of which 4,872 orthogroups are observed to be common between the two clades, while 1,055 are unique to clade 1, and the remaining 3,363 are unique to clade 2. The functions of these unique orthogroups are revealed through COG annotation as shown in **Figure 4B**. The relative abundance of functional classes I (lipid transport and metabolism), G (carbohydrate transport and metabolism), and Q (secondary metabolites biosynthesis, transport, and catabolism) is found to be higher in clade 2 compared to that in clade 1, while the relative abundance of classes D (cell cycle control, cell division, chromosome partitioning), E (amino acid transport and metabolism), H (coenzyme transport and metabolism), L (replication, recombination, and repair), M (cell wall/membrane/envelope biogenesis), and X (Mobilome: prophages, transposons) is higher in clade 1 than that in clade 2 (**Figure 4B**).

It is well documented that antimicrobial resistance and virulence genes are disseminated in the environment according to the function of the respective ecological niche; therefore, we have investigated the distribution of these genes in *B. amyloliquefaciens*. The antimicrobial resistance and virulence genes from different databases have been mined and mapped on the phylogenetic tree (**Figure 5**). To demonstrate the topological structure of the tree more clearly, the outgroup strain has been removed and the tree presented on midpoint rooted. All strains of *B. amyloliquefaciens* including those from foods contain more than one virulence factor. It is observed that *tmrB* and *yuaB* are only existing in clade 2, while *clpP* is prevailing only in

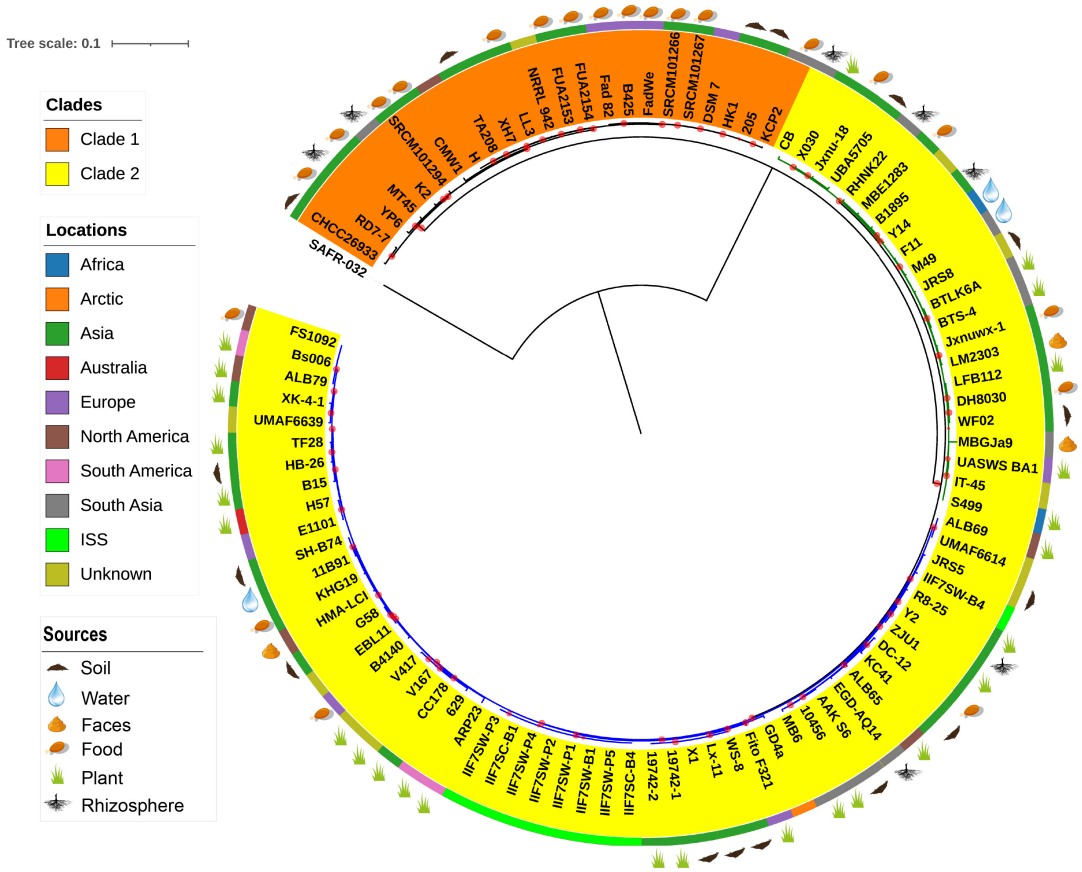

FIGURE 2 | The phylogenetic tree of single-copy core proteins. The maximum-likelihood tree was constructed by the module "CoreTree" of PGCGAP v1.0.21 (Liu et al., 2020) with the best fit model JTT+F+R4. The tree was rooted against *Bacillus pumilus* SAFR-032 as an appropriate outgroup. Red circles on the branch represent bootstrap values larger than 80%. The background color of the labels (colored ranges) represents the clades, and the green and blue branches represent cluster 1 and cluster 2 of clade 2, respectively. The color strip outside the tree describes where the strain was isolated; ISS means International Space Station. The cartoons outside the tree indicate that the strain was isolated from soil, food, rhizosphere, water, or faces of herbivores, or is associated with the plant.

clade 1. The gene *tmrB* is intending an ATP-binding tunicamycin resistance protein found in *B. subtilis* (Noda et al., 1995), while *yuaB* can form a highly hydrophobic coat around *B. subtilis* biofilms (Kobayashi and Iwano, 2012). The gene *clpP* codes for a serine protease involved in proteolysis and is required for growth under stress conditions (Gaillot et al., 2000, 2001). Interestingly, the *B. amyloliquefaciens* strain MBGJa9 has more virulence factors than other strains, and it is seen that *isdA*, *isdB*, *isdC*, *isdD*, *isdE*, *isdF*, *isdG*, and *srtB* form a gene cluster, whose productions participated in the uptake of iron and heme (Skaar and Schneewind, 2004; Skaar et al., 2004).

DISCUSSION

Pan-Genome Assessment of *Bacillus amyloliquefaciens*

Present investigation using the 96 strains of *B. amyloliquefaciens* revealed that it has an extensive pan-genome, and it represents an ample number of genes that were observed to be uniquely associated with each of the divergent species. Population size and respective ecological niche versatility of *B. amyloliquefaciens* are considered to be the most influential factors in determining the pan-genome size, and it can be seen that total genes against the total number of genome sequences are edified up so it is impossible to envisage the size of the full pan-genome. The resulted open pan-genome of *B. amyloliquefaciens* is unsurprising because of the source of isolation and its geographical location, which always adds up new genes to the entire gene pool of species. This species divergence could be an attribute of different mechanisms such as horizontal transfer, transposition, and transformation (Konstantinidis and Tiedje, 2005; Tettelin et al., 2005). On the contrary, observed few core genes in the investigation might be due to the higher number of genomes, the incorporation of genomes from other genera, as well as the inclusion of draft genomes in the data set (Lefebure et al., 2010; Inglin et al., 2018). It is well documented that incomplete, unfinished, or partially assembled genomes have a large impact on the occurrence of core genomes in the analysis as core genomes seem to be very sensitive to the heterogenous data set and poor quality sequences (Mendes-Soares et al., 2014).

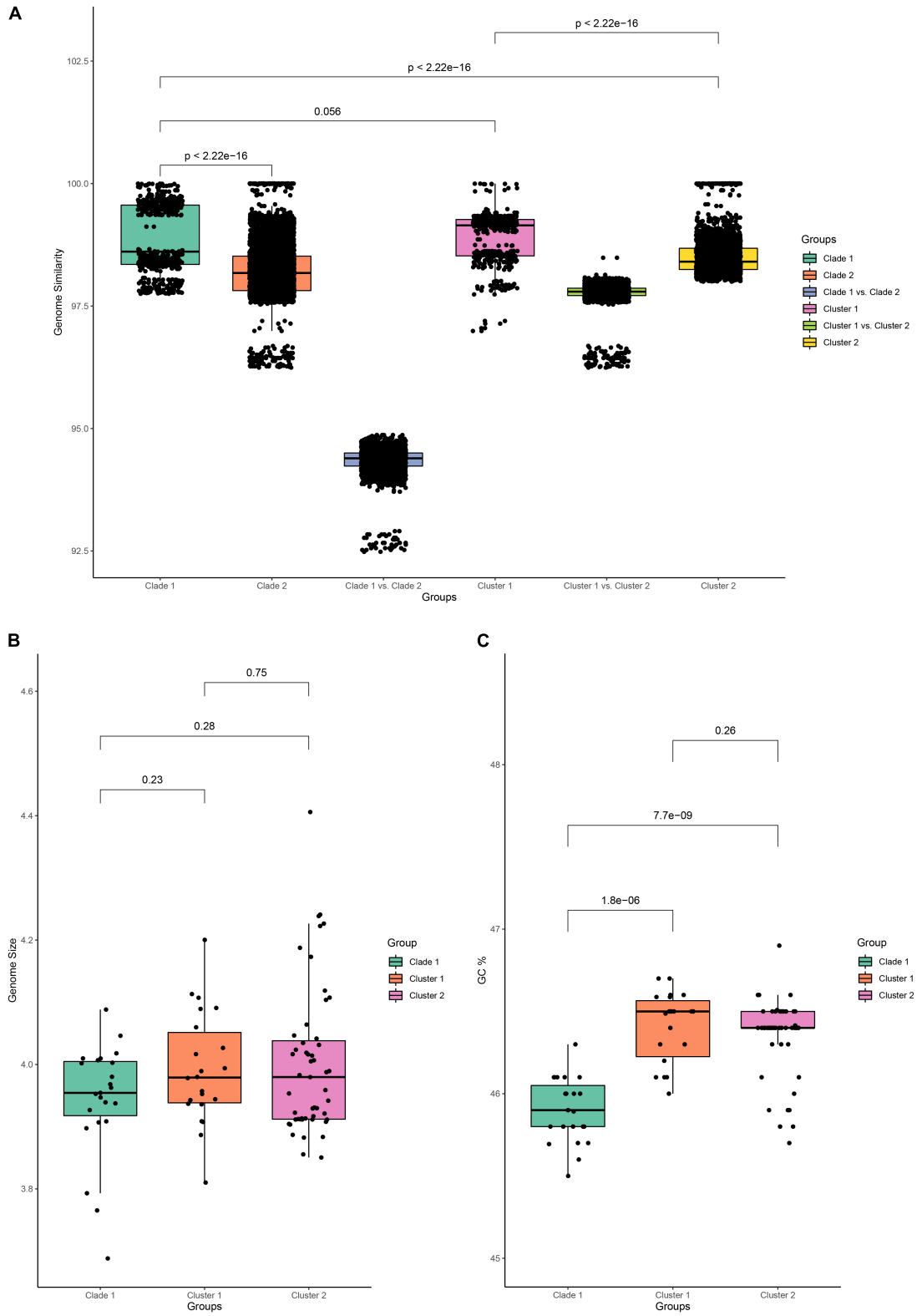

FIGURE 3 | Genome feature of *B. amyloliquefaciens*. **(A)** Genome similarity between all pairs of strains in clade 1, between all pairs of strains in clade 2, between strains in clade 1 and those in clade 2, between strains in cluster 1, between strains in cluster 2, and between strains in cluster 1 and those in cluster 2. **(B)** Genome size of clade 1 and cluster 1 and cluster 2 of clade 2. Wilcox test was performed and marked on top of the box plot. **(C)** GC percent of clade 1, cluster 1, and cluster 2 of clade 2. Wilcox test was performed and the *p*-value was marked on top of the box plot.

The Alliance of *B. amyloliquefaciens* Strains Through Phylogenomics Using Single-Copy Core Proteins and Genomic Comparison: An Evolutionary Assessment

16s rRNA has been widely used for the taxonomy assessment of prokaryotes and has served as the broad context, though the better taxonomic resolution of the microbial species is achieved through "polyphasic approach" and is highly effective (Rosselló-Mora and Amann, 2001; Na et al., 2018). 16s rRNA has limitations as it hampers the phylogenic resolution at the species or subspecies level. The application of genome sequences is highly recommended for the taxonomic understanding of microbial species instead of routinely used DNA–DNA hybridization and 16s rRNA phylogeny (Chun et al., 2018). Therefore, instead of a single gene, genome-based phylogeny called phylogenomics has set up better taxonomic positioning as it uses sets of core

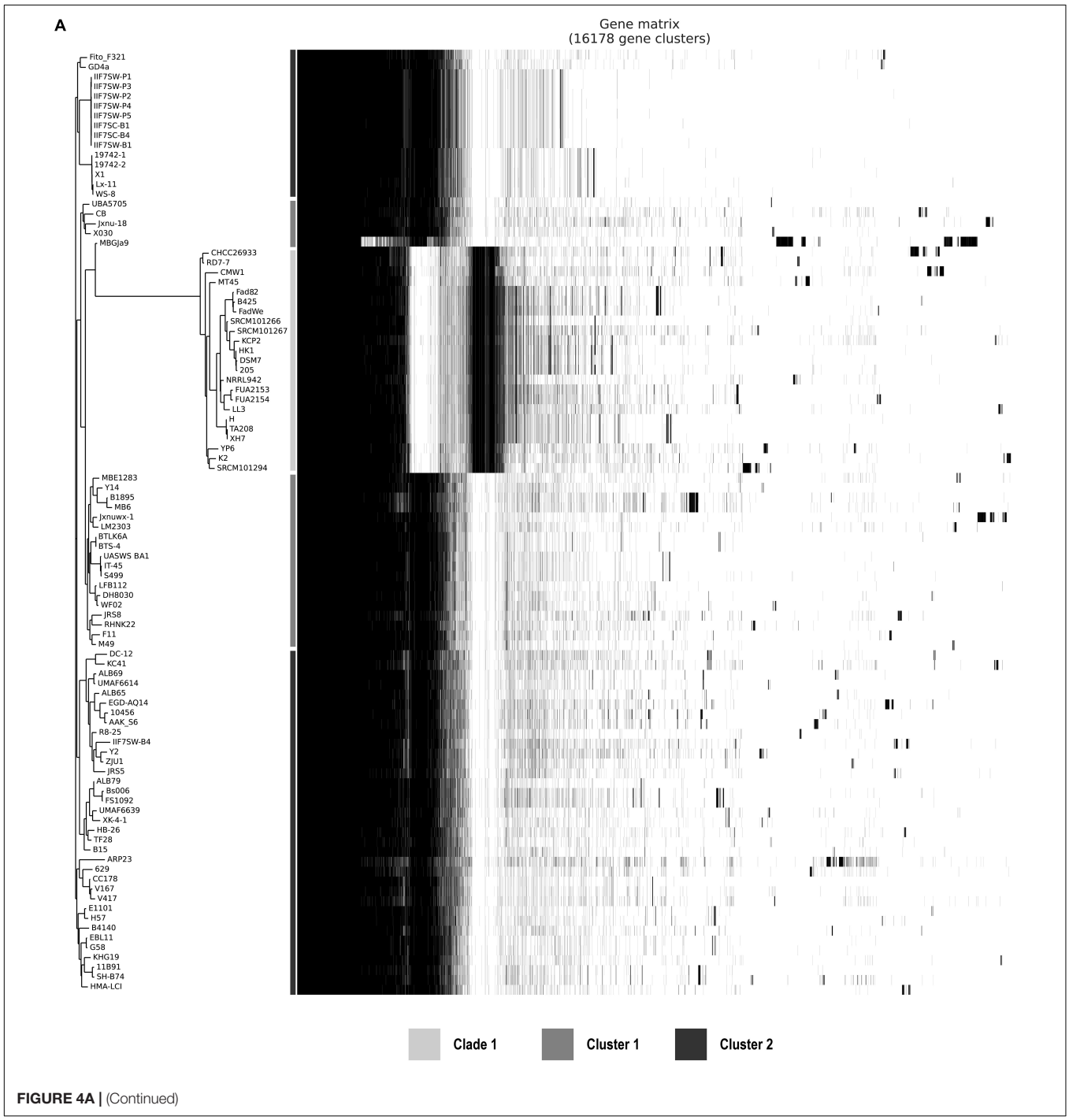

FIGURE 4A | (Continued)

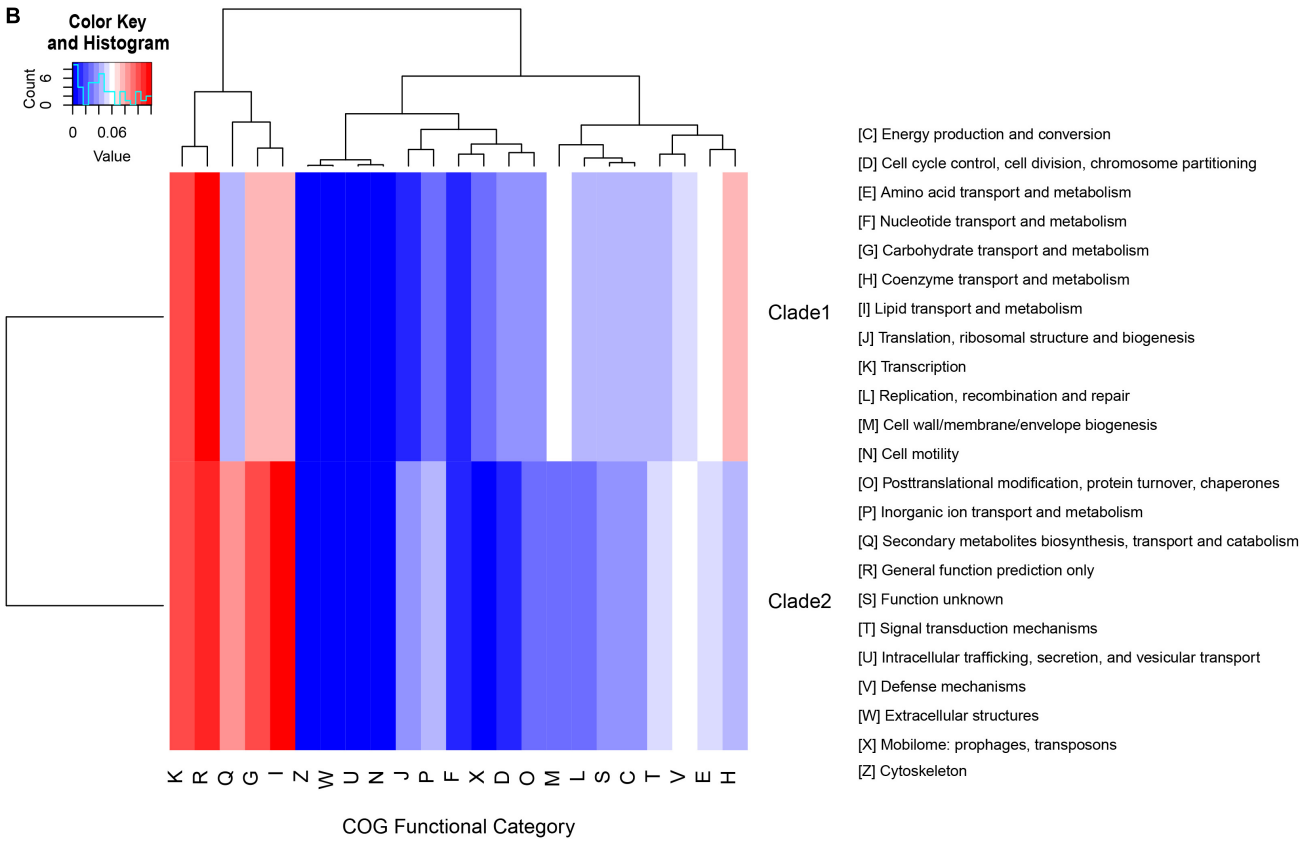

FIGURE 4B | Comparison of gene families between clade 1 and clade 2. The plot shows the tree compared to a matrix with the presence and absence of core and accessory genes **(A)**; the bars between the tree and the matrix show to which clade/cluster the strain belongs. The heat map shows the relative abundance of the function classes corresponding to the unique orthogroups of clade 1 and clade 2 **(B)**.

genes (Eisen and Fraser, 2003). The genome sequences of all *B. amyloliquefaciens* strains are accessible in the Gene Bank NCBI database, which allows us to determine the degree of genome variability among all species as well as distinct out the taxonomic validity of all the isolates and reconstruct their phylogenetic relationship. Two distinct clades were observed when phylogeny was inferred using single copy core protein. Clade 1 comprises 23 strains of *B. amyloliquefaciens*, out of which 56% were food-associated, 17.39% were from soil, and 8.69% were rhizospheric. Clade 2 comprises 73 strains, and it is distinguished into two different clusters, where clusters 1 and 2 comprise 22 and 51 strains of *B. amyloliquefaciens*, respectively. Clade 2 was more enriched with the species of plant origin/host and comprised ∼35.61%, while the strains of soil, food, indoor biome, and rhizosphere origin were 16.43, 10.95, 12.32, and 6.84%, respectively. Two distinct clades were demarcated, one of which was food-associated (clade 1) and the other one plant-associated (clade 2). The selection of core gene sets for accurate phylogeny analysis may vary with the availability of the genome sequences at the time of analysis.

Comparison of the genome similarity between the strains of both clades indicates that the strains grouped together in clade 1 are more similar than those of clade 2. The majority of the plant-associated strains of *B. amyloliquefaciens* are grouped under clade 2, while nonplant-associated strains are mainly found in clade 1, though some scattering is seen with respect to some other ecological niches. Plant-associated strains of *B. amyloliquefaciens* have adopted more modification in their genome, which is directly related to their adaption to the specific plant. Hence, it is believed that the genome size of the plant-associated strains of *B. amyloliquefaciens* is always greater than that of the nonplant-associated *B. amyloliquefaciens* and so the GC % content. Zhang N. et al. (2016) reported that the core genomes of the plant-associated strains of *B. amyloliquefaciens* have more gene contents related to the intermediary metabolism and secondary metabolite biosynthesis as compared to those of nonplant-associated strains. Plant-associated strains also possess specific genes for the synthesis of antibiotics as well as for the utilization of plant-derived substrates.

During the assessment of the core and accessory genes, it was observed that the strains of *B. amyloliquefaciens* grouped in clade 1 have lost many genes that are present in the strains of clade 2 (**Figure 4A**). Exopolysaccharides (EPSs) play very important role in bacteria, specifically those that are plant-associated and have a variety of functions. It helps microorganisms in adherence, pathogenesis, and symbiosis as well as protects from

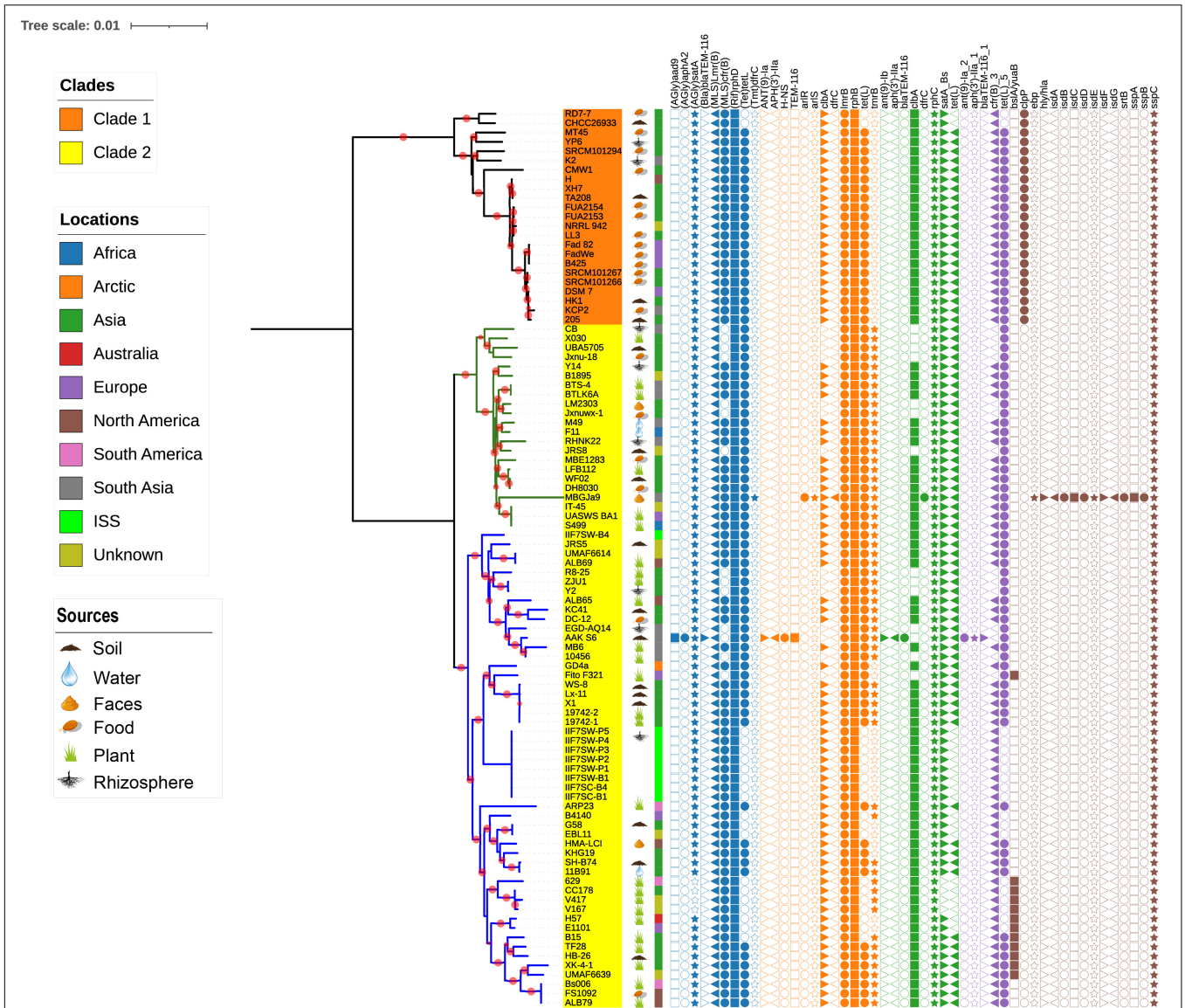

FIGURE 5 | The midpoint rooted phylogenetic tree of single-copy core proteins. The tree was constructed by the module "CoreTree" of PGCGAP v1.0.21 (Liu et al., 2020) with the best fit model JTT+F+R4. Red circles on the branch represent bootstrap values larger than 80%. The background color of the labels represents the clades, and the green and blue branches represent cluster 1 and cluster 2 of clade 2, respectively. The cartoons after the strain name indicate that the strain was isolated from soil, food, rhizosphere, water, or faces of herbivores, or is associated with the plant. The color strip outside the tree describes where the strain was isolated; ISS means International Space Station. The symbols on the right side of the tree represent antibiotic resistance genes or virulence genes from each database [blue: argannot (Gupta et al., 2014), orange: card (Jia et al., 2017), green: NCBI (Michael Feldgarden et al., 2019), purple: resfinder (Zankari et al., 2012), and gray: vfdb (Chen et al., 2016)].

desiccation in some adverse condition (Stingele et al., 1999). The glycosyltransferase gene region comprises the EPS gene cluster, i.e., *epsF-2, epsD, epsI, epsM, epsL,* and *epsJ*, which are involved in the biosynthesis of EPS, and has a profound role in plant-associated strains of *B. amyloliquefaciens*, while it was missing in the strains belonging to clade 1. Plant-associated *B. amyloliquefaciens* strains (clade 2) harbor a certain gene cluster absent in clade 1, which is involved in the biosynthesis of lipopeptides through nonribosomal peptide synthetases (NRPS) including fengycin (*fen*). Gene clusters involved in the synthesis of bacillaene (*bae*) are responsible for the profound antimicrobial activity and are lost in all strains of *B. amyloliquefaciens* in clade 1. The PKS gene cluster, which includes *pksI_2, pksG_2, pksN_2,* and *pksS*, was also found to be present in clade 2 but lost in clade 1 (**Figure 4A**). Some of the genes such as cystathionine beta-lyase (*patB*), putative multidrug resistance ABC transporter ATP-binding/permease protein (*yheI*), cold shock protein (*cspC*), spermidine/spermine N(1)-acetyltransferase (*paiA*), putative sugar phosphate isomerase (*ywlF*), 3-dehydroshikimate dehydratase (asbF),

putative ABC transporter substrate-binding lipoprotein (yhfQ), sirohydrochlorin ferrochelatase (sirB), putative metallo-hydrolase (yflN), dipeptidyl-peptidase 5 (ddp5), L-aspartate oxidase (nadB), putative sporulation hydrolase (cotR), stress response kinase A (srkA), sortase D (srtD), ATP-dependent dethiobiotin synthetase (bioD 1), glycerophosphodiester phosphodiesterase (glypQ), folylpolyglutamate synthase (fpgS), and putative ABC transporter permease (ytrC) were found to be uniquely associated to the strain of *B. amyloliquefaciens* that belongs to clade 1. Hence, the presence of certain gene clusters in clade 2 and their absence in clade 1 conclude that plant-associated strains of *B. amyloliquefaciens* have more abundant gene clusters for intermediary metabolism as well as for antibiotic production compared to the nonplant-associated strains. Niazi et al. (2014) have reported that *B. amyloliquefaciens* subsp. *plantarum*, a rhizobacterium that mends plant growth and stress management, also possesses the more abundant gene cluster that is actively involved in the production of certain hydrolytic enzymes as well as secondary metabolites. It is well documented that the rhizosphere environment has a very dynamic microbial community because of the effect of root exudates and the constant interaction and competition among microbes, as they need to contend with each other for various resources such as nutrient supply, which ultimately leads them to produce various metabolites such as antibiotic and extracellular hydrolases (Bais et al., 2006).

Surveillance of Resistance and Virulence Genes Among all Strains of *B. amyloliquefaciens*

It is documented that bacteria have produced antibiotics for millions of years, which results in the evolution and induction of resistance genes. More precisely, the intensive nonmedical use of antibiotics such as in agricultural and in some industrial applications is not certain and has led to significant dissemination of resistance genes in the environment (Pawlowski et al., 2018). The genomes of all the strains of *B. amyloliquefaciens* were mapped to different databases to evaluate the distribution of antibiotic resistance genes and virulence genes. Many different genes were perceived and were scattered among all the strains of *B. amyloliquefaciens*; also, the observed genes belonged to a variety of resistance classes. The gene (AGlu) *satA* codes for the enzyme aminoglycoside acetyltransferase, and it belongs to the class aminoglycosidese, which is present in almost all the strains independent of its host environment. Aminoglycoside is considered to be part of the broad spectrum of antibiotics, and it acts by inhibiting the protein synthesis, though it works best in synergy with other antimicrobials (Krause et al., 2016). Two genes, *lmrB* and *cfrB*, belonging to the class Macrolide-Lincosamide-Streptogramin B (MLS) were present in most of the strains considered in the investigation. The gene product of *lmrB* and *cfrB* confers specific resistance to lincosamides, such as lincomycin and clindamycin, and synthetic antibiotic linezolid, respectively, (Kim et al., 2001; Toh et al., 2007). The advent of new and more stable macrolide and its vague use could be the key reasons for the induction of such resistance imparting genes, and it provides an opportunity for microbial populations to acquire MLS resistance (Roberts et al., 1999). (Rif) *rphD, rphB*, and *rphC* genes code for trifamycin kinase (phosphotransferase), which confers resistance against rifampin, the most commonly used rifamycin. The enzyme rifampin phosphotransferase present in many environmental bacteria, which used to be induced by selective pressure and nonclinical use of antibiotics, has led to the inactivation of rifampin and ATP to phosphor-rifampin and AMP+Pi (Stogios et al., 2016). More than 40 different tetracycline resistance genes have been reported in numerous bacterial genera of agricultural and industrial use. The dispersion of the *tetL* gene among the bacterial genera was much higher than any other *tet* resistant genes (Roberts, 2005). In the present investigation, the genome sequences of all the strains were mapped against six different databases, i.e., argannot, NCBI, plasmidfinder, card, resfinder, and yfdb, and they reveal the presence of *tetL* genes among all the strains. Colibactin is a genotoxic molecule coded by the *clb* gene cluster in many enteric bacteria, and it is widely distributed in nature (Kawanishi et al., 2020). *clbA* is a plasmid-encoded cfr gene under the control of an inducible promoter reported in *B. velezensis* (*B. amyloliquefaciens* subsp. *plantarum*), while *clbB* and *clbC* are found in *Brevibacillus brevis* and *B. clausii,* respectively, (Hansen et al., 2012). The gene *sspC* codes for cytoplasmic protein known as staphostatin and is present in all the strains of *B. amyloliquefaciens*. It is a very specific and tightly binding inhibitor of staphopain B (SspB). The main function of *sspC* is to protect the cytosolic protein from the degradation executed by misfolded or activated SspB. Shaw et al. (2005) reported that in the absence of sspC protein, major alteration in cellular physiology occurred, and the growth and viability of the microbial cells were impaired. The gene *clpP* is prevailing only in the strains that belong to clade 1, and it codes for the caseinolytic protease proteolytic subunit (ClpP) serine proteases. The ClpP protein confers certain advantages to the microorganisms to sustain in varying environmental conditions as well as stress conditions. ClpC and ClpP are heat shock proteins and are subunits of ATP-dependent proteases reported in *B. subtilis*. The transcription of genes *clpC and* clpP is always negatively regulated under nonstressed condition (Krüger et al., 2001). The virulence and infectivity of a number of microorganisms/pathogens are affected due to the alteration of the ClpP protein function. Clp proteins are highly conserved and have played a very important role in the proteolysis of prokaryotic cell and eukaryotic organelles, though only few reports are available describing the importance of Clp-mediated proteolysis in organisms (Krüger et al., 2001; Moreno-Cinos et al., 2019). Tunicamycin, a nucleoside antibiotic, kills most of the gram-positive bacteria, and it acts by inhibiting the important cell wall component called teichoic acid, which drives the physiology and pathogenesis of microorganisms. The exposure of bacteria toward the sub-inhibitory concentration of tunicamycin leads to the reduction in biofilm production, virulence protein, as well bacterial adhesion and invasion (Zhu et al., 2018). The presence of the *tmrB* gene leads to the production of the TmrB protein, which imparts tunicamycin resistance to *B. subtilis*. The TmrB protein is present in both cytoplasmic and membrane fractions, though it is completely hydrophilic, and it attaches to

the membrane by its C-terminal amphiphilic alpha-helix (Noda et al., 1995). Many plant growth promoting bacteria reported to produce biofilm, which is their key strategy to survive successfully in some harsh conditions as well as in plant rhizosphere. Biofilm formation capability of microorganisms makes them a good biocontrol agent as it leads to the reduction in infection caused by fungal and bacterial pathogens (Hobley et al., 2013). The *bslA/yuaB* gene present in many of the plant-associated strains of *B. amyloliquefaciens* codes for unique surface active protein BslA, which forms a hydrophobic surface layer called hydrophobins. The surface layer regulates the diffusion of various molecules, perception of signaling molecules from other microbial community, as well as nutrient uptake, in addition to imparting the protection to the bacterial cell. The contextual information of ecological and evolutionary facts as well as the application of comparative genomics and the dropping cost of genome sequencing collectively aid to understanding more precisely the structure of microbial diversity and its ecological distribution. Phylogenomics reveals the segregation of all 96 strains of *B. amyloliquefaciens* into two clades. Majority of the plant-associated *B. amyloliquefaciens* strains are grouped in clade 2, while clade 1 accomplishes mostly food-associated strains. The distribution of resistance and virulence genes among all the strains of *B. amyloliquefaciens* has been reported, and it will serve as a benchmark and resourceful information to deduce the hypothesis or conclusion as well as to exploit the potential of any strains through wet-lab experimentation. In future prospectus, we will try to dig out some temporal genes and their occurrence pattern in order to comprehend the significant role of microorganisms as well as the structure of the entire microbial community with its respective environmental niches.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

VP conceived and modeled the study. VP and HL analyzed the data and prepared the methods and results. VP and SP prepared the manuscript. HB and JL corrected the manuscript and inputs. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.724217/full#supplementary-material

Supplementary Figure 1 | A flow chart representation of analyses conducted by different modules of PGCGAP v1.0.21, which integrates some popular software and in-house scripts (Liu et al., 2020).

Supplementary Table 1 | Details of 96 *B. amyloliquefaciens* strains, including their genome size, GC%, scaffolds, CDS, its host and geographical location of the isolates, etc., retrieved from the NCBI database for comparative genome analysis.

Supplementary Table 2 | Data analysis of pan-genome segregation to identify single-copy core proteins, shell genes, soft-core genes, core genes, and cloud genes, including its annotation, genome fragment and order within fragments, accessory order, and accessory order with fragment.

REFERENCES

Alcaraz, L. D., Moreno-Hagelsieb, G., Eguiarte, L. E., Souza, V., Herrera-Estrella, L., and Olmedo, G. (2010). Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics* 11:332. doi: 10.1186/1471-2164-11-332

Baghaee Ravari, S., and Heidarzadeh, N. (2014). Isolation and characterization of rhizosphere auxin producing *Bacilli* and evaluation of their potency on wheat growth improvement. *Arch. Agron. Soil Sci.* 60, 895–905. doi: 10.1080/03650340.2013.856003

Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S., and Vivanco, J. M. (2006). The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu. Rev. Plant Biol.* 57, 233–266. doi: 10.1146/annurev.arplant.57.032905.105159

Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on. *Nucleic Acids Res.* 44, D694–D697. doi: 10.1093/nar/gkv1239

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., et al. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 68, 461–466. doi: 10.1099/ijsem.0.002516

De Wit, P. J. G. M., Van Der Burgt, A., Ökmen, B., Stergiopoulos, I., Abd-Elsalam, K. A., Aerts, A. L., et al. (2012). The genomes of the fungal plant pathogens Cladosporium fulvum and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PLoS Genet* 8:e1003088.

Earl, A. M., Losick, R., and Kolter, R. (2008). Ecology and genomics of *Bacillus subtilis*. *Trends Microbiol.* 16, 269–275.

Eisen, J. A., and Fraser, C. M. (2003). Phylogenomics: intersection of evolution and genomics. *Science* 300, 1706–1707. doi: 10.1126/science.1086292

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., et al. (2019). Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* 63, e483–e419. doi: 10.1128/AAC.00483-19

Gaillot, O., Bregenholt, S., Jaubert, F., Di Santo, J. P., and Berche, P. (2001). Stress-induced ClpP serine protease of *Listeria monocytogenes* is essential for induction of listeriolysin O-dependent protective immunity. *Infect. Immun.* 69, 4938–4943. doi: 10.1128/iai.69.8.4938-4943.2001

Gaillot, O., Pellegrini, E., Bregenholt, S., Nair, S., and Berche, P. (2000). The ClpP serine protease is essential for the intracellular parasitism and virulence of Listeria monocytogenes. *Mol. Microbiol.* 35, 1286–1294. doi: 10.1046/j.1365-2958.2000.01773.x

Gioia, J., Yerrapragada, S., Qin, X., Jiang, H., Igboeli, O. C., Muzny, D., et al. (2007). Paradoxical DNA repair and peroxide resistance gene conservation in *Bacillus pumilus* SAFR-032. *PLoS One* 2:e928. doi: 10.1371/journal.pone.0000928

Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., et al. (2014). ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob.Agents Chemother.* 58, 212–220. doi: 10.1128/aac.01310-13

Hansen, L. H., Planellas, M. H., Long, K. S., and Vester, B. (2012). The order *Bacillales* hosts functional homologs of the worrisome cfr antibiotic resistance gene. *Antimicrob. Agents Chemother.* 56, 3563–3567. doi: 10.1128/aac.00673-12

Hobley, L., Ostrowski, A., Rao, F. V., Bromley, K. M., Porter, M., Prescott, A. R., et al. (2013). BslA is a self-assembling bacterial hydrophobin that coats the

Bacillus subtilis biofilm. *Proc. Natl. Acad. Sci.U.S.A.* 110:13600. doi: 10.1073/pnas.1306390110

Ingle, D. J., Valcanis, M., Kuzevski, A., Tauschek, M., Inouye, M., Stinear, T., et al. (2016). In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microbial. Genomics* 2:e000064. doi: 10.1099/mgen.0.000064

Inglin, R. C., Meile, L., and Stevens, M. J. A. (2018). Clustering of pan-and core-genome of *Lactobacillus* provides novel evolutionary insights for differentiation. *BMC Genomics* 19:284. doi: 10.1186/s12864-018-4601-5

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004

Kawanishi, M., Shimohara, C., Oda, Y., Hisatomi, Y., Tsunematsu, Y., Sato, M., et al. (2020). Genotyping of a gene cluster for production of colibactin and in vitro genotoxicity analysis of *Escherichia coli* strains obtained from the Japan Collection of Microorganisms. *Genes Environ.* 42:12. doi: 10.1186/s41021-020-00149-z

Kim, H. J., Kim, Y., Lee, M. S., and Lee, H. S. (2001). Gene lmrB of *Corynebacterium glutamicum* confers efflux-mediated resistance to lincomycin. *Mol. Cells* 12, 112–116.

Kim, Y., Koh, I., Young, L. M., Chung, W. H., and Rho, M. (2017). Pan-genome analysis of *Bacillus* for microbiome profiling. *Sci. Rep.* 7:10984. doi: 10.1038/s41598-017-11385-9

Kobayashi, K., and Iwano, M. (2012). BslA(YuaB) forms a hydrophobic layer on the surface of *Bacillus subtilis* biofilms. *Mol. Microbiol.* 85, 51–66. doi: 10.1111/j.1365-2958.2012.08094.x

Konstantinidis, K. T., and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2567–2572. doi: 10.1073/pnas.0409727102

Krause, K. M., Serio, A. W., Kane, T. R., and Connolly, L. E. (2016). Aminoglycosides: an overview. *Cold Spring Harb. Perspect. Med.* 6:a027029. doi: 10.1101/cshperspect.a027029

Krüger, E., Zühlke, D., Witt, E., Ludwig, H., and Hecker, M. (2001). Clp-mediated proteolysis in Gram-positive bacteria is autoregulated by the stability of a repressor. *EMBO J.* 20, 852–863. doi: 10.1093/emboj/20.4.852

Lefebure, T., Pavinski Bitar, P. D., Suzuki, H., and Stanhope, M. J. (2010). Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept. *Genome Biol. Evol.* 2, 646–655.

Liu, H., Xin, B., Zheng, J., Zhong, H., Yu, Y., Peng, D., et al. (2020). Build a bioinformatics analysis platform and apply it to routine analysis of microbial genomics and comparative genomics. *Protocol. Exchange*. doi: 10.21203/rs.2.21224/v3

Liu, Y., Chen, L., Zhang, N., Li, Z., Zhang, G., Xu, Y., et al. (2016). Plant-microbe communication enhances auxin biosynthesis by a root-associated bacterium, *Bacillus amyloliquefaciens* SQR9. *Mol. Plant Microbe Interact.* 29, 324–330. doi: 10.1094/mpmi-10-15-0239-r

Mendes-Soares, H., Suzuki, H., Hickey, R. J., and Forney, L. J. (2014). Comparative functional genomics of Lactobacillus spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *J. Bacteriol.* 196, 1458–1470. doi: 10.1128/jb.01439-13

Moreno-Cinos, C., Goossens, K., Salado, I. G., Van Der Veken, P., De Winter, H., and Augustyns, K. (2019). ClpP Protease, a promising antimicrobial target. *Int. J. Mol. Sci.* 20:2232. doi: 10.3390/ijms20092232

Na, S. I., Kim, Y. O., Yoon, S. H., Ha, S. M., Baek, I., and Chun, J. (2018). UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* 56, 280–285. doi: 10.1007/s12275-018-8014-6

Ng, L. C., Sariah, M., Sariam, O., Radziah, O., and Zainal Abidin, M. A. (2016). PGPM-induced defense-related enzymes in aerobic rice against rice leaf blast caused by *Pyricularia oryzae*. *Eur. J. Plant Pathol.* 145, 167–175. doi: 10.1007/s10658-015-0826-1

Niazi, A., Manzoor, S., Asari, S., Bejai, S., Meijer, J., and Bongcam-Rudloff, E. (2014). Genome analysis of *Bacillus amyloliquefaciens* Subsp. plantarum UCMB5113: a rhizobacterium that improves plant growth and stress management. *PLoS One* 9:e104651. doi: 10.1371/journal.pone.0104651

Noda, Y., Takatsuki, A., Yoda, K., and Yamasaki, M. (1995). TmrB protein, which confers resistance to tunicamycin on *Bacillus subtilis*, binds tunicamycin. *Biosci. Biotechnol. Biochem.* 59, 321–322. doi: 10.1271/bbb.59.321

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x

Owusu-Darko, R., Allam, M., Ismail, A., Ferreira, C. A. S., Oliveira, S. D. D., and Buys, E. M. (2020). Comparative genome analysis of *Bacillus sporothermodurans* with Its closest phylogenetic neighbor, *Bacillus oleronius*, and *Bacillus cereus* and *Bacillus subtilis* Groups. *Microorganisms* 8:1185.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Pawlowski, A. C., Westman, E. L., Koteva, K., Waglechner, N., and Wright, G. D. (2018). The complex resistomes of *Paenibacillaceae* reflect diverse antibiotic chemical ecologies. *ISME J.* 12, 885–897. doi: 10.1038/s41396-017-0017-5

Perez, K. J., Viana, J. D. S., Lopes, F. C., Pereira, J. Q., dos Santos, D. M., Oliveira, J. S., et al. (2017). *Bacillus* spp. isolated from puba as a source of biosurfactants and antimicrobial lipopeptides. *Front. Microbiol.* 8:61.

Prajapati, V. S., Trivedi, U. B., and Patel, K. C. (2015). A statistical approach for the production of thermostable and alklophilic alpha-amylase from *Bacillus amyloliquefaciens* KCP2 under solid-state fermentation. *3 Biotech* 5, 211–220. doi: 10.1007/s13205-014-0213-1

Roberts, M. C. (2005). Update on acquired tetracycline resistance genes. *FEMS Microbiol. Lett.* 245, 195–203. doi: 10.1016/j.femsle.2005.02.034

Roberts, M. C., Sutcliffe, J., Courvalin, P., Jensen, L. B., Rood, J., and Seppala, H. (1999). Nomenclature for macrolide and macrolide-lincosamide-streptogramin B resistance determinants. *Antimicrob. Agents Chemother.* 43, 2823–2830. doi: 10.1128/AAC.43.12.2823

Rosselló-Mora, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol. Rev.* 25, 39–67. doi: 10.1111/j.1574-6976.2001.tb00571.x

Shao, J., Li, S., Zhang, N., Cui, X., Zhou, X., Zhang, G., et al. (2015). Analysis and cloning of the synthetic pathway of the phytohormone indole-3-acetic acid in the plant-beneficial *Bacillus amyloliquefaciens* SQR9. *Microbial. Cell Factories* 14:130. doi: 10.1186/s12934-015-0323-4

Sharma, A., and Satyanarayana, T. (2013). Comparative genomics of Bacillus species and its relevance in industrial microbiology. *Genomics Insights* 6, 25–36.

Shaw, L. N., Golonka, E., Szmyd, G., Foster, S. J., Travis, J., and Potempa, J. (2005). Cytoplasmic control of premature activation of a secreted protease zymogen: deletion of staphostatin B (SspC) in Staphylococcus aureus 8325-4 yields a profound pleiotropic phenotype. *J. Bacteriol.* 187, 1751–1762. doi: 10.1128/JB.187.5.1751-1762.2005

Skaar, E. P., Gaspar, A. H., and Schneewind, O. (2004). IsdG and IsdI, heme-degrading enzymes in the cytoplasm of *Staphylococcus aureus*. *J. Biol. Chem.* 279, 436–443. doi: 10.1074/jbc.M307952200

Skaar, E. P., and Schneewind, O. (2004). Iron-regulated surface determinants (Isd) of *Staphylococcus aureus*: stealing iron from heme. *Microbes Infect.* 6, 390–397. doi: 10.1016/j.micinf.2003.12.008

Srivastava, S., Bist, V., Srivastava, S., Singh, P. C., Trivedi, P. K., Asif, M. H., et al. (2016). Unraveling aspects of *Bacillus amyloliquefaciens* mediated enhanced production of rice under biotic stress of *Rhizoctonia solani*. *Front. Plant Sci.* 7:587.

Stingele, F., Newell, J. W., and Neeser, J. R. (1999). Unraveling the function of glycosyltransferases in *Streptococcus thermophilus* Sfi6. *J. Bacteriol.* 181, 6354–6360. doi: 10.1128/JB.181.20.6354-6360.1999

Stogios, P. J., Cox, G., Spanogiannopoulos, P., Pillon, M. C., Waglechner, N., Skarina, T., et al. (2016). Rifampin phosphotransferase is an unusual antibiotic resistance kinase. *Nat. Commun.* 7:11343. doi: 10.1038/ncomms11343

Sun, P., Hui, C., Wang, S., Wan, L., Zhang, X., and Zhao, Y. (2016). *Bacillus amyloliquefaciens* biofilm as a novel biosorbent for the removal of crystal violet from solution. *Colloids Surf. B Biointerfaces* 139, 164–170. doi: 10.1016/j.colsurfb.2015.12.014

Tan, S., Gu, Y., Yang, C., Dong, Y., Mei, X., Shen, Q., et al. (2016). Bacillus amyloliquefaciens T-5 may prevent *Ralstonia solanacearum* infection through competitive exclusion. *Biol. Fertility Soils* 52, 341–351. doi: 10.1007/s00374-015-1079-z

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus*

agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Toh, S.-M., Xiong, L., Arias, C. A., Villegas, M. V., Lolans, K., Quinn, J., et al. (2007). Acquisition of a natural resistance gene renders a clinical strain of methicillin-resistant *Staphylococcus aureus* resistant to the synthetic antibiotic linezolid. *Mol. Microbiol.* 64, 1506–1514. doi: 10.1111/j.1365-2958.2007.05744.x

Verma, A., Singh, V. K., and Gaur, S. (2016). Computational based functional analysis of *Bacillus phytases*. *Comput. Biol. Chem.* 60, 53–58. doi: 10.1016/j.compbiolchem.2015.11.001

Vernikos, G., Medini, D., Riley, D. R., and Tettelin, H. (2015). Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154.

Wang, H., Yang, L., Ping, Y., Bai, Y., Luo, H., Huang, H., et al. (2016). Engineering of a *Bacillus amyloliquefaciens* strain with high neutral protease producing capacity and optimization of its fermentation conditions. *PLoS One* 11:e0146373. doi: 10.1371/journal.pone.0146373

Wang, X., Bai, Y., Cai, Y., and Zheng, X. (2017). Biochemical characteristics of three feruloyl esterases with a broad substrate spectrum from *Bacillus amyloliquefaciens* H47. *Process Biochem.* 53, 109–115. doi: 10.1016/j.procbio.2016.12.012

Wu, B., Changjun, W., Dihong, X., Heng, Z., Huan, Y., Jinping, L., et al. (2016). Effects of *Bacillus amyloliquefaciens* ZM9 on bacterial wilt and rhizosphere microbial communities of tobacco. *Appl. Soil Ecol.*103, 1–12. doi: 10.1016/j.apsoil.2016.03.002

Wu, L., Wu, H.-J., Qiao, J., Gao, X., and Borriss, R. (2015). Novel routes for improving biocontrol activity of *Bacillus* based bioinoculants. *Front. Microbiol.* 6:1395.

Yang, L., Wang, H., Lv, Y., Bai, Y., Luo, H., Shi, P., et al. (2016). Construction of a rapid feather-degrading bacterium by overexpression of a highly efficient alkaline keratinase in its parent strain *Bacillus amyloliquefaciens* K11. *J. Agric. Food Chem.* 64, 78–84. doi: 10.1021/acs.jafc.5b04747

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., et al. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* 67, 2640–2644. doi: 10.1093/jac/dks261

Zhang, J., Xue, Q., Gao, H., Lai, H., and Wang, P. (2016). Bacterial degradation of crude oil using solid formulations of bacillus strains isolated from oil-contaminated soil towards microbial enhanced oil recovery application. *RSC Adv.* 6, 5566–5574. doi: 10.1039/C5RA23772F

Zhang, N., Yang, D., Kendall, J. R. A., Borriss, R., Druzhinina, I. S., Kubicek, C. P., et al. (2016). Comparative genomic analysis of *Bacillus amyloliquefaciens* and *Bacillus subtilis* reveals evolutional traits for adaptation to plant-associated habitats. *Front. Microbiol.* 7:2039. doi: 10.3389/fmicb.2016.02039

Zhu, X., Liu, D., Singh, A. K., Drolia, R., Bai, X., Tenguria, S., et al. (2018). tunicamycin mediated inhibition of wall teichoic acid affects *Staphylococcus aureus* and listeria monocytogenes cell morphology, biofilm formation and virulence. *Front. Microbiol.* 9:1352.

Zühlke, M.-K., Schlüter, R., Henning, A.-K., Lipka, M., Mikolasch, A., Schumann, P., et al. (2016). A novel mechanism of conjugate formation of bisphenol A and its analogues by *Bacillus amyloliquefaciens*: Detoxification and reduction of estrogenicity of bisphenols. *Int. Biodeterior. Biodegradation* 109, 165–173. doi: 10.1016/j.ibiod.2016.01.019

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fgene.2021.764927

Evolutionary and Antigenic Profiling of the Tendentious D614G Mutation of SARS-CoV-2 in Gujarat, India

*Jay Nimavat[1†], Chandrashekar Mootapally[2†], Neelam M. Nathani[2†], Devyani Dave[1], Mukesh N. Kher[3], Mayur S. Mahajan[4], Chaitanya G. Joshi[5] and Vaibhav D. Bhatt[2,4]**

[1]Department of Pharmaceutical Sciences, Saurashtra University, Rajkot, India, [2]School of Applied Sciences and Technology (GTU-SAST), Gujarat Technological University, Ahmedabad, India, [3]L. M. College of Pharmacy, Ahmedabad, India, [4]Atal Incubation Centre, Gujarat Technological University, Ahmedabad, India, [5]Gujarat Biotechnology Research Centre, Department of Science and Technology, Gandhinagar, India

Humankind has suffered many pandemics in history including measles, SARS, MERS, Ebola, and recently the novel Coronavirus disease caused by SARS-CoV-2. As of September 2021, it has affected over 200 million people and caused over 4 million deaths. India is the second most affected country in the world. Up to this date, more than 38 Lakh viral genomes have been submitted to public repositories like GISAID and NCBI to analyze the virus phylogeny and mutations. Here, we analyzed 2349 genome sequences of SARS-CoV-2 submitted in GISAID by a single institute pertaining to infections from the Gujarat state to know their variants and phylogenetic distributions with a major focus on the spike protein. More than 93% of the genomes had one or more mutations in the spike glycoprotein. The D614G variant in spike protein is reported to have a very high frequency of >95% globally followed by the L452R and P681R, thus getting significant attention. The antigenic propensity of a small peptide of 29 residues from 597 to 625 of the spike protein variants having D614 and G614 showed that G614 has a little higher antigenic propensity. Thus, the D614G is the cause for higher viral antigenicity, however, it has not been reported to be effective to be causing more deaths.

Keywords: antigenic propensity, clades, D614G, SARS-CoV-2, spike glycoprotein

OPEN ACCESS

Edited by:
Dhaval K. Acharya, B N Patel Institute of Paramedical, India

Reviewed by:
Arif Ansori, Airlangga University, Indonesia
Alejandro Flores-Alanis, National Autonomous University of Mexico, Mexico

****Correspondence:***
Vaibhav D. Bhatt
bhatt_vbhv@yahoo.co.in

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *26 August 2021*
Accepted: *18 October 2021*
Published: *11 November 2021*

Citation:
Nimavat J, Mootapally C, Nathani NM, Dave D, Kher MN, Mahajan MS, Joshi CG and Bhatt VD (2021) Evolutionary and Antigenic Profiling of the Tendentious D614G Mutation of SARS-CoV-2 in Gujarat, India. Front. Genet. 12:764927. doi: 10.3389/fgene.2021.764927

INTRODUCTION

In the last 2 decades, this is the third instance of a zoonotic coronavirus pandemic. Acute respiratory disease has previously been caused by SARS-CoV, 2002 (Drosten et al., 2003) and MERS-CoV, 2012 (Azhar et al., 2014a; Azhar et al., 2014b) in humans. The novel SARS-CoV-2 virus has recently triggered the coronavirus disease Covid-19. The viral infection began in Wuhan, China in December 2019 and soon became a global outbreak. In a very short span, it has caused significant effects on social and economic activities. Compared to other coronaviruses, this newly emerged SARS-CoV-2 is spreading rapidly, giving challenges to administrative and scientific communities. Influenza, severe respiratory, enteric and neurological complications, elevated white blood cells, and kidney failure are significant indicators of this viral infection. Mammals such as bats are the primary beta coronavirus reservoirs. Due to zoonotic contacts and viral genomic mutations, it is expected to have crossed the species barrier and infected humans. Previous studies indicate that zoonotic infections such as SARS-CoV was transmitted from bats and civets that first infected humans in 2002 (Ksiazek et al., 2003; Marra et al., 2003; Rota et al., 2003; Xu et al., 2004). SARS-CoV-2 is enveloped, contains positive

sense ssRNA, and a genome size of 29–30 kb. It belongs to the coronaviridae family and subfamily beta-coronavirus. The family also comprises of MERS CoV which was originated from camel and later led to human transmission in Saudi Arabia (2012) (Azhar et al., 2014a; Chan et al., 2015; Sabir et al., 2016). These infections from bats are predicted to infect humans due to their change in genomic RNA sequence, especially in the spike glycoprotein region (Song et al., 2005; Menachery et al., 2016). Complete genomic sequences of SARS-CoV-2 isolated from infected patients belonging to different geographical locations allows the understanding of these variations and the corresponding influence on viral infecting potency.

SARS CoV-2 similar to SARS-CoV uses angiotensin-converting enzyme II (ACE2) as a receptor for host cell entry. It has spike glycoproteins on the surface, which has two functional domains, S1 and S2. It helps in host cell receptor binding and fusion of viral membrane with the cellular membrane (Harvey et al., 2021). SARS-CoV-2 spike protein has an ACE2 affinity 10 to 20 times greater than that of SARS-CoV spike protein (Walls et al., 2020; Wrapp et al., 2020). Both SARS-CoV and SARS-CoV-2 use CTD (C-terminal domain) of the S1 domain for receptor binding but SARS-CoV-2 binds more strongly than SARS-CoV. Coronaviruses use two different pathways for host cell entry, First, protease mediated cell surface pathway and second, the endosomal pathway. S protein is cleaved into an S1 subunit for receptor binding and an S2 subunit for membrane fusion by the host proteases. Several cellular proteases including furin, transmembrane protease serine 2 (TMPRSS2) and cathepsin (cat) B/L are important for priming SARS-CoV-2 spike protein to enhance ACE2 mediated viral entry (Hoffmann et al., 2020). Spike protein plays a vital role in the evolution of coronaviruses to escape the host immune system. Spike protein shows a higher amount of antigenicity, which is evident from the fact that convalescent plasma from SARS patients shows a high percentage of anti-S neutralizing antibodies (Liu et al., 2006; Wan et al., 2020). There are a lot of variations observed in spike protein sequence, a major variation in spike protein is a non-synonymous D614G mutation (D-Aspartate, G-Glycine) which has received special attention by several groups due to its dominance (Harvey et al., 2021).

Several studies have reported the phylogenomics of the variant and shown that it is leading to higher transmission, though no less influence on the death rate. Few studies have also shown that the variant has increased cellular entry efficacy to human cells compared to the wild type (Wan et al., 2020). In context to the same, here we assessed the antigenic propensity of the epitope encompassing the D614G mutation considering its high frequency and the segment being earlier reported as immune-dominant peptide in SARS-CoV (Wang et al., 2016; Kim et al., 2020). The region-wise analysis of the variant will provide information of this rapidly spreading variant for possible considerations in protective strategy development. Further, we also compiled the current major spike mutations in the Gujarat state in comparison with their global frequencies.

MATERIALS AND METHOD

SARS-CoV-2 Sequence Retrieval

A total of 2439 sequences of SARS-CoV-2 genome sequences corresponding to the Gujarat state were retrieved from GISAID (https://www.gisaid.org/) hCov-19 database and these sequences represented different districts of Gujarat state. The genomes were sequenced and submitted by Gujarat Biotechnology Research. The complete genome sequences of SARS-CoV-2 reference genome of the Wuhan isolate (GenBank code: MN908947.3) and Bat CoV-RaTG13 (MN996532.1) were retrieved in FASTA format from NCBI. All the retrieved sequences were subjected to BLAST.

Mutation and Phylogenetic Analysis

Spike protein sequences retrieved from GISAID were aligned with reference spike protein sequences using Jalview version 2.11.1.0 (Waterhouse et al., 2009). Mutations were identified and listed using GISAID EpiCoV™ database (Elbe and Buckland-Merrett, 2017). Occurrence of D614G mutation over time was visualized using NextStrain platform (Hadfield et al., 2018) where data is enabled from the GISAID. NextStrain visualization analysis can process up to 3000 genomic sequences at a time. Therefore, for the primary global analysis, they subsample 120 genomes per admin division per month giving results in a more equitable way.

Antigenic Propensity Analysis

A small part of sequence of spike glycoprotein $S_{597\text{-}625}$ from sequences was analyzed using the method described earlier (Kolaskar and Tongaonkar, 1990) provided on an online server by UNIVERSIDAD COMPLUTENSE, MADRID. The interpretation was done as suggested: the average score for the whole protein was used as a cut-off for the then all residues to be considered as potentially antigenic.

RESULTS

Sequence Similarity Studies of SARS-CoV-2 Genomes

A total of 2439 sequences of SARS-CoV-2 were retrieved from GISAID platform. Upon performing nucleotide BLAST of reference SARS-CoV-2 with SARS CoV, at the genomic level SARS-CoV-2 and SARS-CoV were observed to have 79.6% sequence similarity. Sequences retrieved from infected patients by SARS-Cov-2 from GISAID had sequence similarity of about 80% with SARS-CoV, 99.9% with SARS-CoV-2 reference sequence, and 96% with the Bat CoV RaTG13.

Mutation Analysis of SARS-CoV-2 Genomes

Out of 2439 sequences studied, 2400 genomes had a common mutation and there were 9 mutations observed in spike protein that were present in at least 15% of the studied genomes (**Figure 1**). A total of 34 nonsynonymous mutations were observed, with the spike D614G having the highest occurrence in 2400 genomes followed by the nsp12 P4715L mutation observed in 2254 genomes (**Figure 1A**). Nine of the spike mutations had a percentage occurrence in the range of 25–99

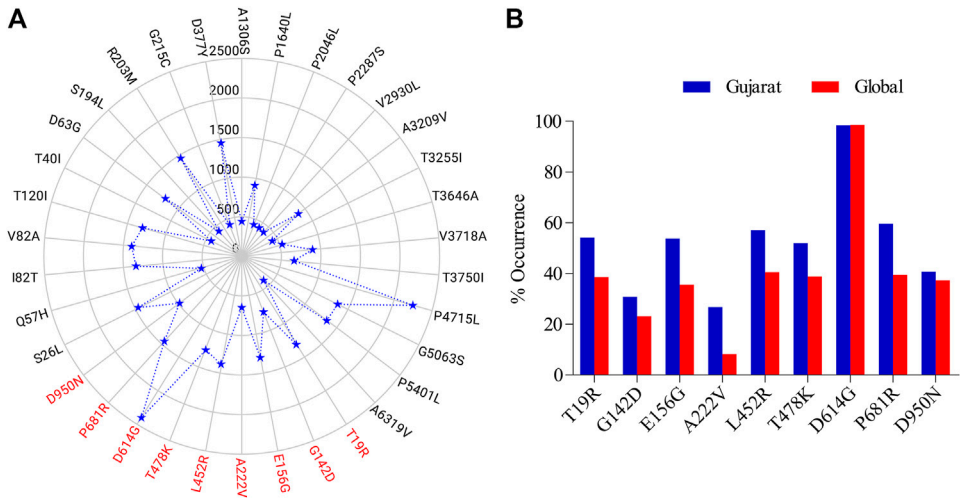

FIGURE 1 | Non synonymous mutations as observed **(A)** in the analysed SARS-CoV-2 genomes (*n* = 2,439) from Gujarat region (submissions by GBRC) compared to the Wuhan SARS-CoV-2 reference, scale represents the number of genomes, those highlighted in red are specific to the spike glycoprotein, **(B)** in the spike glycoprotein of the sequences from Gujarat (*n* = 2,439) and globally (*n* = 3,897,179) in terms of their percent occurrence as on September, 2021.

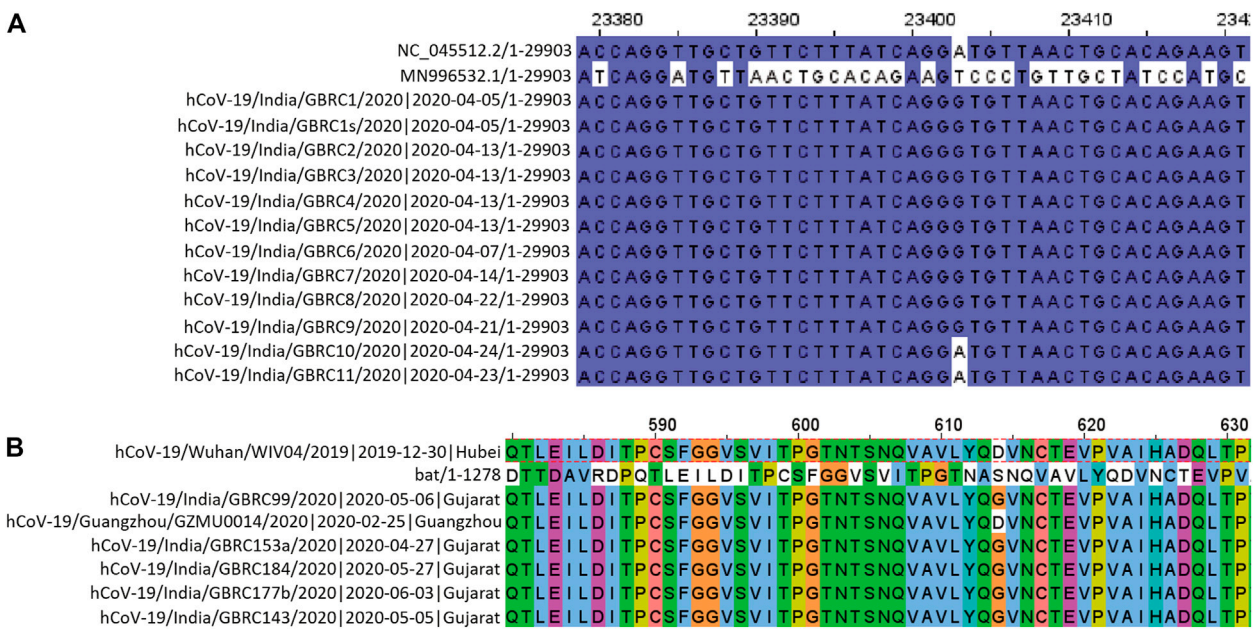

FIGURE 2 | (A) Change in the nucleotide at position 23403. NC_045512.2/1-29903 and MN996532.1/1-29855 are SARS-CoV-2 reference sequence and Bat CoV RaTG13 reference sequence, respectively. **(B)** Amino acid sequence alignment of spike protein encompassing the D614G position.

in the studied genomes (**Figure 1B**). After D614G, P681R occurred in 1455 of the analyzed sequences.

All, except D614G, from the 9 non-synonymous mutations observed, had relatively more frequency in the Gujarat region compared to their global frequencies. While D614G had almost similar frequency with its global value. In the genomic sequence for spike glycoprotein, a single mutation, i.e., from A to G nucleotide was prevalent in the majority of genomes at the nucleotide position number 23403 (**Figure 2**). This change in the nucleotide causes change in amino acid while translating the gene, because of this aspartate is replaced by glycine in the protein sequence (**Figure 2**). Its frequency of occurrence has overall increased with the time. Phylogenomics of genomes based on the D and G variants is depicted at the time of initial data collection and the scenario down to the recent timeline (**Supplementary Figure S1**). Also, there is a clear difference in frequency as observed for D614G

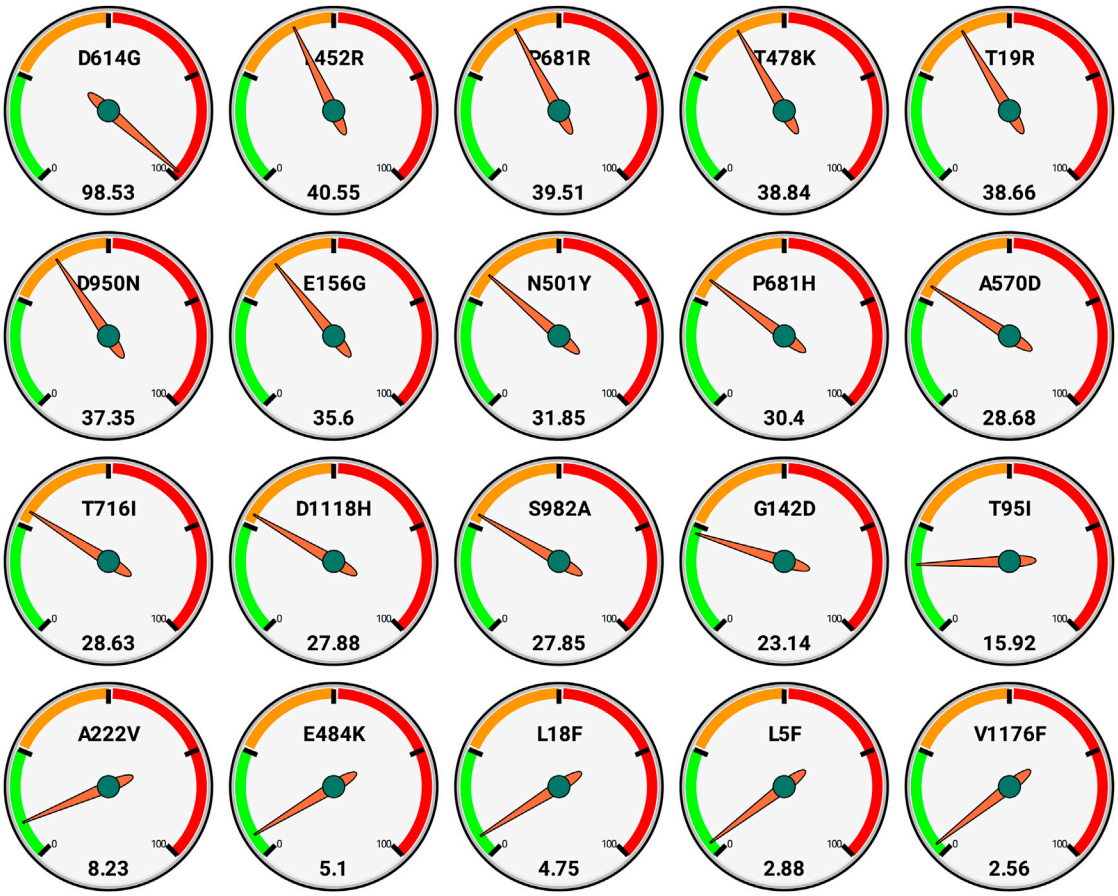

FIGURE 3 | Global frequencies of top 20 non synonymous mutations in the spike protein of the SARS-CoV-2 genomes (n = 3,897,179) as available in GISAID as on date September 30, 2021. Each gauge has a frequency scale (0–100) divided quarterly and represents single mutation (top) along with its respective frequency (bottom).

which was comparatively much higher compared to global during early phase up to July 2020 (**Supplementary Figure S2**) and now the frequencies are almost the same. Such difference currently is observed in the P681R mutation, which is the hallmark mutation of the Delta variant wherein the percentage occurrence is 20% higher in the Gujarat region compared to the global. The P681R has outraced other major mutations in spike proteins and is the second major mutation.

We also report here the top 20 nonsynonymous mutations as per global scenario with their frequencies (**Figure 3**). These also reflect that D613G is the most spread, and further there are around 12 mutations that have >25% occurrence globally.

Antigenicity of Peptides

In the present study, we assessed the antigenic propensity of the peptides of the region $S_{597\text{-}625}$ and the results showed that the variant having G at position number 614 is having a little higher antigenic propensity than the one having D at the same position. Isolate from Wuhan (i.e., reference having D614), a Pangolin CoV and a variant having G614 had antigenic propensity of 1.0822, 1.0822, and 1.0824, respectively. In each sequence of 29 residues, a peptide starting from position 10 to the 25th position comprised a single antigenic determinant (**Figure 4**).

DISCUSSION

A novel corona virus that emerged in December 2019 from Wuhan, China has resulted into a pandemic. In the structure of SARS-CoV-2 virus, spike protein is 1273 amino acid residue in length and forms a trimeric spike on the virion surface. There are many mutations observed in the amino acid composition of the spike protein but primary data shows that strains with S-D614G are more infectious and exhibit high transmission efficiency (Zhang et al., 2020). Regions between amino acid 614 and 621 of SARS-CoV-2 spike proteins were also identified as a B cell epitope by different methods and D614G may affect the antigenicity of this region (Kim et al., 2020). However, there is still much scope to understand how D614G affects antigenic properties of S protein; whether elastase-2 inhibitors and convalescent serum samples of patients can block infection of D614G variant remains unclear.

FIGURE 4 | Antigenic propensity of two different SARS-2COV-2 sequences and Pangolin-CoV. **(A)** Pangolin-CoV, **(B)** SARS-CoV-2 (D614G) Variant, **(C)** Reference sequence at 614 position in spike glycoprotein sequence.

In the current study, we observed that D614G is highly prevalent mutation in the spike protein of genomes from COVID-19 patients of the Gujarat region. D614 is conserved in the reference sequence from Wuhan and a sequence from Guangzhou, while SARS-CoV-2 genome sequences from Gujarat, India showed a very high frequency of this mutation. In addition, it is concurrently seen with other mutations like P681R, L452R, T19R, E156G, T478K. These had >50% of occurrence in Gujarat whereas at the global level only D614G showed very high occurrence (98.53%) of all sequences and the rest of these formerly mentioned mutations had a frequency of 35–40% globally, lower than that in Gujarat. The observation that the P681R was the second most prevalent mutation reveals the recent high dominance of the Delta variant in the region. Additionally, spike D614G was accompanied by high occurrence of the nsp12 P4715L mutation, and this duo variant which is linked to pathogenicity was observed to be not linked positively to fatality rates in Africa (Lamptey et al., 2021). Such duo variants need further attention to assess host-based region-specific response.

Considering the high occurrence of D614G in spike protein of SARS-CoV-2, several groups have assessed the antigenic peptides and it is reported that the peptide $S_{597\text{-}625}$ is one of the major immuno-dominant in humans (Wang et al., 2016; Kim et al., 2020). Antigenic propensity analysis showed that variant spike protein- G_{614} is having a little higher antigenic propensity to the D_{614}. This observation may be one of the reasons for no change in the death rate despite the high spread of the variant. Further studies on spike protein epitopes may provide insights on the potential efficacy of many of the vaccines which may be designed based on the D614 sequence.

Earlier reports have also showed that D614G increases the efficiency of cellular entry for the virus across a broad range of

human cell types, including cells from lung, liver, and colon (Daniloski et al., 2021). They also observed that spike-G_{614} is more resistant to proteolytic cleavage during the production of the protein in the host cell.

Seeing the rise in the cases with D614G mutation and its enhanced transmission, the D614G attracts significant consideration by researchers and healthcare field fellows. In the present work, we attempt to report the mutation analysis of spike protein and the antigenic propensity of D614G mutation in the spike protein of the viral isolates from the Gujarat region.

DATA AVAILABILITY STATEMENT

The original contributions of SARS-CoV-2 genome sequences (n = 2439) presented in the study are publicly available. This data can be found here: https://www.gisaid.org/, the corresponding accession numbers are available from https://covid.gbrc.res.in/ (Sr. No. 1-2439).

AUTHOR CONTRIBUTIONS

JN: Writing original draft and data analysis. CM: Formal analysis, review, and editing. NN: Formal analysis, review, and editing. DD: Data collection. MK: Data collection. MM: Manuscript formatting and data analysisv. CJ: Conceptualization, writing—review and editingv. VB: Conceptualization, writing—review and editingv. All authors contributed to the manuscript and approved the submitted version.

ACKNOWLEDGMENTS

The authors are thankful to Gujarat Biotechnology Research Centre for providing the data of Covid-19 samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.764927/full#supplementary-material

Supplementary Figure S1 | Phylogenomic distribution of D and G variants **(A)** upto the studied genomes viz., July 2020 and **(B)** scenario down to the recent timeline viz., September 2021.

Supplementary Figure S2 | Non synonymous mutations observed in the spike glycoprotein sequence of analysed genomes from Gujarat ($n = 315$) compared to the globally submitted genome sequences of SARS-CoV-2 compared to Wuhan reference by EpiCoVTM on GISAID platform as on June 30, 2020.

REFERENCES

Azhar, E. I., El-Kafrawy, S. A., Farraj, S. A., Hassan, A. M., Al-Saeed, M. S., Hashem, A. M., et al. (2014). Evidence for Camel-To-Human Transmission of MERS Coronavirus. *N. Engl. J. Med.* 370 (26), 2499–2505. PubMed PMID: 24896817. doi:10.1056/NEJMoa1401505

Azhar, E. I., Hashem, A. M., El-Kafrawy, S. A., Sohrab, S. S., Aburizaiza, A. S., Farraj, S. A., et al. (2014). Detection of the Middle East Respiratory Syndrome Coronavirus Genome in an Air Sample Originating from a Camel Barn Owned by an Infected Patient. *mBio* 5 (4), e01450–14. doi:10.1128/mBio.01450-14

Chan, J. F. W., Lau, S. K. P., To, K. K. W., Cheng, V. C. C., Woo, P. C. Y., and Yuen, K.-Y. (2015). Middle East Respiratory Syndrome Coronavirus: Another Zoonotic Betacoronavirus Causing SARS-like Disease. *Clin. Microbiol. Rev.* 28 (2), 465–522. doi:10.1128/CMR.00102-14

Daniloski, Z., Jordan, T. X., Ilmain, J. K., Guo, X., Bhabha, G., tenOever, B. R., et al. (2021). The Spike D614G Mutation Increases SARS-CoV-2 Infection of Multiple Human Cell Types. *Elife* 10, e65365, 2021 . PubMed PMID: 33570490. doi:10.7554/eLife.65365

Drosten, C., Günther, S., Preiser, W., van der Werf, S., Brodt, H.-R., Becker, S., et al. (2003). Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* 348 (20), 1967–1976. PubMed PMID: 12690091. doi:10.1056/NEJMoa030747

Elbe, S., and Buckland-Merrett, G. (2017). Data, Disease and Diplomacy: GISAID's Innovative Contribution to Global Health. *Glob. Challenges* 1 (1), 33–46. doi:10.1002/gch2.1018

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* 34 (23), 4121–4123. doi:10.1093/bioinformatics/bty407

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., et al. (2021). SARS-CoV-2 Variants, Spike Mutations and Immune Escape. *Nat. Rev. Microbiol.* 19 (7), 409–424. doi:10.1038/s41579-021-00573-0

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181 (2), 271–280e8. doi:10.1016/j.cell.2020.02.052

Kim, S.-J., Nguyen, V.-G., Park, Y.-H., Park, B.-K., and Chung, H.-C. (2020). A Novel Synonymous Mutation of SARS-CoV-2: Is This Possible to Affect Their Antigenicity and Immunogenicity? *Vaccines* 8 (2), 220, 2020 . PubMed PMID:. doi:10.3390/vaccines8020220

Kolaskar, A. S., and Tongaonkar, P. C. (1990). A Semi-empirical Method for Prediction of Antigenic Determinants on Protein Antigens. *FEBS Lett.* 276 (1-2), 172–174. doi:10.1016/0014-5793(90)80535-Q

Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., et al. (2003). A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* 348 (20), 1953–1966. PubMed PMID: 12690092. doi:10.1056/NEJMoa030781

Lamptey, J., Oyelami, F. O., Owusu, M., Nkrumah, B., Idowu, P. O., Adu-Gyamfi, E. A., et al. (2021). Genomic and Epidemiological Characteristics of SARS-CoV-2 in Africa. *Plos Negl. Trop. Dis.* 15 (4), e0009335. doi:10.1371/journal.pntd.0009335

Liu, W., Fontanet, A., Zhang, P. H., Zhan, L., Xin, Z. T., Baril, L., et al. (2006). Two-Year Prospective Study of the Humoral Immune Response of Patients with Severe Acute Respiratory Syndrome. *J. Infect. Dis.* 193 (6), 792–795. doi:10.1086/500469

Marra, M. A., Jones, S. J. M., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S. N., et al. (2003). The Genome Sequence of the SARS-Associated Coronavirus. *Science* 300 (5624), 1399–1404. doi:10.1126/science.1085953

Menachery, V. D., Yount, B. L., Sims, A. C., Debbink, K., Agnihothram, S. S., Gralinski, L. E., et al. (2016). SARS-like WIV1-CoV Poised for Human Emergence. *Proc. Natl. Acad. Sci. USA* 113 (11), 3048–3053. doi:10.1073/pnas.1517719113

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., et al. (2003). Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *Science* 300 (5624), 1394–1399. doi:10.1126/science.1085952

Sabir, J. S., Lam, T. T., Ahmed, M. M., Li, L., Shen, Y., Abo-Aba, S. E., et al. (2016). Co-circulation of Three Camel Coronavirus Species and Recombination of MERS-CoVs in Saudi Arabia. *Science* 351 (6268), 81–84. doi:10.1126/science.aac8608

Song, H.-D., Tu, C.-C., Zhang, G.-W., Wang, S.-Y., Zheng, K., Lei, L.-C., et al. (2005). Cross-host Evolution of Severe Acute Respiratory Syndrome Coronavirus in palm Civet and Human. *Proc. Natl. Acad. Sci.* 102 (7), 2430–2435. doi:10.1073/pnas.0409608102

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181 (2), 281–292e6. doi:10.1016/j.cell.2020.02.058

Wan, Y., Shang, J., Graham, R., Baric, R. S., and Li, F. (2020). Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J. Virol.* 94 (7), e00127–20. doi:10.1128/jvi.00127-20

Wang, Q., Zhang, L., Kuwahara, K., Li, L., Liu, Z., Li, T., et al. (2016). Immunodominant SARS Coronavirus Epitopes in Humans Elicited Both Enhancing and Neutralizing Effects on Infection in Non-human Primates. *ACS Infect. Dis.* 2 (5), 361–376. doi:10.1021/acsinfecdis.6b00006

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* 25 (9), 1189–1191. doi:10.1093/bioinformatics/btp033

Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., et al. (2020). Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. *Science* 367 (6483), 1260–1263. doi:10.1126/science.abb2507

Xu, R.-H., He, J.-F., Evans, M. R., Peng, G.-W., Field, H. E., Yu, D.-W., et al. (2004). Epidemiologic Clues to SARS Origin in China. *Emerg. Infect. Dis.* 10 (6), 1030–1037. doi:10.3201/eid1006.030852

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Peng, H., Quinlan, B. D., et al. (2020). SARS-CoV-2 Spike-Protein D614G Mutation Increases Virion Spike Density and Infectivity. *Nat. Commun.* 11 (1), 6013–6019. doi:10.1038/s41467-020-19808-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 December 2021
doi: 10.3389/fgene.2021.770350

Molecular Diagnosis of Muscular Dystrophy Patients in Western Indian Population: A Comprehensive Mutation Analysis Using Amplicon Sequencing

Komal M. Patel[1], *Arpan D. Bhatt*[1], *Krati Shah*[2], *Bhargav N. Waghela*[1], *Ramesh J. Pandit*[1], *Harsh Sheth*[3], *Chaitanya G. Joshi*[1] *and Madhvi N. Joshi*[1]*

[1]Gujarat Biotechnology Research Centre, Department of Science and Technology, Government of Gujarat, Gandhinagar, India, [2]ONE-Centre for Rheumatology and Genetics, Vadodara, India, [3]Foundation for Research in Genetics and Endocrinology (FRIGE), Ahmedabad, India

Muscular Dystrophies (MDs) are a group of inherited diseases and heterogeneous in nature. To date, 40 different genes have been reported for the occurrence and/or progression of MDs. This study was conducted to demonstrate the application of next-generation sequencing (NGS) in developing a time-saving and cost-effective diagnostic method to detect single nucleotide variants (SNVs) and copy number variants (CNVs) in a single test. A total of 123 cases clinically suspected of MD were enrolled in this study. Amplicon panel-based diagnosis was carried out for 102 (DMD/BMD) cases and the results were further screened using multiplex ligation-dependent probe amplification (MLPA). Whilst in the case of LGMD (N = 19) and UMD (N = 2), only NGS panel-based analysis was carried out. We identified the large deletions in 74.50% (76/102) of the cases screened with query DMD or BMD. Further, the large deletion in *CAPN3* gene (N = 3) and known SNV mutations (N = 4) were identified in LGMD patients. Together, the total diagnosis rate for this amplicon panel was 70.73% (87/123) which demonstrated the utility of panel-based diagnosis for high throughput, affordable, and time-saving diagnostic strategy. Collectively, present study demonstrates that the panel based NGS sequencing could be superior over to MLPA.

Keywords: next generation sequencing (NGS), multiplex ligation-dependent probe amplification (MLPA), duchenne muscular dystrophy (DMD), becker muscular dystrophy (BMD), limb-girdle muscular dystrophies, congenital muscular dystrophies (CMDs)

OPEN ACCESS

Edited by:
Dhaval K. Acharya, B N Patel Institute of Paramedical, India

Reviewed by:
Corrado Italo Angelini, University of Padua, Italy
Dusanka Savic Pavicevic, University of Belgrade, Serbia

***Correspondence:**
Madhvi N. Joshi
madhvimicrobio@gmail.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *03 September 2021*
Accepted: *10 November 2021*
Published: *03 December 2021*

Citation:
Patel KM, Bhatt AD, Shah K, Waghela BN, Pandit RJ, Sheth H, Joshi CG and Joshi MN (2021) Molecular Diagnosis of Muscular Dystrophy Patients in Western Indian Population: A Comprehensive Mutation Analysis Using Amplicon Sequencing. Front. Genet. 12:770350. doi: 10.3389/fgene.2021.770350

HIGHLIGHTS

- Muscular Dystrophies (MDs) are genetically heterogeneous diseases.
- Loss of function mutations in the *DMD* gene causes non-functional dystrophin protein that progresses various MDs.
- The customized amplicon panel consisting of genes targeting 29 MDs was used to detect large deletions in the *DMD* gene and novel deletion in the *CAPN3* gene.
- NGS-based study provides eligibility of patients for currently available treatment such as exon skipping.

INTRODUCTION

Muscular dystrophy is a genetically heterogeneous group of neuromuscular diseases that result in degradation of skeletal muscles, progressive muscle weakness, loss of ambulation, cardiac attack, and respiratory failure (Wang et al., 2019). To date, more than 30 different types of MDs are known and can be classified based on the onset of disease, clinical manifestations, mode of inheritance, and severity of the disease (Gaina et al., 2019). Duchenne Muscular Dystrophy (OMIM # 310200) is the most common, rapidly progressive, and severe neuromuscular disease. It inherits in an X-linked recessive manner affecting 1 in 3,500 male children with onset before 5 years' age (Wu et al., 2017; Mohammed et al., 2018; Zhang Y. et al., 2019). Becker Muscular Dystrophy (BMD) is a less severe form of the disease caused by the mutation in the *DMD* gene and slow progressive with an incidence rate of 1 in 20,000 male children (Aartsma-Rus et al., 2016; Wu et al., 2017; Mohammed et al., 2018; Kong et al., 2019; Wang et al., 2019). The *DMD* (dystrophin) is a large gene encompassing 79 exons and spanning approximately 2.5 Mb of the genomic DNA. Loss of function mutations in the *DMD* gene causes an impaired dystrophin protein which disturbs the membrane complex and myofiber loss. In contrast, patients with BMD have a shorter or less functioning form of the dystrophin protein, which makes the disease less severe and slow progressive (Wicklund, 2013; Wu et al., 2017; Wang et al., 2019). DMD and BMD are caused due to various mutations like large deletions (60%), SNVs and INDELs (30%), and duplications in the *DMD* gene (5–7%) (Mohammed et al., 2018; Wang et al., 2019).

Limb-Girdle Muscular Dystrophies (LGMDs) are another heterogeneous group of MDs consisting of around 30 subtypes, vary with genetic and clinical characteristics (Iyadurai and Kissel, 2016). LGMDs are progressive and characterized by weakness of the shoulder and pelvic girdle muscles. The incidence rate of LGMDs is approximately 1 in 14,500 to 123,000 (Pegoraro and Hoffman, 2012; Murphy and Straub, 2015; Nallamilli et al., 2018) [*https://rarediseases.org/rare-diseases/limb-girdle-muscular-dystrophies*, last accessed July 29, 2020.] The inheritance pattern of LGMDs is both autosomal dominant (AD-LGMD) and autosomal recessive (AR-LGMD). AR-LGMDs are more frequent than AD-LGMDs. LGMD associated proteins includes dystrophin-glycoprotein complex (DGC) and play a pivotal role in membrane stability. The mutations in MD genes causes disturbance in DGC proteins that destabilizes the membrane and eventually muscle degradation (Murphy and Straub, 2015).

Congenital Muscular Dystrophies (CMDs) are another group of muscular dystrophies that are also heterogeneous and affect newborns with an incidence of 1:10,000 to 1:50,000. Common symptoms of CMDs include hypotonia, scoliosis, motor delay, and muscle weakness from birth or infancy. Moreover, mutation in multiple genes causes CMDs (Valencia et al., 2013).

Routinely, Multiplex Ligation Dependent Probe Amplification (MLPA) or array-CGH (aCGH) diagnostics tests are being used to detect large CNVs (Deletions/Duplications) in MDs (Zhang K. et al., 2019). The results of these diagnostic tests further requires targeted sequencing to detect SNVs in the DMD/BMD cases. In several cases in which large deletions can be a cause of LGMDs, an aCGH is an exclusive option (Zhang K. et al., 2019). Performing aCGH in all referred cases would be time consuming and expensive. In a developing country like India, cost-effective and a single screening approach to detect CNVs, and SNVs, can be a boon. Furthermore, variety of therapies for DMD patients are available and few are under development, which requires an utmost knowledge of breakpoints for deletions and targeted mutations. Hence, timely and precise diagnosis of MDs helps clinicians to enroll eligible patients for therapy. The diagnosis of specific subtype of MDs using Next-generation sequencing (NGS) can be a timely and affordable approach which improves clinical prognosis (Bello and Pegoraro, 2016; Okubo et al., 2016; Zhang K. et al., 2019). Further, the NGS platforms also identified the novel variants as well as confirmation of hard-to-detect variants (Sheikh and Yokota, 2020). Recent studies suggest that the utilization of a high-throughput method using NGS platform is more suitable for clinical diagnosis (Okubo et al., 2016; Aravind et al., 2019). Moreover, the detection of large duplications is a major challenge for single-point diagnostic strategy (Okubo et al., 2016). In the present study, a total of 123 subjects (including both patients and female carriers) with suspected MDs were evaluated using an amplicon-based panel for its diagnostic specificity to detect CNVs and SNVs. Results of CNV analysis for the *DMD* gene were compared with MLPA. Further, we aimed to identify CNVs and SNVs type of mutation with our customized amplicon panel for different types of muscular dystrophies.

MATERIALS AND METHODS

Sample Collection and Genomic DNA Isolation

A total of 123 unrelated patients suspected of MD [DMD/BMD (N = 82), LGMD (N = 19), and UMD (N = 2)] and possible carriers (N = 20) were enrolled in this study. These cases were recruited in the study through screening camps across the state of Gujarat by collaborative efforts of the Indian Muscular Dystrophy Society (IMDS), Rashtriya Bal Swasthya Karyakram, and Gujarat Biotechnology Research Centre (GBRC). Informed and written consent was derived from the patients and their relatives after Genetic counseling. We have included patients clinically suspected with DMD/BMD with the following indications 1) significantly high serum creatine phosphokinase (CPK- >200 U/L (Aujla and Patel, 2020); 2) difficulty in walking, waddling gait, toe walk, Gower's sign or loss of ambulation; 3) and progressive muscle weakness. Patients with evident proximal muscle weakness mainly the shoulder girdle and pelvic were included in the study with query LGMD. Uncertain Muscular Dystrophies (UMDs) were included in the study for amplicon sequencing. Blood samples were collected in EDTA vacutainer in a standard blood collection setup. Genomic DNA was extracted from blood samples using the QIAamp DNA Blood Mini Kit (QIAGEN, Germany) as per the manufacturer's instructions. DNA

quantitation was done on Qubit 4 Fluorometer (Thermo Fisher Scientific, IN) using dsDNA BR (broad range) assay kit (Thermo Fisher Scientific, IN). For the data analysis, the baseline was generated using amplicon sequencing of 10 healthy male controls.

Customized Multi-Gene Panel

In the present study, a custom Ion AmpliSeq™ Panel which covers *DMD, SGCA, SGCB, SGCG, SGCD, CAPN3, ISPD, TCAP, TMEM43, TRIM32, FKRP, MYOT, POMT1, FKTN, POMT2, POMGNT1, DAG1, LMNA, ANO5, LAMA2, COL6A1, COL6A2, COL6A3, FHL1, DYSF, LARGE, TRAPPC11*, and *EMD* genes which are reported earlier and significantly associated with different pathologies of muscular dystrophies was designed. The association of these genes with different phenotypic abnormalities of MDs has been indicated in **Supplementary Table S1**. The panel comprises a total of 1,312 amplicon primer pools targeting the coding and untranslated regions (UTRs) with 10bp flanking regions of the mentioned genes.

Targeted Sequencing

A total of 123 cases (includes 102 DMD/BMD/Carrier, 2 UMD, and 19 LGMD) were screened by targeted sequencing using a custom amplicon panel. For each sample, 50 ng of DNA was amplified with custom primer pools using Ion Ampliseq™ HiFi Mix (Thermo Fisher Scientific, IN). This was followed by partial digestion, adaptor + barcode ligation, and library amplification. Libraries were purified using AgencourtAMPure XP (Beckman Coulter, United States). Purified libraries were quantified by Qubit dsDNA HS assay kit (Thermo Fisher Scientific, IN) and then pooled in equimolar concentrations. Emulsion PCR of pooled and diluted libraries was carried out using the Ion OneTouch™ 2 System (Thermo Fisher Scientific, IN) followed by enrichment of template-positive Ion Sphere™ Particles on an Ion OneTouch™ ES system (Thermo Fisher Scientific, IN). Sequencing was carried out on the Ion Proton™ and Ion S5™ systems using Ion PI and Ion 530 chips respectively, with an average depth of 80x.

MLPA

A total of 102 DMD/BMD/Carrier subjects were screened through MLPA for all exon deletions and duplications in the human *DMD* gene. MLPA was performed using SALSA MLPA kit P034/P035 (MRC-Holland, Netherlands) as per the manufacturer's instructions. Fragment analysis was performed on the 3500xL Genetic Analyzer (Applied Biosystems, United States) and MLPA data were analysed using Coffalyser Software (MRC-Holland, Netherlands).

Analysis of Single Nucleotide Variants

Analysis of the raw sequences was performed using Ion Torrent Suite software v5.12 on the Ion torrent server with the incorporated standard pipeline. Variant analysis pipeline includes, signal processing, base calling, quality score assignment, adaptor trimming, PCR duplicate removal, and read alignment to the human reference genome (hg19 genome build). Variants were identified with Torrent Variant Caller plugin software and the Coverage Analysis plugin software obtained coverage analysis. The poor quality and intronic mutations were discarded from the datasets. Annotation of the high-quality variants was performed using the Ion Reporter server system. Clinically known and reported variants like pathogenic or likely pathogenic were identified from the ClinVar database (Landrum et al., 2014). To check strand biases and sequencing errors in the variant calling, alignments were visualized and the presence of mutations in the datasets against the reference genome was confirmed using Integrative Genomics Viewer (IGV) (Robinson et al., 2011). Classification of variants was carried out as per the American College of Medical Genetics and Genomics recommendations (ACMG) for standard interpretation and reporting of sequence variations (Richards et al., 2015).

Identification of Large Homozygous and Heterozygous Deletions

For identification of large deletion/s, we used a CNV detection workflow available on the Ion Reporter Server system. For the CNV detection workflow, the base line was created using 10 healthy normal individual male samples. This baseline control was used as a reference to analyse CNV in patient samples and female carriers.

Analysis of Reading Frame

In the *DMD* gene, in-frame and out-of-frame mutation patterns were analyzed with a reading frame checker of online available DMD database [Leiden Muscular Dystrophy Pages. https://www.dmd.nl/, last accessed July 29, 2020.].

RESULTS

A total of 123 suspected MD patients and suspected female carriers were enrolled in the study. The mean age of onset for DMD and BMD cases was 13 and 22 years, respectively. Average CPK levels in clinically confirmed DMD and BMD cases were 6,551.08 U/L and 1,459.8 U/L, respectively. The normal range of CPK is 20–200 U/L (Aujla and Patel, 2020). The graph of CPK vs age is shown in **Figure 1**.

The Large Deletion in DMD/BMD

CNV analysis of the NGS data of suspected DMD and BMD cases revealed large deletions in 76/102 (74.5%) cases, which included 69 patients and 7 female carriers. As per the reading frame rule, considering only exon deletion, 63/102 (61.76%) and 13/102 (12.75%) cases were categorized into DMD and BMD, respectively. The majority of deletions (78.94%) were in the distal hotspot region (Exon 42–55) and proximal hotspot deletions were between exon 2–19 (10.52%). Two patients showed very large deletion including both proximal and distal hotspot regions. No deletion was observed in exon 61–79. The results of CNV analysis (*DMD* gene) using the NGS panel were concordant with the results obtained using MLPA (**Supplementary Table S2**). The deletion pattern of all positive

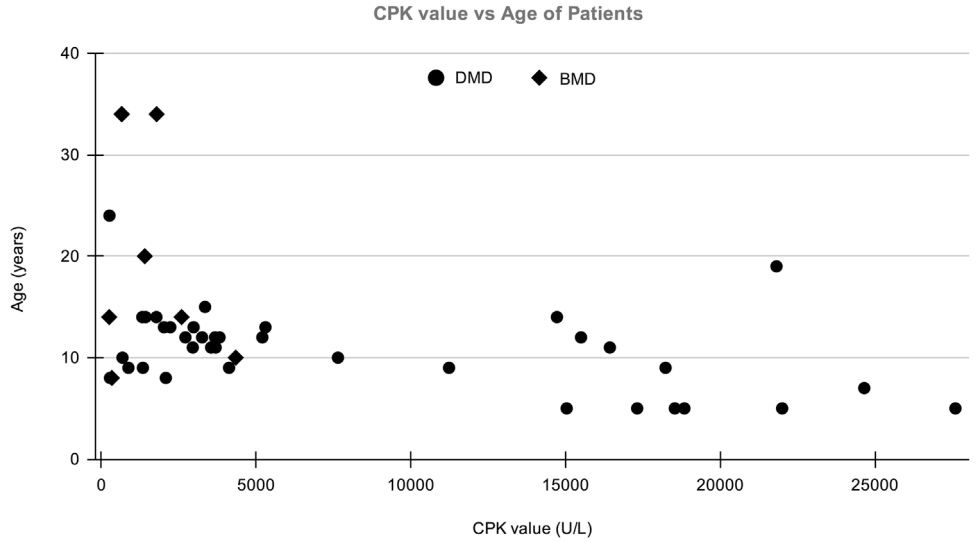

FIGURE 1 | The distribution of patient's Age vs CPK value in which *X*-axis shows CPK value (U/L) of patients and *Y*-axis shows the age of DMD/BMD patients in our study.

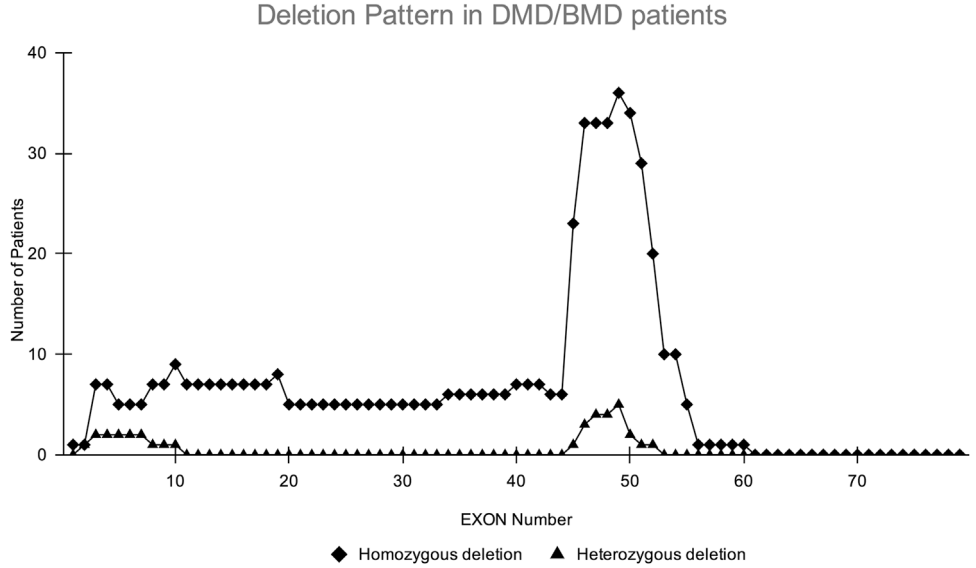

FIGURE 2 | Homozygous and heterozygous deletion (female carrier) pattern in DMD/BMD patients. The *X*-axis shows exon number of *DMD* gene and *Y*-axis shows the number of patients showing deletion in the study. The highest deletions frequency was found in exons 45–52.

cases is depicted in **Figure 2**. The largest deletion was seen in (exons 1 to 60 (P30) followed by exons 3 to 44 (P41), exons 3 to 42 (P46), and exons 3 to 41 (P55). In the proximal region, the most frequently deleted exon is 10 (10/76, 13.15%) followed by exon 3, and exon 4 (9/76, 11.84%), while in the distal region, the most common deleted exon is 49 (41/76, 53.94%), followed by exon 48 and exon 47 (37/76, 48.68%). Single exon deletion was observed in 9/76 (11.84%) patients, where the most common deletion was observed in exon 45 followed by exon 51. More than one exon deletion was identified in 67/76 patients (88.15%). The distribution of the mutation pattern is shown in **Figure 3**. In this study, exon 45 to 52 was identified as a major deletion hotspot region. Representative IGV tool image of large deletion is shown in **Supplementary Figure S1**.

The Large Deletion in LGMD

Amplicon sequencing analysis of 19 LGMD suspected cases revealed a homozygous deletion of exon 17 to 24 of the

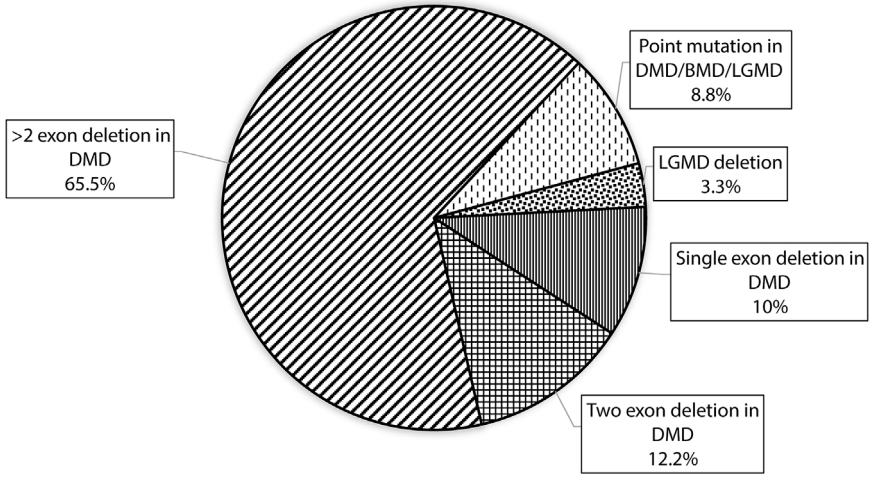

FIGURE 3 | Figure showing the distribution of mutation patterns with designed custom amplicon panel in our study.

TABLE 1 | Summary of point mutations in LGMD and UMD cases identified using panel-based NGS sequencing.

Patient ID	CPK total	Age	Locus	Location	Function	Exon	Gene-cDNA	ClinVar's clinical significance	dbSNP ID	Clinical phenotypesof disease
P52	95	40	chr21: 47537326	Exonic	missense	11	*COL6A2*-c.1012C>T	Uncertain significance	rs775751831	LGMD R22 Collagen6-related/ Ullrich CMD type 1
P60	NA	9	chr6: 129704300	Exonic	missense	35	*LAMA2*-c.4993G>A	Uncertain significance	rs373997222	CMD due to partial LAMA2 deficiency
P83	3,675	18	chr9: 134397500	Exonic	missense	19	*POMT1*-c.1958C>T	Pathogenic	rs149682171	LGMD R5 γ-sarcoglycan-related
P92	1,631	28	chr2: 71788881	Splice site	unknown	23	*DYSF*-c.2217-1G>T	Pathogenic	rs886044379	LGMD R2 dysferlin-related
P97	653	21	chr21: 47419593	Exonic	missense	27	*COL6A1*-c.1763C>T	Uncertain significance	rs759442615	LGMD R22 Collagen6-related/ Ullrich CMD type 1
P99	995	32	chr4: 52895918	Exonic	missense	3	*SGCB*-c.355A>T	Uncertain significance	rs762412447	LGMD R4 β-sarcoglycan-related
P100	679	34	chr4: 52895918	Exonic	missense	3	*SGCB*-c.355A>T	Uncertain significance	rs762412447	LGMD R4 β-sarcoglycan-related
P109	3,673	12	chrX: 32398743	Exonic	missense	34	*DMD*-c.4729C>T	Pathogenic	rs863224999	DMD

DMD, duchenne muscular dystrophy; LGMD, Limb-girdle muscular dystrophies; CMD, congenital muscular dystrophy; CPK, creatine phosphokinase; NA, not available.
Note, 2018 Note: LGMDs, were described according to new nomenclature proposed by ENMC, Consortium (Straub et al., 2018) and Bethlem myopathy was described as a type of LGMD (Angelini et al., 2018).

CAPN3 gene in 3 patients (15.78%, 3/19). It was further confirmed by visualizing in IGV (**Supplementary Figure S2**), which supports the results of CNV workflow. *CAPN3* deletion results are consistent with the clinical presentation of LGMD type 2A disease.

Single Nucleotide Variation

MLPA and NGS (CNV) negative cases (N = 44) were further considered for SNV analysis where in patient P109, a pathogenic hemizygous mutation was found in the *DMD* gene. This mutation causes a premature translational stop signal at codon 4,729 (p. Arg1577*) of the *DMD* gene, which results in a disrupted protein product. Truncating variants in the DMD gene are known to be pathogenic as per the Clinvar database (Landrum et al., 2014). This variant previously has been reported in individuals affected with DMD (Mah et al., 2011; Yang et al., 2013). In the case of suspected LGMD, mutations in 3 different genes were identified in 4 patients (P99 and P100 are sisters). Two pathogenic mutations were observed in *POMT1* (P83) and *DYSF* (P92) genes and other 3 VUS mutations were observed in *LAMA2* (P60), and *SGCB* (P99 and P100) genes. SNV analysis also revealed a missense Variant of Uncertain Significance (VUS) in 1 UMD and 1 suspected LGMD case in *COL6A2* (P52) and *COL6A1* (P97) gene, respectively (**Table 1**) which causes Bethlem myopathy 1 (LGMD R22 Collagen6-related) disease as per the Clinvar database (Landrum et al., 2014).

DISCUSSION

In this study, we showed the utility of an amplicon panel to detect CNVs and SNVs to diagnose a heterogeneous group of MDs in patients and carriers. The accurate diagnosis of different types of muscular dystrophies using a single method such as Sanger sequencing or MLPA is a big challenge due to the complex mutational spectrum. MLPA being a first-line test for the diagnosis of the most common type of MD (DMD/BMD), to detect SNVs sequencing is mandatory. However, Sanger sequencing of the large coding region becomes laborious as well as costly (Wang et al., 2014; Wei et al., 2014). Hence, NGS could be the better alternative in terms of cost, since per base sequencing cost has decreased drastically (Pareek et al., 2011). Also, in cases with LGMDs and CMDs, neuromuscular disease-specific panels at a lower cost can be beneficial in developing countries like India. Previously, many studies have been published for NGS-based approaches for MD (Lim et al., 2011; Wang et al., 2014; Wei et al., 2014; Alame et al., 2016). However, there are very few such studies reported for the Indian population (Aravind et al., 2019; Ganapathy et al., 2019; Polavarapu et al., 2019).

We customized an amplicon panel consisting of genes targeting 29 different types of muscular dystrophies. One of the major objectives of the present study was to detect point mutation and CNVs in suspected, DMD/BMD patients/carriers, LGMD, and CMD using an NGS-based amplicon panel. Our CNV results of the *DMD* gene are consistent with the MLPA results. Our findings support the idea that NGS-based diagnosis methods could be routinly employed as a single diagnostic screening method for the most frequent type of MDs. Further, different MDs can be characterized by genotype and phenotype correlation. Earlier reports highlighted the importance of respective mutations in DMD patients and their mutation-specific therapies (Kohli et al., 2020). Identification of mutation patterns in the Indian cohort could improve the therapeutic management. In the mutation analysis, deletion was observed almost in each exon of *DMD* gene except 61–79 exons, where some deletions are reported in very low frequency in Leiden Open Variation database (LOVD) (Aartsma-Rus et al., 2006). Such as deletions are exon 19–45 (P15), 10–19 (P27 and P108), 1–60 (P30), and in 3–41 (P55). Furthermore, two novel out-of-frame deletions (exon 8–30 in P38 and 46–55 in P61) observed in our study are not reported in the LOVD database. The majority of deletions were observed in the hotspot region of exon 45–52. Interestingly, during sample collection, three patients were enrolled phenotypically as a DMD patients however, our results concluded them as BMD patients with the in-frame mutation. The variants were confirmed in reading frame checker of LOVD database (Takeshima et al., 1994). In SNV analysis, a nonsense variant (c.4729C>T in exon 34) was observed in the *DMD* gene in patient P109. This point mutation has been recorded in the LOVD database as a pathogenic variant, which leads to a premature termination codon (p. Arg1577*) and hence forms a truncated protein. Earlier report suggests that the patients with such mutation are affected with DMD (Esterhuizen et al., 2014). Further, we have found four SNVs, *COL6A2*-c.1012C>T, *LAMA2*-c.4993G>A,*COL6A1*-c.1763C>T, *SGCB*-c.355A>T, reported as VUS in Clinvar database due to their conflicting reports and prediction from various computational tools and require further characterizations. NGS analysis of LGMD patients for CNV analysis revealed a homozygous deletion in the *CAPN3* gene in exon 17 to 24 which is already reported in our previous study (Bhatt et al., 2019). Identification of the same mutation in 3 patients in the current study accelerates the proof of novel variants in our population. Mutation in the *CAPN3* gene leads to the most common form of autosomal recessive LGMD-2A type of Muscular Dystrophy. Deletion in 17–24 exons results in short truncated non-functional CAPN3 protein (Bhatt et al., 2019). *CAPN3* gene regulates the instruction for Calpain-3 enzyme production that enzyme found in sarcomeres structure of muscle cells which are the basic unit for muscle contraction (Kramerova et al., 2007).

Characterization of the mutational landscape in the population may increase the success of current therapeutics and may provide direction to develop novel drug candidates. Antisense oligonucleotide (AON)-mediated exon skipping approach is currently developed to restore reading frame rule which produces partially function protein in DMD patients. FDA approved drugs such as EXONDYS 51 and VYONDYS 53 are commercially available to treat the patient who has a confirmed mutation to skip exon 51 and 53 respectively [http://www.aetna.com/cpb/medical/data/900_999/0911.html, last accessed 29 July 2020.]. The limitation of the present study is the very low frequency of point mutations and therefore further sampling is required to validate such mutation with our panel. Moreover, from a total of 123 cases, no mutation is detected in 36 cases suggesting further testing is required for other neuromuscular diseases which may rule out in our panel.

CONCLUSION

In conclusion, our finding showed the NGS platform could be a future diagnostic tool for identifying disease-causing mutation/s in different Muscular dystrophies, which are currently diagnosed using multiple methods. The analysis of CNV in the *DMD* gene concludes that our custom panel is superior to the MLPA method. NGS-based diagnosis is not only time-saving but also cost-effective method when compared with traditional testing strategies.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=692346. Accession number is PRJNA692346.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Foundation for Research in Genetics and Endocrinology, Institute of Human Genetics. Written

informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

CJ and MJ: Conceptualization, Methodology, Supervision; KP and AB: Data curation, analysis and interpretation of data, Writing—Original draft preparation; RP, BW, and HS: Thoroughly editing of the manuscript and interpretation of data; KS: Genetic counselling of patients.

FUNDING

This work was funded by the Department of Science and Technology, Government of Gujarat. Grant Number is GBRC/GOG/DST/JD1/HLT/2017-18/06.

ACKNOWLEDGMENTS

We are thankful to the patients and their families for registering themselves and providing consent for participation in this study. We also thank the Department of Science and Technology (DST), Government of Gujarat, Gandhinagar for the financial assistance. We would like to thank Mr. Bakulesh Nagar for helping us in the successful conduction of the screening camp. We would also want to thank Rashtriya Bal Swasthya Karyakram (RBSK) and the Indian Muscular Dystrophy Society (IMDS) for their contribution to camp conduction. We would also like to thank our lab mate, Ms. Mital Patel, for participating in camps and Dr. ApurvasinhPuvar who helped us get over any obstacles we faced during different phases of the study. We would also like to thank the Institute of Human Genetics for providing ethical committee approval for conducting research on muscular dystrophy.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.770350/full#supplementary-material

Supplementary Figure S1 | Integrative Genomics Viewer (IGV) snapshot depicting a representative image of Ion Reporter software (v5.12) CNV analysis results in DMD gene where red bar showing the deleted region of exon 3-4 in the patient.

Supplementary Figure S2 | Integrative Genomics Viewer (IGV) snapshot depicting a representative image of Ion Reporter software (v5.12) CNV analysis results in CAPN3 gene where red bar showing the deleted region of exon 17-24 in the patient.

REFERENCES

Aartsma-Rus, A., Van Deutekom, J. C., Fokkema, I. F., Van Ommen, G. J., and Den Dunnen, J. T. (2006). Entries in the Leiden Duchenne Muscular Dystrophy Mutation Database: an Overview of Mutation Types and Paradoxical Cases that Confirm the reading-frame Rule. *Muscle Nerve.* 34, 135–144. doi:10.1002/mus.20586

Aartsma-Rus, A., Ginjaar, I. B., and Bushby, K. (2016). The Importance of Genetic Diagnosis for Duchenne Muscular Dystrophy. *J. Med. Genet.* 53, 145–151. doi:10.1136/jmedgenet-2015-103387

Alame, M., Lacourt, D., Zenagui, R., Mechin, D., Danton, F., Koenig, M., et al. (2016). Implementation of a Reliable Next-Generation Sequencing Strategy for Molecular Diagnosis of Dystrophinopathies. *J. Mol. Diagn.* 18, 731–740. doi:10.1016/j.jmoldx.2016.05.003

Angelini, C., Giaretta, L., and Marozzo, R. (2018). An Update on Diagnostic Options and Considerations in Limb-Girdle Dystrophies. *Expert Rev. Neurotherapeutics* 18 (9), 693–703. doi:10.1080/14737175.2018.1508997

Aravind, S., Ashley, B., Mannan, A., Ganapathy, A., Ramesh, K., Ramachandran, A., et al. (2019). Targeted Sequencing of the DMD Locus: A Comprehensive Diagnostic Tool for All Mutations. *Indian J. Med. Res.* 150, 282–289. doi:10.4103/ijmr.IJMR_290_18

Aujla, R. S., and Patel, R. (2020). *Creatine Phosphokinase*. Treasure Island, FL: StatPearls Publishing. doi:10.5772/intechopen.85339

Bello, L., and Pegoraro, E. (2016). Genetic Diagnosis as a Tool for Personalized Treatment of Duchenne Muscular Dystrophy. *Acta Myol.* 35, 122–127.

Bhatt, A. D., Puvar, A., Shah, K., Joshi, C. G., and Joshi, M. (2019). A Case of Limb Girdle Muscular Dystrophy Type 2A from India: Copy Number Variation Analysis Using Targeted Amplicon Sequencing. *J. Clin. Diagn. Res.* 13. doi:10.7860/jcdr/2019/40923.12812

Esterhuizen, A. I., Wilmshurst, J. M., Goliath, R. G., and Greenberg, L. J. (2014). Duchenne Muscular Dystrophy: High-Resolution Melting Curve Analysis as an Affordable Diagnostic Mutation Scanning Tool in a South African Cohort. *S Afr. Med. J.* 104, 779–784. doi:10.7196/samj.8257

Gaina, G., Budisteanu, M., Manole, E., and Ionica, E. (2019). *Clinical and Molecular Diagnosis in Muscular Dystrophies*. London: IntechOpen. doi:10.5772/intechopen.85339

Ganapathy, A., Mishra, A., Soni, M. R., Kumar, P., Sadagopan, M., Kanthi, A. V., et al. (2019). Multi-gene Testing in Neurological Disorders Showed an Improved Diagnostic Yield: Data from over 1000 Indian Patients. *J. Neurol.* 266, 1919–1926. doi:10.1007/s00415-019-09358-1

Iyadurai, S. J. P., and Kissel, J. T. (2016). The Limb-Girdle Muscular Dystrophies and the Dystrophinopathies. *CONTINUUM: Lifelong Learn. Neurol.* 22, 1954–1977. doi:10.1212/con.0000000000000406

Kohli, S., Saxena, R., Thomas, E., Singh, K., Bijarnia Mahay, S., Puri, R. D., et al. (2020). Mutation Spectrum of Dystrophinopathies in India: Implications for Therapy. *Indian J. Pediatr.* 87, 495–504. doi:10.1007/s12098-020-03286-z

Kong, X., Zhong, X., Liu, L., Cui, S., Yang, Y., and Kong, L. (2019). Genetic Analysis of 1051 Chinese Families with Duchenne/Becker Muscular Dystrophy. *BMC Med. Genet.* 20, 139. doi:10.1186/s12881-019-0873-0

Kramerova, I., Beckmann, J. S., and Spencer, M. J. (2007). Molecular and Cellular Basis of Calpainopathy (Limb Girdle Muscular Dystrophy Type 2A). *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1772, 128–144. doi:10.1016/j.bbadis.2006.07.002

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: Public Archive of Relationships Among Sequence Variation and Human Phenotype. *Nucl. Acids Res.* 42, D980–D985. doi:10.1093/nar/gkt1113

Lim, B. C., Lee, S., Shin, J.-Y., Kim, J.-I., Hwang, H., Kim, K. J., et al. (2011). Genetic Diagnosis of Duchenne and Becker Muscular Dystrophy Using Next-Generation Sequencing Technology: Comprehensive Mutational Search in a Single Platform. *J. Med. Genet.* 48, 731–736. doi:10.1136/jmedgenet-2011-100133

Mah, J. K., Selby, K., Campbell, C., Nadeau, A., Tarnopolsky, M., McCormick, A., et al. (2011). A Population-Based Study of Dystrophin Mutations in Canada. *Can. J. Neurol. Sci.* 38, 465–474. doi:10.1017/s0317167100011896

Mohammed, F., Elshafey, A., Al-Balool, H., Alaboud, H., Al Ben Ali, M., Baqer, A., et al. (2018). Mutation Spectrum Analysis of Duchenne/Becker Muscular Dystrophy in 68 Families in Kuwait: The Era of Personalized Medicine. *PloS one* 13, e0197205. doi:10.1371/journal.pone.0197205

Murphy, A. P., and Straub, V. (2015). The Classification, Natural History and Treatment of the Limb Girdle Muscular Dystrophies. *Jnd* 2, S7–S19. doi:10.3233/jnd-150105

Nallamilli, B. R. R., Chakravorty, S., Kesari, A., Tanner, A., Ankala, A., Schneider, T., et al. (2018). Genetic Landscape and Novel Disease Mechanisms from a largeLGMDcohort of 4656 Patients. *Ann. Clin. Transl Neurol.* 5, 1574–1587. doi:10.1002/acn3.649

Okubo, M., Minami, N., Goto, K., Goto, Y., Noguchi, S., Mitsuhashi, S., et al. (2016). Genetic Diagnosis of Duchenne/Becker Muscular Dystrophy Using Next-Generation Sequencing: Validation Analysis of DMD Mutations. *J. Hum. Genet.* 61, 483–489. doi:10.1038/jhg.2016.7

Pareek, C. S., Smoczynski, R., and Tretyn, A. (2011). Sequencing Technologies and Genome Sequencing. *J. Appl. Genet.* 52, 413–435. doi:10.1007/s13353-011-0057-x

Pegoraro, E., and Hoffman, E. P. (2012). "Limb-Girdle Muscular Dystrophy Overview – RETIRED CHAPTER, FOR HISTORICAL REFERENCE ONLY," in *GeneReviews® [Internet]*. Editors M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. H. Bean, G. Mirzaa, et al. (Seattle, WA: University of Washington, Seattle), 1993–2021.

Polavarapu, K., Preethish-Kumar, V., Sekar, D., Vengalil, S., Nashi, S., Mahajan, N. P., et al. (2019). Mutation Pattern in 606 Duchenne Muscular Dystrophy Children with a Comparison between Familial and Non-familial Forms: a Study in an Indian Large Single-center Cohort. *J. Neurol.* 266, 2177–2185. doi:10.1007/s00415-019-09380-3

Richards, S., Aziz, N., Aziz, N., Bale, S., Bick, D., Das, S., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–423. doi:10.1038/gim.2015.30

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative Genomics Viewer. *Nat. Biotechnol.* 29, 24–26. doi:10.1038/nbt.1754

Sheikh, O., and Yokota, T. (2020). Advances in Genetic Characterization and Genotype-Phenotype Correlation of Duchenne and Becker Muscular Dystrophy in the Personalized Medicine Era. *Jpm* 10 (3), 111. doi:10.3390/jpm10030111

Straub, V., Murphy, A., Udd, B., Corrado, A., Aymé, S., Bönneman, C., et al. (2018). 229th ENMC International Workshop: Limb Girdle Muscular Dystrophies - Nomenclature and Reformed Classification Naarden, the Netherlands, 17-19 March 2017. *Neuromuscul. Disord.* 28 (8), 702–710. doi:10.1016/j.nmd.2018.05.007

Takeshima, Y., Nishio, H., Narita, N., Wada, H., Ishikawa, Y., Ishikawa, Y., et al. (1994). Amino-terminal Deletion of 53% of Dystrophin Results in an Intermediate Duchenne-Becker Muscular Dystrophy Phenotype. *Neurology* 44 (9), 1648–1651. doi:10.1212/wnl.44.9.1648

Valencia, C. A., Ankala, A., Rhodenizer, D., Bhide, S., Littlejohn, M. R., Keong, L. M., et al. (2013). Comprehensive Mutation Analysis for Congenital Muscular Dystrophy: a Clinical PCR-Based Enrichment and Next-Generation Sequencing Panel. *PloS one* 8, e53083. doi:10.1371/journal.pone.0053083

Wang, D., Gao, M., Zhang, K., Jin, R., Lv, Y., Liu, Y., et al. (2019). Molecular Genetics Analysis of 70 Chinese Families with Muscular Dystrophy Using Multiplex Ligation-dependent Probe Amplification and Next-Generation Sequencing. *Front. Pharmacol.* 10, 814. doi:10.3389/fphar.2019.00814

Wang, Y., Yang, Y., Liu, J., Chen, X.-C., Liu, X., Wang, C.-Z., et al. (2014). Whole Dystrophin Gene Analysis by Next-Generation Sequencing: a Comprehensive Genetic Diagnosis of Duchenne and Becker Muscular Dystrophy. *Mol. Genet. Genomics* 289, 1013–1021. doi:10.1007/s00438-014-0847-z

Wei, X., Dai, Y., Yu, P., Qu, N., Lan, Z., Hong, X., et al. (2014). Targeted Next-Generation Sequencing as a Comprehensive Test for Patients with and Female Carriers of DMD/BMD: a Multi-Population Diagnostic Study. *Eur. J. Hum. Genet.* 22, 110–118. doi:10.1038/ejhg.2013.82

Wicklund, M. P. (2013). The Muscular Dystrophies. *CONTINUUM: Lifelong Learn. Neurol.* 19, 1535–1570. doi:10.1212/01.con.0000440659.41675.8b

Wu, B., Wang, L., Dong, T., Jin, J., Lu, Y., Wu, H., et al. (2017). Identification of a Novel DMD Duplication Identified by a Combination of MLPA and Targeted Exome Sequencing. *Mol. Cytogenet.* 10, 8–6. doi:10.1186/s13039-017-0301-0

Yang, J., Li, S. Y., Li, Y. Q., Cao, J. Q., Feng, S. W., Wang, Y. Y., et al. (2013). MLPA-based Genotype-Phenotype Analysis in 1053 Chinese Patients with DMD/BMD. *BMC Med. Genet.* 14, 29–9. doi:10.1186/1471-2350-14-29

Zhang, K., Yang, X., Lin, G., Han, Y., and Li, J. (2019a). Molecular Genetic Testing and Diagnosis Strategies for Dystrophinopathies in the Era of Next Generation Sequencing. *Clinica Chim. Acta* 491, 66–73. doi:10.1016/j.cca.2019.01.014

Zhang, Y., Yang, W., Wen, G., Wu, Y., Jing, Z., Li, D., et al. (2019b). Application Whole Exome Sequencing for the Clinical Molecular Diagnosis of Patients with Duchenne Muscular Dystrophy; Identification of Four Novel Nonsense Mutations in Four Unrelated Chinese DMD Patients. *Mol. Genet. Genomic Med.* 7, e622. doi:10.1002/mgg3.622

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

GLOSSARY

ANO5 anoctamin 5

CAPN3 calpain 3

COL6A1 collagen type VI alpha 1 chain

COL6A2 collagen type VI alpha 2 chain

COL6A3 collagen type VI alpha 3 chain

DAG1 dystroglycan 1

DMD Duchenne Muscular Dystrophy

DYSF dystrophy-associated fer-1-like protein

EMD Emerin

FHL1 Four-and-a-Half Lim Domains 1

FKRP fukutin related protein

FKTN fukutin

ISPD isoprenoid synthase domain-containing protein

LMNA lamin A/C

LAMA2 laminin subunit alpha 2

LARGE LARGE xylosyl- and glucuronyltransferase 1

MYOT myotilin

NGS Next Generation Sequencing

POMT1 protein O-mannosyltransferase 1

POMT2 protein O-mannosyltransferase 2

POMGNT1 protein O-linked mannose N-acetylglucosaminyltransferase 1 (beta 1,2-)

SGCA Sarcoglycan Alpha

SGCB Sarcoglycan Beta (43 kDa Dystrophin-Associated Glycoprotein)

SGCG Sarcoglycan Gamma

SGCD Sarcoglycan Delta

TCAP Titin-Cap associated protein

TMEM43 Transmembrane Protein 43

TRAPPC11 trafficking protein particle complex 11

TRIM32 tripartite motif containing 32

VUS Variants of Uncertain Significance

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fgene.2021.800587

Methyltransferase as Antibiotics Against Foodborne Pathogens: An *In Silico* Approach for Exploring Enzyme as Enzymobiotics

Varish Ahmad[1], Aftab Ahmad[1], Mohammed F. Abuzinadah[2], Salwa Al-Thawdi[3] and Ghazala Yunus[4]*

[1]Health Information Technology Department, Faculty of Applied Studies, King Abdulaziz University, Jeddah, Saudi Arabia, [2]Department of Medical Laboratory Technology, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, [3]Department of Biology, College of Science, University of Bahrain, Sakhir, Bahrain, [4]Department of Basic Science, University of Hail, Hail, Saudi Arabia

The development of resistance in microbes against antibiotics and limited choice for the use of chemical preservatives in food lead the urgent need to search for an alternative to antibiotics. The enzymes are catalytic proteins that catalyze digestion of bacterial cell walls and protein requirements for the survival of the cell. To study methyltransferase as antibiotics against foodborne pathogen, the methyltransferase enzyme sequence was modeled and its interactions were analyzed against a membrane protein of the gram-positive and gram-negative bacteria through *in silico* protein–protein interactions. The methyltransferase interaction with cellular protein was found to be maximum, due to the maximum PatchDock Score (15808), which was followed by colicin (12864) and amoxicillin (4122). The modeled protein has found to be interact more significantly to inhibit the indicator bacteria than the tested antibiotics and antimicrobial colicin protein. Thus, model enzyme methyltransferase could be used as enzymobiotics. Moreover, peptide sequences similar to this enzyme sequence need to be designed and evaluated against the microbial pathogen.

Keywords: methyltransferase, antimicrobial, drug resistant, protein–protein interaction, enzymobiotics

OPEN ACCESS

Edited by:
Saumya Patel,
Gujarat University, India

Reviewed by:
Khurshid Ahmad,
Yeungnam University, South Korea
M. Wahajuddin,
Central Drug Research Institute (CSIR), India

****Correspondence:***
Varish Ahmad
vaahmad@kau.edu.sa

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *23 October 2021*
Accepted: *17 November 2021*
Published: *03 January 2022*

Citation:
Ahmad V, Ahmad A, Abuzinadah MF, Al-Thawdi S and Yunus G (2022) Methyltransferase as Antibiotics Against Foodborne Pathogens: An In Silico Approach for Exploring Enzyme as Enzymobiotics. Front. Genet. 12:800587. doi: 10.3389/fgene.2021.800587

INTRODUCTION

In fact, chemotherapy has renovated the treatments not only against bacterial disease but also fungal diseases. However, many pathogens become protective against available antibiotics and pose a threat to the health of humans and animals. Various alternatives to antibiotics such as probiotics, nanobiotics, antimicrobial peptides or bacteriocin, CRISPR-Cas, quorum-sensing inhibitors, phage therapy, and immunotherapy exist (Kumar et al., 2021).

The enzymes are proteinaceous molecules and known as biocatalysts or endopeptidases. Recent research has reported that enzymes could be used as a special class of antimicrobial enzymobiotics, against microbial infections and to control the drug-resistant microbes.

Enzymes play a significant role in the expression of cellular proteins, cell wall polysaccharides, nucleic acids, and other cellular metabolites that are required for the survival of the cell. The use of enzymes as bacteriophage holins and their membrane-disrupting activity, anti-staphylococcal lytic enzymes, and membrane-targeted antibiotics have been recently highlighted by much research.

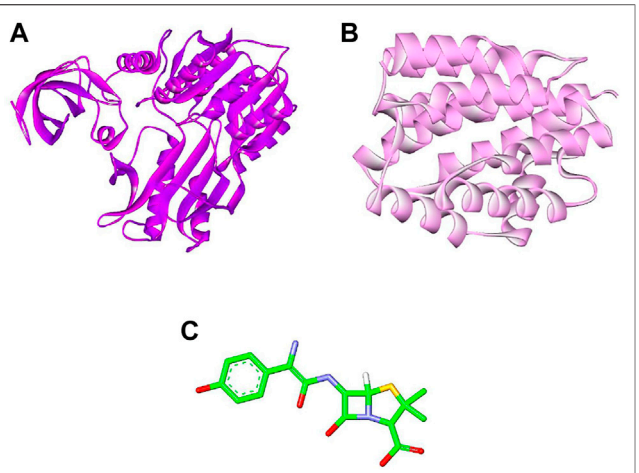

FIGURE 1 | Selected molecules as ligand **(A)** methyltransferase (purple color), **(B)** colicin protein (light pink color), and **(C)** amoxicillin (green color).

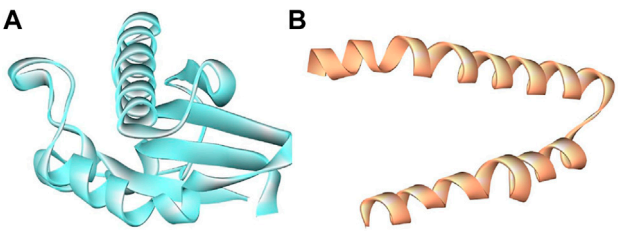

FIGURE 2 | Selected biomolecules as receptor **(A)** *E. coli* membrane protein (light turquoise color) and **(B)** *Staphylococcus aureus* membrane protein (golden color).

Enzymes like glucose oxidase, hydrogen peroxidase, and protease have inhibitory effects on microbial pathogens. The bacteriostatic and inhibitory effect on biofilm formation against many pathogenic bacteria like *P. aeruginosa*, *S. aureus*, and methicillin-resistant *S. aureus*, with a patented formulation of glucose oxidase, was recently explored by Cooper (2013). The glucose oxidase reported inhibiting *Staphylococcus* cells more potent than the *P. aeruginosa* cells (Cooper, 2013). Bacterial lipopolysaccharides (LPS) are involved in maintaining intestinal homeostasis and mediate potent pro-inflammatory toxins/mediators. Thus, apical brush borders rich in alkaline phosphatase are analyzed as a significant de-phosphorylation molecule for the neutralization of LPS in addition to un-methylated cytosine-guanosine dinucleotides and flagellin, resulting in reduced toxicity and inflammatory responses (Drago-Serrano et al., 2012).

A recent study was conducted to investigate the effect of dietary proteases on nutrient digestibility, growth performance, crude protein digestibility, enzyme (pepsin, pancreatic amylase, and trypsin) activities, plasma total proteins, intestinal villus heights, intestinal morphology, and the expression levels of specific genes. Significant increases in growth performance have been observed which were attributed to better intestinal development, enhanced protein digestibility, and improved nutrient transport efficiencies. The supplementation of proteases (200 and 300 mg/kg) within the diet increased the ratio of villus heights to crypt depth significantly, especially in the duodenum, jejunum, and ileum, and induced higher expression levels of the peptide transporter 1 (PepT1) within the duodenum region (Zuo et al., 2015).

The microorganisms are very sensitive to utilizing the nutrients from crops through microbial enzymatic actions. The main enzymes that help to initiate the deterioration are the first attacker on the cell wall, and they are popularly known as cell wall degrading enzymes. The cell wall degrading enzymes could be employed as antibiotics. Many proteinaceous molecules like bacteriocin produced from plants, animals, and microbes have been tested as potential therapeutic molecules. Initially, a homogeneous microbial population has grown and started the deterioration that is further exposed to the new environment to favor the growth of other pathogens. This resulted in the growth of heterogeneous microbial populations on the same habitats to initiate spoilage or pathogenesis by damaging the cellular components, thus helping tissue attack and microbial dissemination (Kikot et al., 2009). To inhibit the growth of these foodborne pathogens, chemical preservatives are not the preferred choice for food (Chukwuka et al., 2010).

Moreover, the microbial resistance against currently used antibiotics has raised serious human and animal health issues globally. This antimicrobial-resistant is well reported in many microbes including bacteria and fungi against last-line antibiotics, signifying a future loss of the therapeutic option to treat the infections. Many scientific strategies have also been tested to combat the drug-resistant microbes (Sartelli et al., 2017). The developed countries and developing countries have many challenges that can spread and stimulate the emergence of multidrug-resistant pathogen among microbial populations. The drug-resistant bacteria like *Pseudomonas aeruginosa* (*P. aeruginosa*), *K. pneumonia*, *Streptococcus pneumoniae*, and *Staphylococcus aureus* have been well reported and recognized as a global threat (Ventola, 2015). Managing these challenges need many scientific efforts that explore microbial resistance and the designing of effective controlling strategies such as active surveillance that stop the development and spread of drug-resistant microbes in the country. Moreover, improper use of antibiotics, infection inhibition, and control safeguards should also be improved to limit further spread. Therefore, it is highly important to explore the alternatives to antibiotics, such as the use of antimicrobial peptides, bacitracin, or lactic acid bacteria must be promoted for the primary control of microbes. Therefore, this *in silico* based study explores the use of the enzyme methyltransferase sequence as antibiotics against gram-negative and gram-positive bacteria.

MATERIALS AND METHODS

Preparation of Ligand Molecules

The structural information of amoxicillin (AMX) was retrieved from the DrugBank database (https://go.drugbank.com/drugs/

TABLE 1 | Showing quality assessment results of modeled 3D structures.

Model	MolProbity score	Clash score	Ramachandran favored	Ramachandran outliers	Rotamer outliers	C-Beta deviations
Ideal case	As low as possible	0	>98%	<0.2%	<1%	0
Methyltransferase	1.04	0.70	95.55%	1.11%	0.51%	1
Membrane protein (*Escherichia coli*)	1.77	4.56	95.16%	0.81%	1.83%	2
Membrane protein (*Staphylococcus aureus*)	0.50	0.00	98.11%	0.00%	0.00%	0

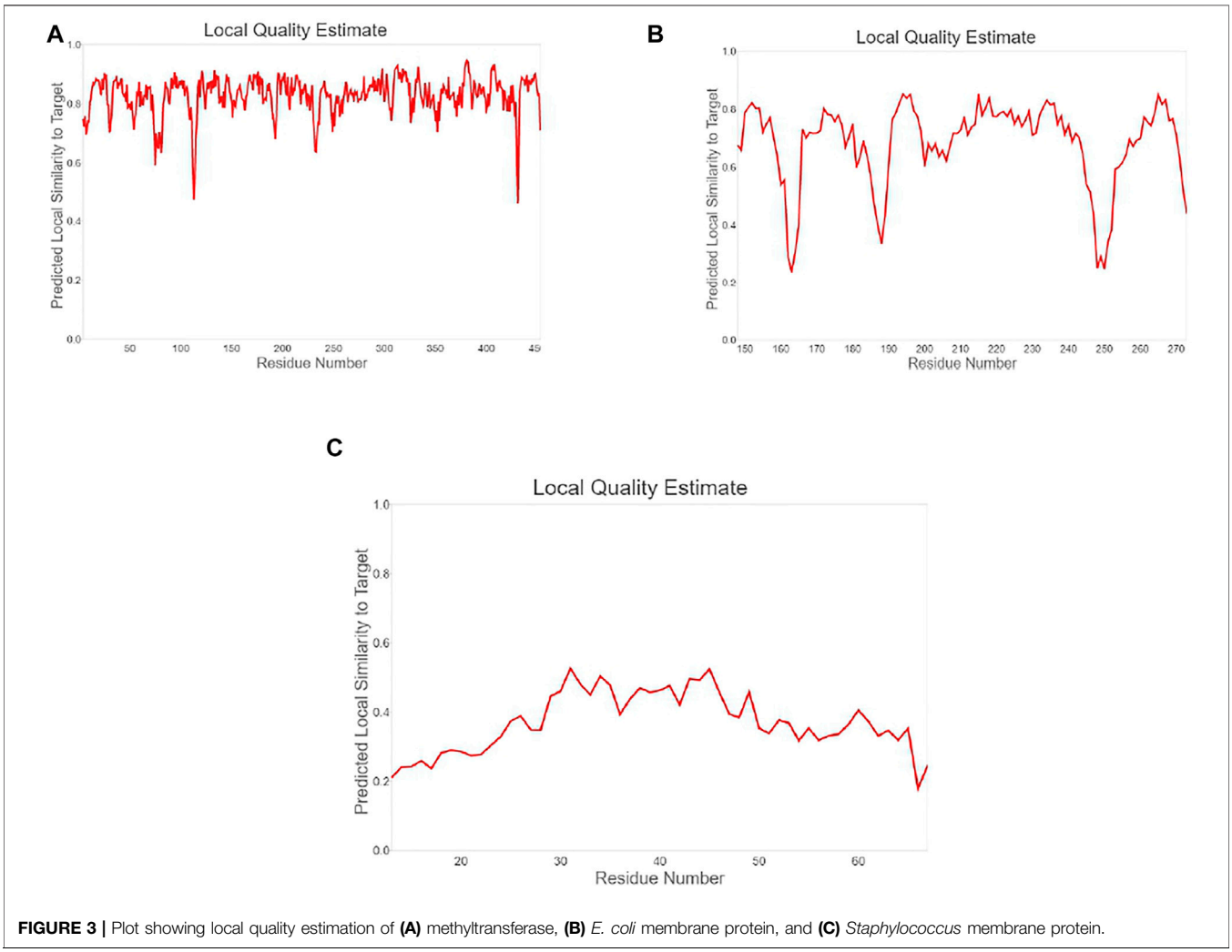

FIGURE 3 | Plot showing local quality estimation of **(A)** methyltransferase, **(B)** *E. coli* membrane protein, and **(C)** *Staphylococcus* membrane protein.

DB01060) (Wishart et al., 2008) (**Figure 1C**). The 3D structure of 23S rRNA [uracil (1939)-C (5)]-methyltransferase RlmD (*Pediococcus acidilactici*) (Sequence ID: WP_004165491.1) sequence was obtained from the National Centre for Biotechnology Information (NCBI) and modeled using the SWISS MODEL server (Waterhouse et al., 2018) (**Figure 1A**). The colicin structure was downloaded from the Protein Data Bank (RCSB PDB - 1COL: refined structure of the pore-forming domain of colicin a at 2.4 angstroms resolution) (**Figure 1B**).

Preparation of Receptor Molecules (SWISS-MODEL Workspace/GMQE)

We have accessed the PDB database for the receptor molecules but did not find them. Therefore, the 3D structures of membrane protein (*Escherichia coli*) (accession no.: APJ97041.1) (**Figure 2A**) and membrane protein (*Staphylococcus aureus*) (accession no.: KII21430.1) (**Figure 2B**) were modeled after retrieving their sequences in the FASTA format from the NCBI and provided as an input

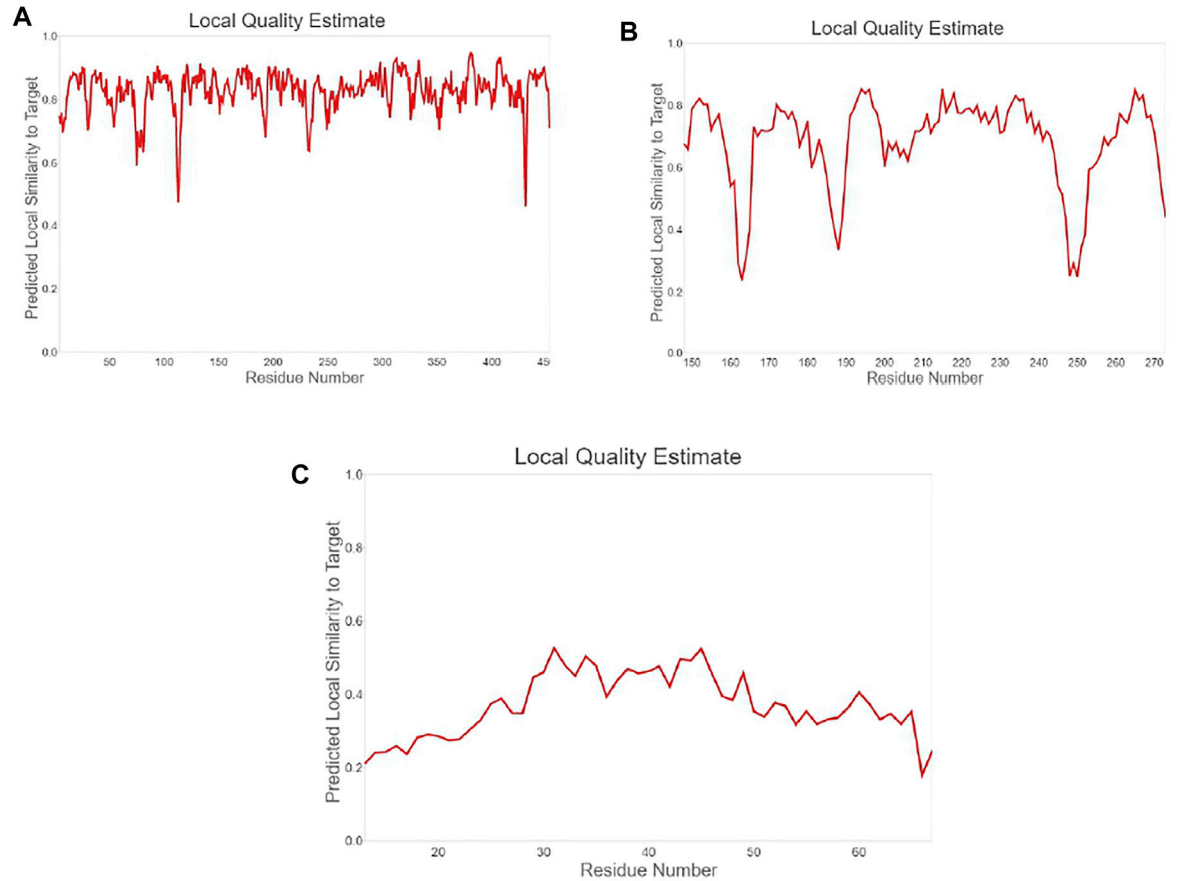

FIGURE 4 | Plot showing local quality estimation of **(A)** methyltransferase, **(B)** *E. coli* membrane protein, and **(C)** *Staphylococcus* membrane protein.

for the SWISS MODEL server on the basis of homology approaches (detailed information available in Supplementary Materials).

Model Evaluation

All modeled 3D structures were evaluated using the MolProbity version 4.4 assessment tool integrated in the SWISS-MODEL server (Chen et al., 2010). All-atom structure validation for macromolecular crystallography was carried out.

Molecular Docking

The molecular interaction analysis were executed using the PatchDock online server (https://bioinfo3d.cs.tau.ac.il/PatchDock/) (Duhovny et al., 2002; Schneidman-Duhovny et al., 2005). PatchDock uses a geometry-based molecular docking algorithm as a scoring function. All figures were generated using Discovery Studio Visualizer 2020 (Ventola, 2015; Dassault Systèmes, 2020).

MDS Experimentation

The docking results *E. coli*_mem, *E. coli*_mem + COL complexes, and *E. coli*_mem + MT of complexes were further analyzed by MDS studies using advanced computational techniques. Thus, the MDS environment was created, and simulation study was conducted for 50 nanoseconds (ns) using the GROningen MAchine for Chemical Simulations (GROMACS) tool (2018 version) (Van Der Spoel et al., 2005) developed by the University of Groningen, Netherlands. The simulation in water for complexes *E. coli*_mem, *E. coli*_mem + MT, and *E. coli*_mem + COL was performed by using GROMACS standard protocol.

The simulation for selected complexes, initially, the pdb2gmx module, was utilized and required *E. coli*_mem, *E. coli*_mem + MT, and *E. coli*_mem + COL topology files to be generated, followed by OPLS-AA/L all-atom force field selection.

The solvation step was performed by creating a unit water willed cell cubic box. The energy was minimized by addition of Na+ and Cl- ions for stabilization of the system. Equilibrium setup for the (all complexes) system was essential and created, followed by two-step ensembles NVT and NPT (constant N, number of particles; V, volume; P, pressure; T, temperature) providing constancy and stabilization of the system through complete simulation (Gupta et al., 2020).

GORMACS have many packages, for *E. coli*_mem, *E. coli*_mem + MT, and *E. coli*_mem + COL complexes. MDS analysis, root mean square deviation (RMSD) was

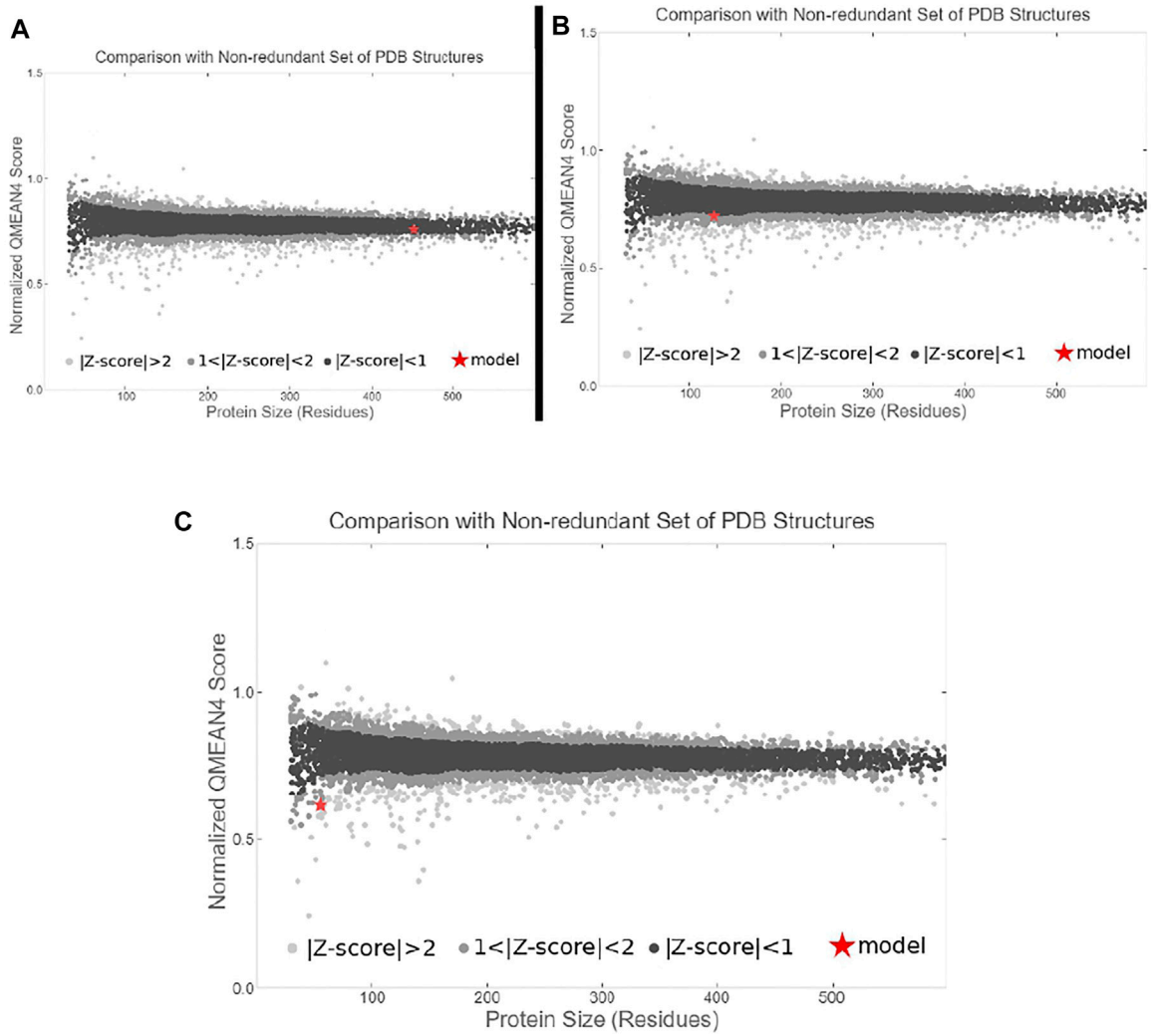

FIGURE 5 | Plot showing local quality comparison with non-redundant set of PDB structure **(A)** methyltransferase, **(B)** *E. coli* membrane protein, and **(C)** *Staphylococcus* membrane protein.

analyzed by gmx rms (Kufareva and Abagyan, 2012), root mean square fluctuation (RMSF) was analyzed by gmx rmsf for, gmx gyrate for the calculation of radius of gyration (Rg) (Kuzmanic and Zagrovic, 2010), and gmx. Finally, after a successful 50-ns simulation run, trajectory files and graphical plots were generated by using the xmgrace program (Turner, 2005).

RESULTS

To study the binding interaction of methyltransferase with the cellular protein of gram-positive and gram-negative bacterial pathogens, the ligand and receptor structure was created using SWISS-MODEL (**Figures 1A–C**, **2A–C**). Similarly, the ligand 3D structure of membrane protein (*Escherichia coli*) and (accession no.: APJ97041.1) and membrane protein (*Staphylococcus aureus*) (accession no.: KII21430.1) was modeled after retrieving their sequences in the FASTA format from the NCBI and provided as an input for the SWISS-MODEL server on the basis of homology approaches. All modeled 3D structures were evaluated using the MolProbity version 4.4 assessment tool integrated into the SWISS-MODEL server. The model structure information is represented in **Table 1**, and the local quality of models is represented in **Figures 3**, **4**.

The quality comparison with a non-redundant set of PDB structures is also performed, which is shown in **Figure 5**. The stability of modeled ligand molecules and receptors was confirmed by the Ramachandran plot (**Figure 6**), which shows that the modeled 3D structure of membrane protein (*Staphylococcus aureus*) was the best-predicted structure that had 98.11% amino acid residues in the favored region with no C-Beta deviation (**Table 1** and **Figure 7A–C**).

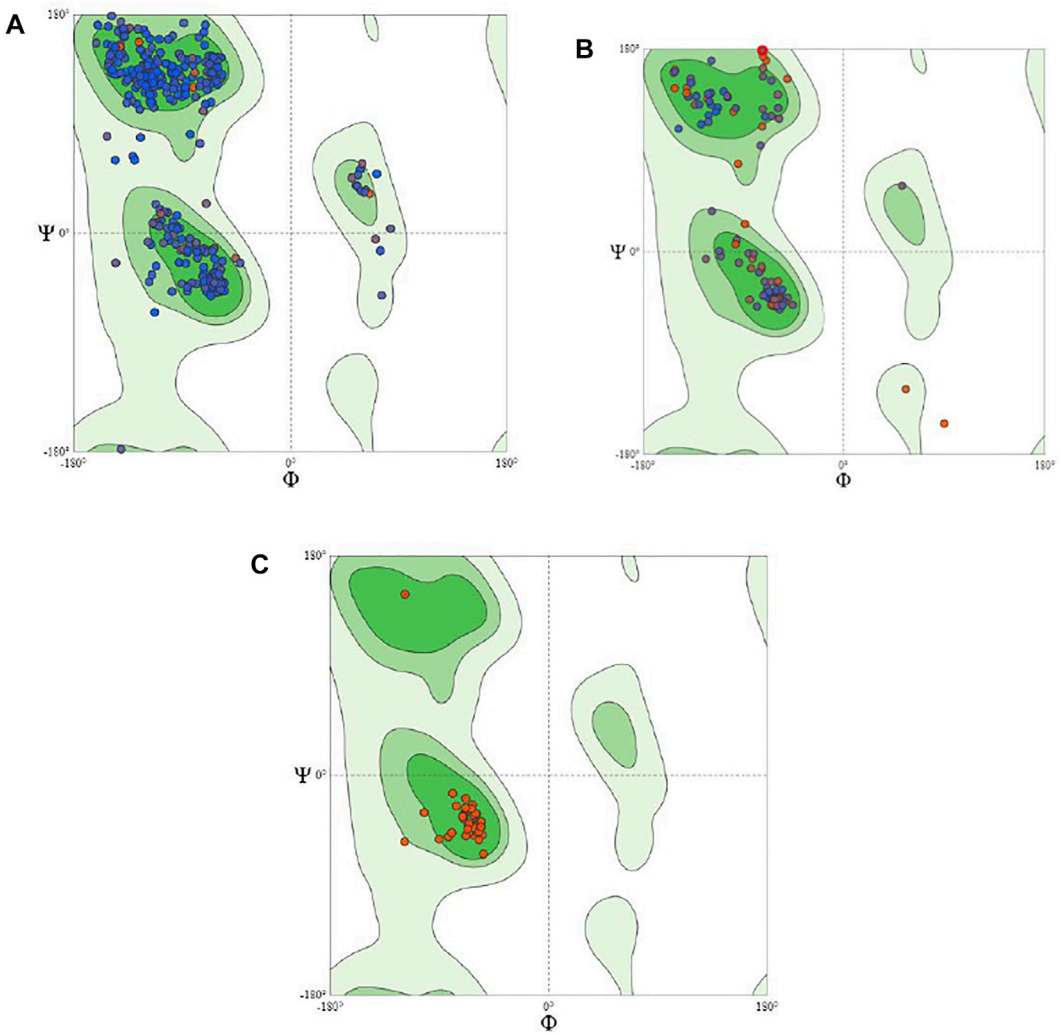

FIGURE 6 | Showing Ramachandran plot for modeled 3D structures of structures **(A)** methyltransferase, **(B)** *E. coli* membrane protein, and **(C)** *Staphylococcus* membrane protein.

Furthermore, the molecular interaction analysis was executed using the PatchDock online server (https://bioinfo3d.cs.tau.ac.il/PatchDock/) (Duhovny et al., 2002; Schneidman-Duhovny et al., 2005; Ansari et al., 2020) (**Tables 2** and **3**).

E. coli membrane protein interaction with methyl transferase with colicin protein and amoxicillin were analyzed and 3D graphics generated using Discovery Studio Visualizer 2020 (**Figure 8**).

The methyltransferase interaction with cellular protein was found to be maximum, due to the maximum PatchDock Score (15808), which was followed by colicin (12864) and amoxicillin (4122) (**Table 2**; **Figures 7A–C**). Moreover, the interaction bond was stabilized through the hydrogen bond between methyltransferase, colicin, amoxicillin, and cellular protein; EC: LYS163:N—MT:GLY153:O, EC:ARG150:NH1—COL:ASP24:OD2, and AMX:O—EC:TYR207. Similarly, interaction study with methytransferase, colicin, and amoxicillin with cellular protein of bacterial pathogen *S. aureus* were found to the maximum with methyltransferase (PatchDock Score: 14024), followed by colicin (PatchDock Score 12790) and amoxicillin (**Table 3**; **Figures 8A–C**).

MDS Analysis

Furthermore, after the MDS total experimentation 50 ns run, the analysis was done on the basis of obtained data from RMSD, RMSF, and Rg plot analysis, revealing deviation, fluctuation, and stability of *E. coli*_mem, *E. coli*_mem + MT, and *E. coli*_mem + COL complexes during the whole simulation period. The RMSD values for selected simulated molecules ranged between 0.15 and 0.4 nm (**Figure 7A**). The observed RMSD values for *E. coli* _mem, *E. coli*_ mem + MT, and *E. coli*_mem + COL complexes were between 0.2 and 0.25 nm, 0.3–0.4 nm, and 0.2–0.25 nm, respectively. In comparison with *E. coli*_mem, the *E. coli*_mem + MT, and *E. coli*_mem +

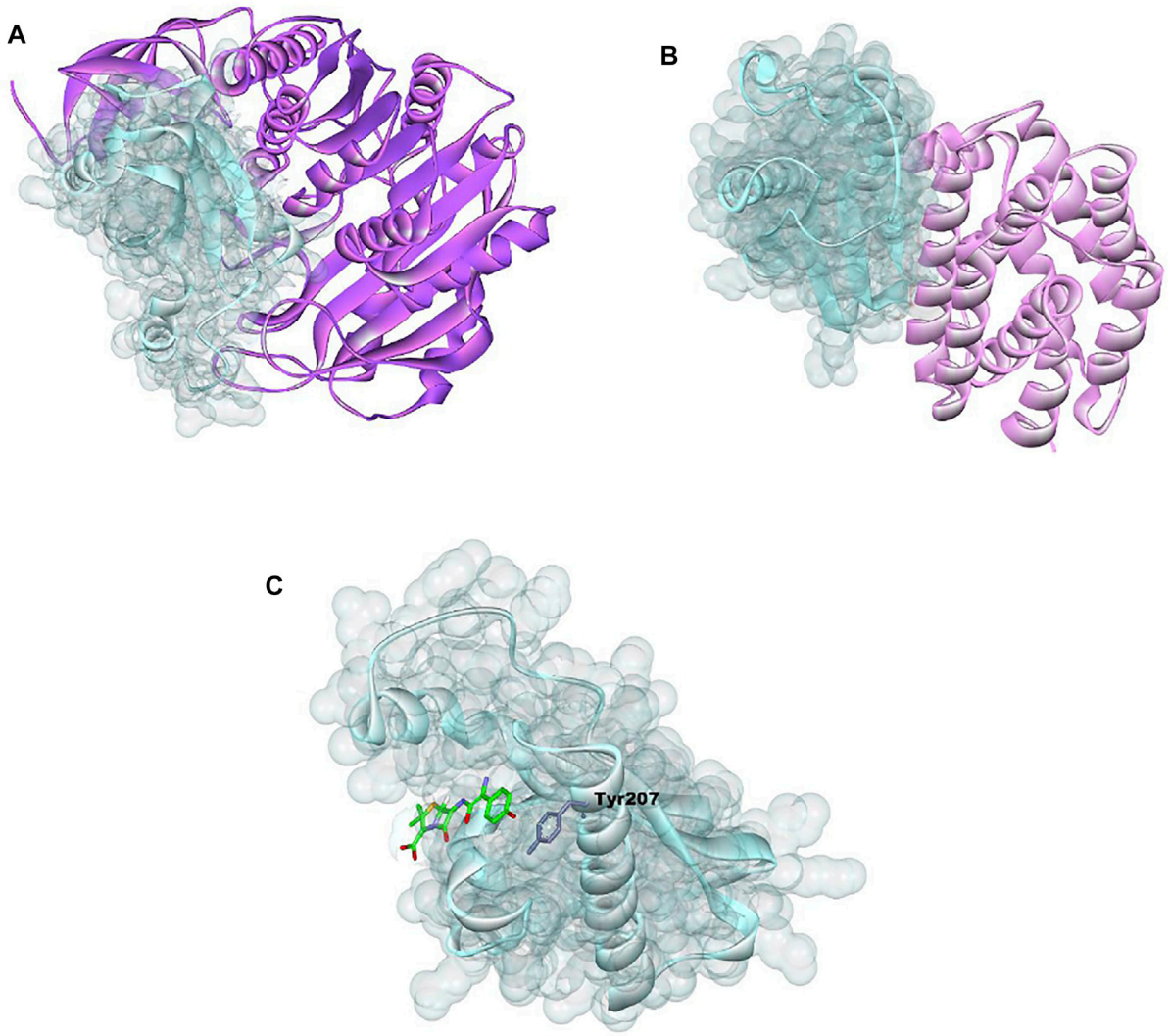

FIGURE 7 | Showing 3D visualization of *E. coli* membrane protein (shown in light turquoise color with solid ribbon pattern) interaction with **(A)** methyltransferase (purple color in ribbon pattern), **(B)** colicin protein (light pink color in ribbon pattern), and **(C)** amoxicillin (green color in stick pattern).

COL complexes showed stability after 20 ns until 50 ns (**Figure 7A**). The *E. coli*_mem + MT complex average RMSD value was higher than those of the *E. coli*_mem and *E. coli*_mem + COL complexes.

RMSF calculation per atom showed a value that ranged between 0.1 and 0.7 nm for protease, *E. coli*_mem, *E. coli*_mem + MT, and *E coli*_mem + COL complexes, and it was observed that for most of the residues, the RMSF value remains near 0.1 nm for all complexes (**Figure 9A–C**). Furthermore, few fluctuations were observed at the 2000- and 3000-atom regions. (**Figure 9B**).

Radius of gyration (Rg) analysis is very important for the assessment of the compactness and stability of the protein structure during the whole simulation period. The *E. coli*_mem Rg plot shows an average value of approximately 1.5 nm. *E. coli*_mem + MT and *E. coli*_mem + COL remain stable and show average values near 2.0 and 2.25 nm, respectively. No major fluctuation was observed in Rg plot analysis (**Figure 9C**).

DISCUSSION

Currently, a formulation of glucose oxidase with other active ingredients was patented. The formulation has been described to has the inhibitory potential against foodborne bacteria *Staphylococcus* cells and biofilm formation of *P. aeruginosa, S. aureus*, and methicillin-resistant *S. aureus*. Honey is rich in nutrient value, glucose oxidase which inhibits *P. aeruginosa*.

TABLE 2 | *E. coli* membrane protein interaction with methyltransferase, colicin protein, and amoxicillin. In the hydrogen bond column, EC, *E. coli* membrane protein; MT, methyltransferase; COL, colicin protein; AMX, amoxicillin.

Serial number	Ligand molecule	PatchDock score	Hydrogen bonds	Hydrogen bonds length (Angstrom)
1	Methyltransferase	15808	EC:LYS163:N—MT:GLY153:O	2.91143
			EC:LYS240:NZ—MT:THR251:OG1	3.15602
			MT:SER111:OG—EC:GLU217:OE1	2.58874
			MT:ASN252:N—EC:ARG201:O	2.39953
			MT:ARG387:NH2—EC:ASP206:O	1.93498
			MT:ARG387:NH2—EC:THR210:OG1	2.16363
			EC:LYS163:CA—MT:ARG152:O	3.48398
			MT:ARG418:CD—EC:SER235:O	2.69943
			MT:HIS445:CE1—EC:THR210:OG1	3.19989
2	Colicin protein	12864	EC:ARG150:NH1—COL:ASP24:OD2	2.88023
			EC:THR152:OG1—COL:GLU17:OE2	3.36734
			EC:LYS245:NZ—COL:LYS6:O	2.9338
			EC:LYS246:NZ—COL:ASN47:OD1	3.30611
			COL:LYS6:NZ—EC:SER255:O	3.01983
			EC:ARG150:CD—COL:GLU17:O	2.74897
			COL:LYS97:CE—EC:HIS179:O	2.22174
			COL:GLY166:CA—EC:GLY274:OXT	3.08932
3	Amoxicillin	4122	AMX:O—EC:TYR207	3.93411

TABLE 3 | *Staphylococcus* membrane protein interaction with methyltransferase, colicin protein, and amoxicillin. In the hydrogen bond column, STP, *Staphylococcus* membrane protein; MT, methyltransferase; COL, colicin protein; AMX, amoxicillin.

Serial number	Ligand molecule	PatchDock score	Hydrogen bonds	Hydrogen bonds length (Angstrom)
1	Methyltransferase	14024	STP:LYS36:NZ—MT:GLN286:OE1	3.12548
			STP:GLN42:NE2—MT:SER154:O	3.36677
			STP:LYS45:NZ—MT:LYS151:O	3.28368
			STP:LYS45:NZ—MT:ARG152:O	2.83486
			MT:GLN135:NE2—STP:MET34:SD	3.7311
			MT:SER154:N—STP:GLN42:OE1	2.89249
			MT:ASN252:ND2—STP:LYS45:O	2.9043
			STP:HIS37:CE1—MT:GLN166:OE1	3.75744
			MT:HIS445:CE1—STP:VAL35:O	2.55974
2	Colicin protein	12790	COL:PRO132:CD—STP:ASP26:OD1	2.24668
3	Amoxicillin	3544	STP:ARG58:NH1—AMX:O	1.67897
			AMX:O—STP:LEU51:O	3.26094
			AMX:N—STP:PHE55	3.93687

Moreover, breeding of novel honeybee species that have produced more glucose oxidases in order to increase the antibacterial efficacy of the product (Bucekova et al., 2014).

The enzymes like endopeptidases themselves are required for the normal growth of bacteria. Moreover, it also destroyed the bacterial cell well in presence of beta-lactam antibiotics like penicillin (Shin et al., 2016). The antimicrobial potential of many ribosomally synthesized proteins named bacteriocin has been reported to kill or inhibit the growth of gram-positive and gram-negative bacteria (Ahmad et al., 2014; Ahmad et al., 2019). In this study, the protein–protein interaction study was conducted between the methyltransferase and cellular membrane protein of gram-positive bacteria and gram-negative bacteria. The antibiotics amoxicillin and peptide antibiotic colicin, a well-known antimicrobial peptide were studied as a positive control. Colicin is the first potential antimicrobial peptide produced from *E. coli* bacteria that has a bactericidal effect by forming pores in the inner membrane of nonhost *E. coli* and damaging the DNA and RNA (Cascales et al., 2007; Jin et al., 2018). The enterococin A, the protein–protein interactions of methyltransferase with cell protein of *E. coli* was observed to be more significant than that of the colicin and amoxicillin as it has been observed to be the maximum PatchDock score (15808), which is followed by

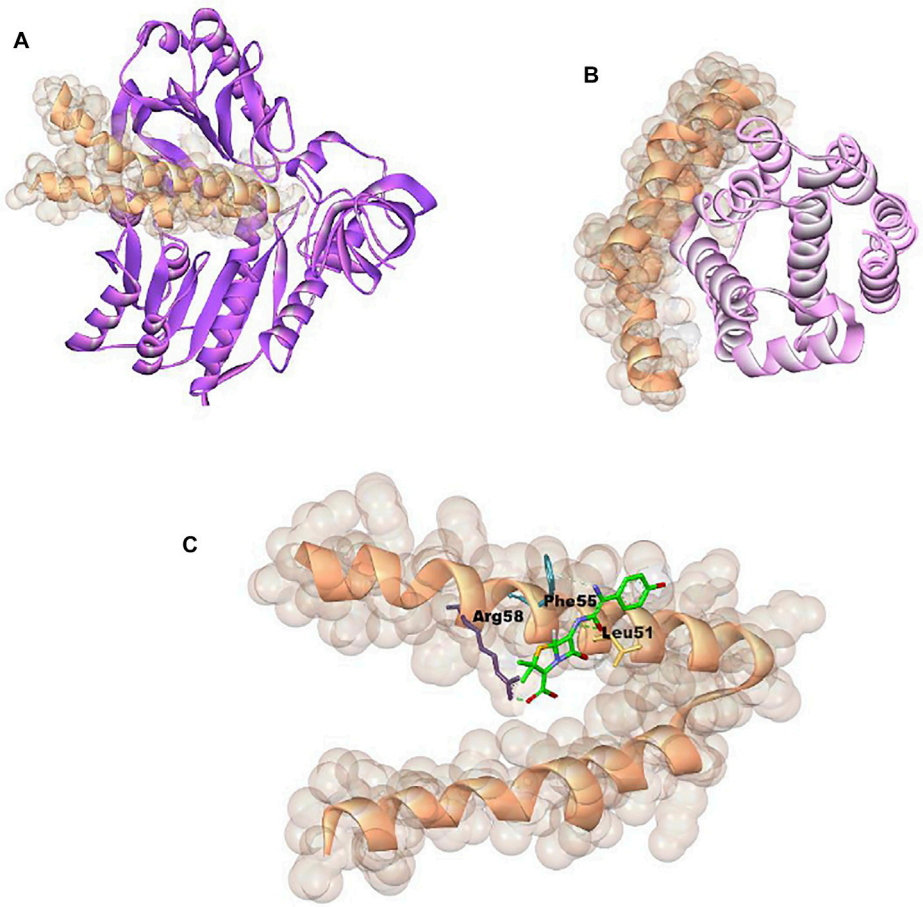

FIGURE 8 | Showing 3D visualization of *Staphylococcus aureus* membrane protein (shown in golden color with solid ribbon pattern) interaction with **(A)** methyltransferase (purple color in ribbon pattern), **(B)** colicin protein (light pink color in ribbon pattern), and **(C)** amoxicillin (green color in stick pattern).

colicin (12864) and amoxicillin (4122). The interaction of cellular protein was also significant for colicin as compared to the amoxicillin (**Figure 6**).

The endogenous alkaline phosphatase (IAP) enzyme usually localizes to the apical brush border and participates in the dephosphorylation of bacterial LPS in addition to the unmethylated cytosine-guanosine dinucleotides and flagellin, leading to reduced bacterial toxicity and inflammation responses (Vaishnava and Hooper, 2007).

In animals, the endogenous levels of IAP are reported to decrease at the weaning stage; hence pathogenic gram-negative bacteria (through an LPS-mediated mechanism) can upregulate inflammatory responses, leading to a symptomatic diarrhea. To address this issue, the use of exogenous IAP overexpression systems to modulate the animal's overall IAP levels, promote gut health, and reduce the associated diarrhea has been suggested.

In a recent study, the effect of intestinal alkaline phosphatase (IAP) and sodium butyrate on LPS-induced intestinal inflammation was evaluated in pigs. The exogenous IAP was able to complement endogenous IAP levels and downregulate LPS-induced inflammatory responses *via* the RelA/p65 (NF-κB) route, demonstrating that such a treatment may indeed be beneficial in attenuating LPS-induced intestinal inflammation.

Colicin is a well-reported antimicrobial peptide that showed strong inhibition as compared to the repurposing antibiotics (Cascales et al., 2007; Jin et al., 2018). The study sequence of methyltransferase has shown a greater inhibitory potential. Recently, Ahmad et al. reported an antimicrobial peptide of 51 kDa from *Lysinibacillus*, with close sequence similarity to the methyltransferase (Ahmad et al., 2019). The interactions of methyltransferase, colicin, and amoxicillin were also studied with membrane protein of a gram-positive bacterial indicator, *Staphylococcus aureus* (**Table 3** and **Figure 7**). The methyltransferase has also been observed to interact more significantly than colicin and amoxicillin. The interaction of colicin with *S. aureus* membrane protein was also observed to

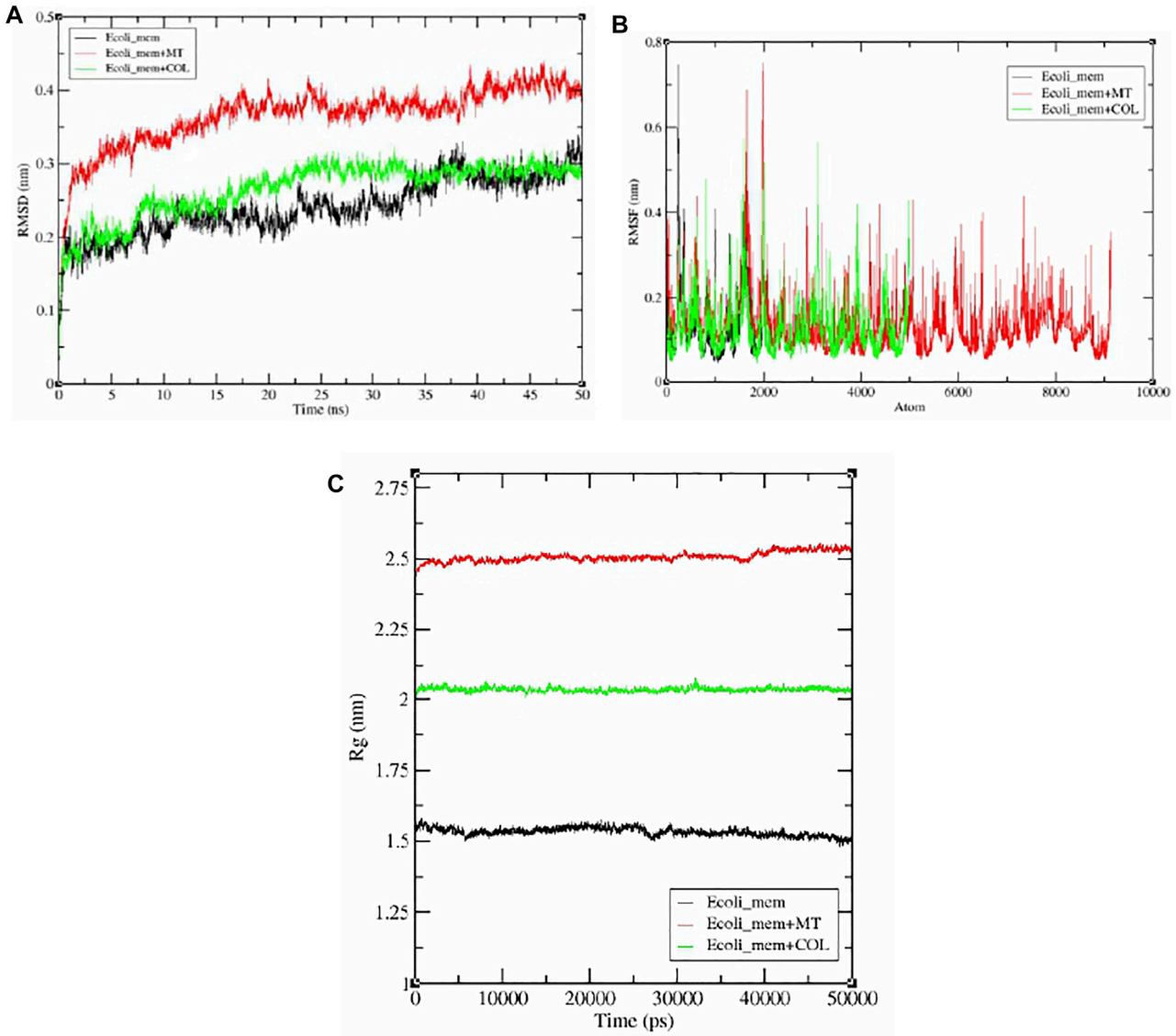

FIGURE 9 | Graphical representation **(A)** RMSD plot of *E. coli*_mem (black color), *E. coli*_mem + MT (red color), and *E. coli*_mem + COL (green color) and showing deviation and stability during 100-ns period. **(B)** RMSF plot with fluctuation per residues. **(C)** Radius of gyration (Rg) plot showing compactness of *E. coli*_mem, *E. coli*_mem + MT, and *E. coli*_mem + COL molecule during 50000ps simulation. Where nm, nanometer; ns, nanosecond; ps, picosecond.

be significant, but it was less than methyltransferase. The antibiotic amoxicillin has also been observed significantly, but it was less than the interaction of methyltransferase. ent A-col E1, an antimicrobial peptide, was recently reported against *S. aureus* (Simons et al., 2020; Fathizadeh et al., 2020).

CONCLUSION

Enzymes play a significant role for the expression of cellular proteins, cell wall polysaccharides, nucleic acids, and other cellular metabolites that are required for the survival of the cell. The use of enzymes as bacteriophage holins and their membrane-disrupting activity, anti-staphylococcal lytic enzymes, and membrane-targeted enzybiotics has recently been highlighted by much research. This *in silico* based study also explores the use of methyltransferase against gram-negative and gram-positive bacteria. The active sequences of this enzyme need to be explored. In this regard, we recommended the designing of short peptides using the methyltransferase sequence and evaluation of antimicrobial potential of these peptides that could be beneficial to develop a peptide-based enzymobiotic against gram-positive and gram-negative bacteria and pathogenic bacteria.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization, VA and AA; methodology, VA and GY; data analysis, VA; writing—original draft preparation, VA and AA; writing—review and editing, MA and SA-T.

FUNDING

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia under Grant No. KEP-26-130-42.

ACKNOWLEDGMENTS

The authors extend their appreciation to Dr Qazi Mohammad Sajid Jamal, Department of Health Informatics, College of Public Health and Health Informatics, Qassim University, Al Bukayriyah, Saudi Arabia for supporting computational analysis.

REFERENCES

Ahmad, V., Ahmad, K., Baig, M. H., Al-Shwaiman, H. A., Al Khulaifi, M. M., Elgorban, A. M., et al. (2019). Efficacy of a Novel Bacteriocin Isolated from Lysinibacillus Sp. Against Bacillus Pumilus. *LWT* 102, 260–267. doi:10.1016/j.lwt.2018.12.021

Ahmad, V., Muhammad Zafar Iqbal, A. N., Haseeb, M., and Khan, M. S. (2014). Antimicrobial Potential of Bacteriocin Producing Lysinibacillus Jx416856 against Foodborne Bacterial and Fungal Pathogens, Isolated from Fruits and Vegetable Waste. *Anaerobe* 27, 87–95. doi:10.1016/j.anaerobe.2014.04.001

Ansari, M. A., Jamal, Q. M. S., Rehman, S., Almatroudi, A., Alzohairy, M. A., Alomary, M. N., et al. (2020). TAT-peptide Conjugated Repurposing Drug against SARS-CoV-2 Main Protease (3CLpro): Potential Therapeutic Intervention to Combat COVID-19. *Arabian J. Chem.* 13 (11), 8069–8079. doi:10.1016/j.arabjc.2020.09.037

Bucekova, M., Valachova, I., Kohutova, L., Prochazka, E., Klaudiny, J., and Majtan, J. (2014). Honeybee Glucose Oxidase-Its Expression in Honeybee Workers and Comparative Analyses of its Content and H2O2-Mediated Antibacterial Activity in Natural Honeys. *Naturwissenschaften* 101, 661–670. doi:10.1007/s00114-014-1205-z

Cascales, E., Buchanan, S. K., Duche´, D., Kleanthous, C., Lloubès, R., Postle, K., et al. (2007). Colicin Biology. *Microbiol. Mol. Biol. Rev.* 71 (1), 158–229. doi:10.1128/MMBR.00036-06

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., et al. (2010). MolProbity: All-Atom Structure Validation for Macromolecular Crystallography. *Acta Cryst.* D66, 12–21. doi:10.1107/S0907444909042073

ChukwukaOkonko, I. O., and Adekunle, A. A. (2010). Microbial Ecology of Organisms Causing Pawpaw (*Carica papaya* L.). Fruit Decay In Oyo State, Nigeria. *Amer.-Eur. J. Toxicol. Sci.* 2, 43–50.

Cooper, R. A. (2013). Inhibition of Biofilms by Glucose Oxidase, Lactoperoxidase and Guaiacol: the Active Antibacterial Component in an Enzyme Alginogel. *Int. Wound J.* 10, 630–637. doi:10.1111/iwj.12083

Dassault Systèmes (2020). *BIOVIA Discovery Studio Visualizer*. [version 2020]. San Diego: Dassault Systèmes.

Drago-Serrano, M. E., de la Garza-Amaya, M., Luna, J. S., and Campos-Rodríguez, R. (2012). Lactoferrin-lipopolysaccharide (LPS) Binding as Key to Antibacterial and Antiendotoxic Effects. *Int. Immunopharmacol.* 12, 1–9. doi:10.1016/j.intimp.2011.11.002

Duhovny, D., Nussinov, R., and Wolfson, H. J. (2002). "Efficient Unbound Docking of Rigid Molecules," in Proceedings of the 2'nd Workshop on Algorithms in Bioinformatics(WABI), Rome, Italy, September 17–21, 2002. Editors R. Guigó and D. Gusfield (Springer-Verlag), 185–200. doi:10.1007/3-540-45784-4_14

Fathizadeh, H., Saffari, M., Esmaeili, D., Moniri, R., and Salimian, M. (2020). Evaluation of Antibacterial Activity of Enterocin A-Colicin E1 Fusion Peptide. *Iran J. Basic Med. Sci.* 23 (11), 1471–1479. doi:10.22038/ijbms.2020.47826.11004

Gupta, S., Tiwari, N., Verma, J., Waseem, M., Subbarao, N., and Munde, M. (2020). Estimation of a Stronger Heparin Binding Locus in Fibronectin Domain III14using Thermodynamics and Molecular Dynamics. *RSC Adv.* 10 (34), 20288–20301. doi:10.1039/D0RA01773F

Jin, X., Kightlinger, W., Kwon, Y.-C., and Hong, S. H. (2018). Rapid Production and Characterization of Antimicrobial Colicins Using Escherichia Coli-Based Cell-free Protein Synthesis. *Synth. Biol.* 3 (1), ysy004. doi:10.1093/synbio/ysy004

Kikot, G. E., Hours, R. A., and Alconada, T. M. (2009). Contribution of Cell wall Degrading Enzymes to Pathogenesis of Fusarium Graminearum : a Review. *J. Basic Microbiol.* 49, 231–241. doi:10.1002/jobm.200800231

Kufareva, I., and Abagyan, R. (2012). Methods of Protein Structure Comparison. *Methods Mol. Biol.* 857, 231–257. doi:10.1007/978-1-61779-588-6_10

Kumar, M., Sarma, D. K., Shubham, S., Kumawat, M., Verma, V., Nina, P. B., et al. (2021). Futuristic Non-antibiotic Therapies to Combat Antibiotic Resistance: A Review. *Front. Microbiol.* 12, 609459. doi:10.3389/fmicb.2021.609459

Kuzmanic, A., and Zagrovic, B. (2010). Determination of Ensemble-Average Pairwise Root Mean-Square Deviation from Experimental B-Factors. *Biophys. J.* 98 (5), 861–871. doi:10.1016/j.bpj.2009.11.011

Sartelli, M., Chichom-Mefire, A., Labricciosa, F. M., Abu-Zidan, F. M., Adesunkanmi, A. K., Ansaloni, L., et al. (2017). Erratum to: The Management of Intra-abdominal Infections from a Global Perspective: 2017 WSES Guidelines for Management of Intra-abdominal Infections. *World J. Emerg. Surg.* 12, 36. doi:10.1186/s13017-017-0148-z

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: Servers for Rigid and Symmetric Docking. *Nucleic Acids Res.* 33, W363–W367. doi:10.1093/nar/gki481

Shin, J. M., Gwak, J. W., Kamarajan, P., Fenno, J. C., Rickard, A. H., and Kapila, Y. L. (2016). Biomedical Applications of Nisin. *J. Appl. Microbiol.* 120, 1449–1465. doi:10.1111/jam.13033

Simons, A., Alhanout, K., and Duval, R. E. (2020). Bacteriocins, Antimicrobial Peptides from Bacterial Origin: Overview of Their Biology and Their Impact against Multidrug-Resistant Bacteria. *Microorganisms* 8, 639. doi:10.3390/microorganisms8050639

Turner, P. (2005). *XMGRACE, Version 5.1. 19*. Beaverton: Center for Coastal and LandMargin Research, Oregon Graduate Institute of Science and Technology.

Vaishnava, S., and Hooper, L. V. (2007). Alkaline Phosphatase: Keeping the Peace at the Gut Epithelial Surface. *Cell Host Microbe* 2, 365–367. doi:10.1016/j.chom.2007.11.004

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* 26 (16), 1701–1718. doi:10.1002/jcc.20291

Ventola, C. L. (2015). The Antibiotic Resistance Crisis: Part 1: Causes and Threats. *P T* 40 (4), 277–283.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* 46, W296–W303. doi:10.1093/nar/gky427

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* 36, D901–D906. doi:10.1093/nar/gkm958

Zuo, J., Ling, B., Long, L., Li, T., Lahaye, L., Yang, C., et al. (2015). Effect of Dietary Supplementation with Protease on Growth Performance, Nutrient

Digestibility, Intestinal Morphology, Digestive Enzymes and Gene Expression of Weaned Piglets. *Anim. Nutr.* 1, 276–282. doi:10.1016/j.aninu.2015.10.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership