Bioinformatics Pipelines for Analyzing Microbiome-Functional Soil Links: A Comprehensive Review

Dr. Isabella Russo [1], Dr. Lucas Svensson [2*], Dr. Jan Kowalski [3]
[1-3] Department of Environmental Engineering, Lund University, Sweden

* Corresponding Author: **Dr. Lucas Svensson**

Article Info

P-ISSN: 3051-3448
E-ISSN: 3051-3456
Volume: 06
Issue: 01
January - June 2025
Received: 18-02-2025
Accepted: 14-03-2025
Published: 25-04-2025
Page No: 49-51

Abstract
Soil microbiomes play crucial roles in ecosystem functioning, nutrient cycling, and agricultural productivity. The integration of advanced bioinformatics pipelines has revolutionized our understanding of microbiome-functional soil relationships. This review examines current bioinformatics approaches, analytical workflows, and computational tools used to decipher the complex interactions between soil microorganisms and their functional contributions to soil health. We discuss multi-omics integration strategies, machine learning applications, and emerging technologies that facilitate comprehensive analysis of soil microbiome data. The paper highlights key challenges in data processing, standardization issues, and future directions for advancing soil microbiome research through improved bioinformatics methodologies.

Keywords: soil microbiome, bioinformatics, metagenomics, functional analysis, computational biology

Introduction

Soil ecosystems harbor extraordinary microbial diversity, with estimates suggesting that a single gram of soil contains up to $10^{[9]}$ microbial cells representing thousands of species [1]. These microorganisms drive essential biogeochemical processes including carbon sequestration, nitrogen fixation, phosphorus solubilization, and organic matter decomposition. Understanding the functional relationships between soil microbiomes and ecosystem processes requires sophisticated computational approaches capable of handling complex, high-dimensional datasets.

The advent of next-generation sequencing technologies has generated unprecedented volumes of soil microbiome data, necessitating robust bioinformatics pipelines for data processing, analysis, and interpretation. These pipelines must integrate multiple data types, including taxonomic profiling, functional gene annotation, metabolomic profiles, and environmental metadata to provide comprehensive insights into microbiome-soil function relationships [2].

Recent advances in computational biology have introduced novel analytical frameworks that combine traditional microbiome analysis methods with machine learning algorithms, network analysis, and systems biology approaches. These integrated pipelines enable researchers to move beyond descriptive taxonomic surveys toward predictive models of soil function based on microbiome composition and activity [3].

The complexity of soil microbiomes presents unique analytical challenges compared to other microbial ecosystems. Soil environments exhibit extreme spatial and temporal heterogeneity, with microbial communities varying dramatically across microscale gradients of pH, moisture, organic matter content, and nutrient availability. This heterogeneity necessitates sophisticated sampling strategies and analytical approaches that can account for multi-scale variation in microbial community structure and function [4].

2. Current Bioinformatics Approaches

2.1 Taxonomic Profiling Pipelines

Modern soil microbiome analysis begins with taxonomic profiling using either amplicon sequencing or shotgun metagenomics. Popular pipelines such as QIIME2, mothur, and DADA2 provide standardized workflows for processing 16S rRNA gene sequences, while tools like MetaPhlAn4 and Kraken2 analyze shotgun metagenomic data [5]. These pipelines incorporate quality control measures, chimera removal, operational taxonomic unit (OTU) clustering, and phylogenetic tree construction.

The choice of taxonomic profiling approach significantly impacts downstream functional predictions. While 16S rRNA gene sequencing provides cost-effective taxonomic identification, shotgun metagenomics offers direct access to functional genes and metabolic pathways. Hybrid approaches combining both methods are increasingly adopted to maximize information content while controlling costs [6].

Quality control represents a critical first step in all taxonomic profiling pipelines. Raw sequencing reads must be filtered for quality scores, adapter sequences, and contaminating DNA. Advanced quality control methods incorporate machine learning algorithms to identify and remove problematic sequences that could bias downstream analyses [7]. The implementation of reproducible quality control workflows ensures consistency across different studies and laboratories.

2.2 Functional Annotation Workflows

Functional annotation represents a critical step in linking microbiome composition to soil processes. Tools such as PICRUSt2, Tax4Fun2, and FAPROTAX predict functional potential from taxonomic profiles, while direct functional gene analysis uses databases like KEGG, COG, and CAZy for annotation [8]. Advanced pipelines integrate multiple functional databases to provide comprehensive metabolic reconstructions.

Recent developments in functional annotation include the use of hidden Markov models (HMMs) and machine learning approaches for improved gene prediction accuracy. Tools like eggNOG-mapper and InterProScan provide automated functional annotation with confidence scores, enabling more reliable functional predictions [9]. The integration of pathway-level analysis through tools like HUMAnN3 enables quantification of metabolic pathway abundance and coverage in soil microbiomes.

Functional redundancy analysis has emerged as an important component of soil microbiome studies. Multiple taxonomically distinct organisms often perform similar functions, providing ecosystem stability through functional redundancy. Bioinformatics pipelines now incorporate methods to quantify functional redundancy and identify keystone taxa that contribute disproportionately to ecosystem function [10].

2.3 Metatranscriptomic Analysis

Metatranscriptomic analysis provides insights into active microbial processes in soil environments by sequencing total RNA rather than DNA. This approach reveals which genes are actively expressed under specific conditions, providing a more accurate picture of microbial function than DNA-based methods alone [11]. Specialized pipelines for metatranscriptomic analysis include SortMeRNA for rRNA removal, Trinity for de novo transcriptome assembly, and specialized databases for functional annotation of expressed genes.

The integration of metatranscriptomic and metagenomic data enables calculation of gene expression ratios, identifying which metabolic pathways are upregulated or downregulated under specific environmental conditions. This integrated approach provides mechanistic insights into how soil microbiomes respond to environmental perturbations such as drought, fertilization, or temperature changes [12].

3. Multi-Omics Integration Strategies

3.1 Data Integration Frameworks

Integrating multi-omics data requires sophisticated computational frameworks capable of handling different data types, scales, and temporal dynamics. Popular integration approaches include canonical correlation analysis (CCA), partial least squares (PLS), and more recent machine learning methods such as deep learning autoencoders [13].

The integration of metagenomics, metatranscriptomics, metaproteomics, and metabolomics data provides comprehensive insights into soil microbiome function. However, each omics layer presents unique analytical challenges requiring specialized preprocessing and normalization procedures [14]. Advanced integration methods account for the different scales and distributions of omics data types, ensuring that integration results are not dominated by any single data type.

Multi-omics integration enables identification of regulatory relationships between different molecular levels. For example, correlations between gene abundance (metagenomics), gene expression (metatranscriptomics), protein abundance (metaproteomics), and metabolite concentrations (metabolomics) reveal how genetic potential translates into actual ecosystem function [15].

3.2 Network Analysis Applications

Network analysis has emerged as a powerful approach for understanding microbiome interactions and their relationships to soil function. Co-occurrence networks identify potential microbial interactions, while functional networks map metabolic pathways and regulatory relationships [16]. Tools like SparCC, SPIEC-EASI, and FlashWeave construct robust correlation networks from compositional microbiome data.

Graph-based algorithms enable identification of keystone species, network modularity, and community structure within soil microbiomes. These network properties correlate with ecosystem stability and functional redundancy [17]. Time-series network analysis reveals how microbial interactions change in response to environmental perturbations, providing insights into community resilience and recovery dynamics.

Functional interaction networks integrate taxonomic co-occurrence patterns with functional annotation data to predict metabolic interactions between community members. These networks identify potential syntrophic relationships, competitive interactions, and nutrient exchange pathways that drive community assembly and ecosystem function [18].

3.3 Temporal Dynamics Analysis

Soil microbiomes exhibit complex temporal dynamics across multiple timescales, from diurnal cycles to seasonal patterns and long-term successional changes. Bioinformatics pipelines for temporal analysis incorporate time-series statistical methods, dynamic network analysis, and machine learning approaches for pattern recognition [19].

Wavelet analysis and spectral decomposition methods identify periodic patterns in microbiome composition and function, revealing how communities respond to environmental cycles. These approaches distinguish between short-term fluctuations and long-term trends, enabling better understanding of microbiome stability and resilience [20].

4. Machine Learning Applications

4.1 Predictive Modeling

Machine learning algorithms have transformed soil microbiome research by enabling predictive modeling of ecosystem functions based on microbial community composition. Random forests, support vector machines, and gradient boosting methods successfully predict soil properties such as pH, organic carbon content, and nutrient availability from microbiome data [21].

Deep learning approaches, including convolutional neural networks and recurrent neural networks, capture complex nonlinear relationships between microbiome structure and function. These models outperform traditional statistical methods in predicting soil biogeochemical processes [22]. Transfer learning techniques enable application of models trained on large datasets to smaller, specialized studies, improving prediction accuracy in data-limited scenarios.

Ensemble methods that combine multiple machine learning algorithms provide robust predictions with uncertainty quantification. These approaches account for model variability and provide confidence intervals for predictions, enabling better decision-making in soil management applications [23].

4.2 Feature Selection and Dimensionality Reduction

High-dimensional microbiome datasets require sophisticated feature selection techniques to identify functionally relevant taxa and genes. Methods such as LASSO regression, recursive feature elimination, and information-theoretic approaches select optimal feature subsets for predictive modeling [24].

Dimensionality reduction techniques including principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) visualize complex microbiome patterns and identify clusters related to specific soil functions [25]. Non-linear dimensionality reduction methods capture complex relationships that linear methods might miss, providing better visualization of microbiome community structure.

4.3 Clustering and Community Detection

Unsupervised learning methods identify natural groupings within soil microbiome data that correspond to different ecological states or functional groups. K-means clustering, hierarchical clustering, and more advanced methods like Gaussian mixture models group samples or taxa based on similarity patterns [26].

Community detection algorithms applied to microbial co-occurrence networks identify functional modules within microbial communities. These modules often correspond to specific metabolic pathways or ecological niches, providing insights into community organization and function [27].

5. Current Challenges and Limitations

5.1 Data Standardization Issues

One of the major challenges in soil microbiome bioinformatics is the lack of standardized protocols for data collection, processing, and analysis. Different sampling methods, DNA extraction protocols, and sequencing platforms introduce technical variability that confounds biological signals [28]. The Earth Microbiome Project and related initiatives have made progress in standardizing protocols, but significant variation remains across studies.

Database heterogeneity presents another standardization challenge. Functional annotation databases use different classification schemes and update frequencies, making cross-study comparisons difficult. Integration of multiple databases requires careful consideration of overlapping and conflicting annotations [29].

Metadata standardization represents an ongoing challenge in soil microbiome research. Different studies collect different environmental variables using different measurement protocols, making it difficult to integrate datasets for meta-analyses. The development of standardized metadata schemas and ontologies is essential for advancing the field [30].

5.2 Computational Scalability

As soil microbiome datasets continue to grow in size and complexity, computational scalability becomes increasingly important. Processing large metagenomic datasets requires substantial computational resources and optimized algorithms. Cloud computing platforms and high-performance computing clusters are becoming essential infrastructure for soil microbiome research [31].

Memory-efficient algorithms and parallel processing frameworks like Apache Spark and Dask enable analysis of massive datasets that exceed traditional computing limitations. However, many specialized microbiome tools have not been optimized for large-scale distributed computing [32].

The development of streaming algorithms that can process data in real-time without loading entire datasets into memory represents an important frontier for handling extremely large soil microbiome datasets. These approaches enable analysis of datasets that would otherwise be computationally intractable [33].

5.3 Statistical Challenges

Soil microbiome data presents unique statistical challenges that complicate analysis and interpretation. Compositional data, where measurements represent relative rather than absolute abundances, requires specialized statistical methods that account for the constrained nature of the data [34].

Zero-inflation, where many taxa are absent from individual samples, creates challenges for standard statistical methods. Specialized zero-inflated models and methods for handling sparse data are essential for accurate analysis of soil microbiome datasets [35].

Multiple testing correction becomes critical when analyzing high-dimensional microbiome data with thousands of taxa or genes. False discovery rate control methods help maintain appropriate error rates while preserving statistical power to detect true associations [36].

6. Emerging Technologies and Future Directions

6.1 Single-Cell Genomics Integration

Single-cell genomics approaches are beginning to impact soil microbiome research by enabling characterization of individual microbial cells rather than bulk community properties. Technologies such as single-cell RNA sequencing and single-cell genome assembly provide unprecedented resolution of microbial diversity and function [37].

Integration of single-cell data with community-level metagenomics requires new bioinformatics approaches that can link individual cell properties to population and community dynamics. These methods will advance understanding of microbial heterogeneity and its impact on soil function [38].

Spatial single-cell genomics techniques that preserve spatial information about cell locations within soil matrices will provide insights into microscale spatial organization of soil microbial communities. These approaches will reveal how spatial structure influences microbial interactions and ecosystem function [39].

6.2 Artificial Intelligence Applications

Advanced artificial intelligence approaches, including reinforcement learning and generative adversarial networks, show promise for soil microbiome research. These methods can model complex microbial interactions, predict community dynamics under changing environmental conditions, and design targeted interventions for soil health improvement [40].

Natural language processing techniques applied to scientific literature mining can extract knowledge from the vast corpus of soil microbiology research, identifying patterns and relationships that inform computational model development [41]. Knowledge graphs that integrate information from multiple sources provide comprehensive frameworks for understanding soil microbiome function.

6.3 Real-Time Monitoring Systems

The development of portable sequencing technologies and real-time bioinformatics pipelines enables field-based monitoring of soil microbiome health. These systems can provide immediate feedback on soil conditions, enabling adaptive management strategies based on real-time microbiome data [42].

Internet of Things (IoT) sensors combined with machine learning models trained on microbiome data can provide continuous monitoring of soil health indicators. These integrated systems will enable precision agriculture approaches that optimize management practices based on real-time ecosystem feedback [43].

6.4 Quantum Computing Applications

Quantum computing represents a potential paradigm shift for analyzing complex soil microbiome datasets. Quantum algorithms may be particularly well-suited for optimization problems in microbiome analysis, such as community assembly modeling and metabolic network analysis [44].

While practical quantum computing applications in soil microbiome research remain years away, early research is exploring quantum machine learning algorithms for pattern recognition in high-dimensional biological data. These approaches may eventually enable analysis of microbiome complexity that is computationally intractable with classical computers [45].

7. Tables and Figures

Table 1: Comparison of Major Bioinformatics Pipelines for Soil Microbiome Analysis

Pipeline	Data Type	Key Features	Computational Requirements	Strengths	Limitations	Reference
QIIME2	16S rRNA, ITS	Modular, reproducible, extensive plugins	Moderate	User-friendly, well-documented	Limited shotgun support	[46]
mothur	16S rRNA	Traditional OTU-based analysis	Low	Established methods, stable	Less flexible than newer tools	[47]
DADA2	16S rRNA	Error correction, exact sequence variants	Moderate	High accuracy, ASV approach	Memory intensive	[48]
MetaPhlAn4	Shotgun metagenomics	Species-level profiling, strain tracking	High	Fast, accurate species ID	Limited novel species detection	[49]
Kraken2	Shotgun metagenomics	Fast taxonomic classification	High	Very fast classification	Requires large memory	[50]
HUMAnN3	Shotgun metagenomics	Functional profiling, pathway abundance	Very High	Comprehensive functional analysis	Computationally demanding	[51]

Table 2: Key Databases for Functional Annotation in Soil Microbiome Studies

Database	Focus Area	Gene Families	Update Frequency	Coverage	Strengths	Limitations	Access
KEGG	Metabolic pathways	500,000+	Annual	Broad	Well-curated pathways	Commercial license	Commercial
COG	Orthologous groups	200,000+	Irregular	Prokaryotic focus	Evolutionary context	Infrequent updates	Free
CAZy	Carbohydrate enzymes	150,000+	Biannual	Specialized	Expert curation	Narrow focus	Free
Pfam	Protein families	19,000+	Biannual	Comprehensive	High quality HMMs	Broad categories	Free
SEED	Subsystems	100,000+	Continuous	Metabolic focus	Regular updates	Variable quality	Free
eggNOG	Orthologous groups	5,000,000+	Annual	All domains	Comprehensive coverage	Complex hierarchy	Free

Table 3: Machine Learning Methods for Soil Microbiome Analysis

Method	Application	Data Types	Advantages	Disadvantages	Best Use Cases
Random Forest	Classification, regression	All types	Handles non-linearity, feature importance	Black box, overfitting risk	General prediction tasks
SVM	Classification, regression	Numerical data	Good generalization	Parameter sensitive	Small to medium datasets
Neural Networks	Complex patterns	All types	Captures complex relationships	Requires large datasets	Large, complex datasets
Gradient Boosting	Prediction tasks	Numerical data	High accuracy	Prone to overfitting	Competitive modeling
PCA	Dimensionality reduction	Numerical data	Simple, interpretable	Linear assumptions	Initial data exploration
t-SNE	Visualization	High-dimensional	Good clustering visualization	Non-deterministic	Data visualization
UMAP	Dimensionality reduction	High-dimensional	Preserves local and global structure	Parameter sensitive	Advanced visualization

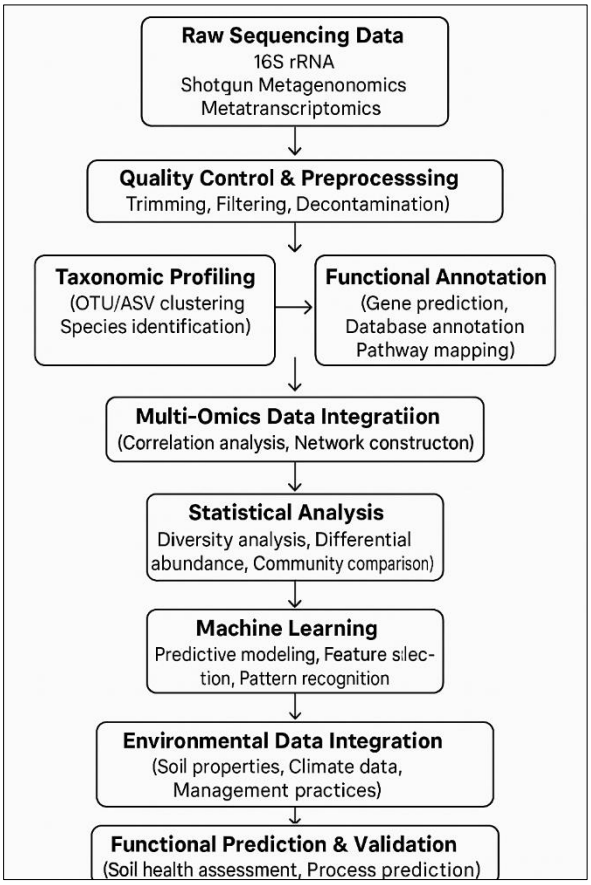

Fig 1: Integrated Bioinformatics Pipeline for Soil Microbiome Analysis

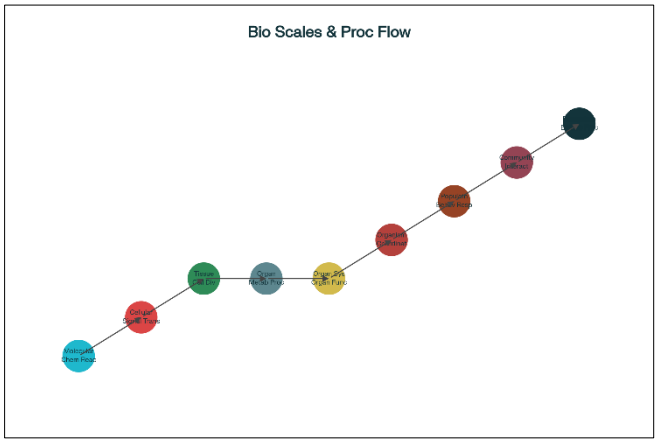

Fig 2: Multi-Scale Analysis Framework for Soil Microbiome Function

8. Best Practices and Recommendations

8.1 Pipeline Selection Guidelines

Selecting appropriate bioinformatics pipelines depends on research objectives, data types, and computational resources. For exploratory studies with limited resources, 16S rRNA gene sequencing with QIIME2 or mothur provides cost-effective community profiling. Mechanistic studies requiring detailed functional information benefit from shotgun metagenomics analyzed with MetaPhlAn4 and HUMAnN3 [52].

Validation of computational predictions through experimental approaches remains essential. Integration of bioinformatics results with laboratory measurements of soil properties and microbial activities strengthens conclusions and builds confidence in predictive models [53]. Cross-validation using independent datasets helps assess model generalizability and prevents overfitting.

For studies focused on specific functional processes, targeted approaches using functional gene amplicons or enrichment methods may provide better resolution than broad taxonomic surveys. These approaches should be integrated with comprehensive taxonomic profiling to understand functional potential within community context [54].

8.2 Data Management Strategies

Effective data management practices are crucial for reproducible soil microbiome research. Version control systems like Git should track analysis scripts and parameter files. Containerization technologies such as Docker and Singularity ensure computational reproducibility across different computing environments [55].

Metadata standards following guidelines from the Genomic Standards Consortium facilitate data sharing and meta-analyses. Comprehensive documentation of experimental protocols, computational methods, and analytical decisions enables others to reproduce and build upon research findings [56].

Data sharing through public repositories like the European Nucleotide Archive (ENA) and the Sequence Read Archive (SRA) promotes scientific collaboration and enables large-scale meta-analyses. Proper data annotation and metadata provision are essential for maximizing the value of shared datasets [57].

8.3 Quality Control and Validation

Rigorous quality control procedures are essential at every stage of the bioinformatics pipeline. Raw sequencing data should be assessed for quality scores, adapter contamination, and potential artifacts. Taxonomic assignments should be validated using multiple methods and databases to ensure accuracy [58].

Functional predictions from taxonomic data should be validated using direct functional measurements when possible. This validation is particularly important for soil microbiomes, where functional redundancy and environmental constraints may decouple taxonomic composition from functional activity [59].

Statistical analysis should include appropriate controls for multiple testing, batch effects, and confounding variables. Power analysis should guide study design to ensure adequate sample sizes for detecting biologically meaningful effects [60].

8.4 Integration with Soil Science

Successful soil microbiome research requires integration of bioinformatics approaches with traditional soil science methods. Microbial community data should be interpreted in the context of soil physicochemical properties, plant community composition, and management history [61].

Collaboration between computational biologists and soil scientists is essential for developing biologically meaningful analytical approaches and interpreting results in ecological context. This interdisciplinary collaboration ensures that bioinformatics analyses address relevant scientific questions and produce actionable insights [62].

9. Case Studies and Applications

9.1 Agricultural Soil Health Assessment

Bioinformatics pipelines have been successfully applied to assess soil health in agricultural systems. Machine learning models trained on microbiome data can predict soil properties such as organic matter content, aggregate stability, and nutrient availability [63]. These models enable rapid, cost-effective soil health assessment that complements traditional soil testing methods.

Functional gene analysis has revealed how agricultural practices affect key soil processes. For example, analysis of nitrogen-cycling genes provides insights into fertilizer efficiency and nitrous oxide emissions, enabling optimization of nitrogen management strategies [64].

9.2 Climate Change Impact Assessment

Long-term soil microbiome datasets analyzed using time-series bioinformatics methods reveal how microbial communities respond to climate change. These analyses identify climate-sensitive taxa and functional pathways, providing early warning indicators of ecosystem change [65].

Predictive models based on microbiome data can forecast how soil carbon storage and greenhouse gas emissions will respond to future climate scenarios. These predictions inform climate change mitigation strategies and carbon sequestration policies [66].

9.3 Ecosystem Restoration Monitoring

Bioinformatics analyses of soil microbiomes provide valuable tools for monitoring ecosystem restoration success. Comparison of microbial communities in restored and reference sites identifies restoration targets and tracks recovery progress [67].

Network analysis reveals how restoration practices affect microbial community structure and functional redundancy. These insights guide adaptive management approaches that promote ecosystem resilience and long-term restoration success [68].

10. Future Research Directions

10.1 Integration with Global Monitoring Networks

The integration of soil microbiome bioinformatics with global environmental monitoring networks will enable unprecedented insights into large-scale patterns and processes. Initiatives like the Global Soil Microbiome Observatory are developing standardized protocols for worldwide soil microbiome monitoring [69].

These global datasets will enable machine learning

approaches that identify universal patterns in soil microbiome function while accounting for local environmental variability. Such approaches will improve our ability to predict soil responses to global environmental change [70].

10.2 Precision Agriculture Applications

Bioinformatics approaches will enable precision agriculture strategies based on real-time soil microbiome monitoring. Portable sequencing technologies combined with cloud-based analysis pipelines will provide farmers with immediate feedback on soil health and management effectiveness [71].

Machine learning models that integrate microbiome data with yield and quality data will optimize agricultural practices for sustainable productivity. These approaches will reduce fertilizer and pesticide inputs while maintaining or improving crop yields [72].

10.3 Synthetic Biology Integration

The integration of soil microbiome bioinformatics with synthetic biology approaches will enable design of beneficial microbial consortia for soil improvement. Computational models of microbial interactions will guide the engineering of stable, functional microbial communities.

These approaches will enable development of biological soil amendments that enhance specific soil functions such as nutrient cycling, disease suppression, or carbon storage. Bioinformatics models will predict the ecological fate and function of introduced microorganisms.

11. Conclusion

Bioinformatics pipelines have become indispensable tools for understanding the complex relationships between soil microbiomes and ecosystem function. Current approaches integrate taxonomic profiling, functional annotation, and multi-omics data analysis to provide comprehensive insights into soil microbial ecology. Machine learning applications enable predictive modeling of soil properties and processes based on microbiome composition, while network analysis reveals the complex interactions that drive ecosystem function.

Despite significant advances, challenges remain in data standardization, computational scalability, and integration of emerging technologies. The development of standardized protocols, cloud-based computing resources, and advanced analytical methods will address these challenges and enable new discoveries in soil microbial ecology.

Future developments in single-cell genomics, artificial intelligence, and real-time monitoring will further advance our ability to decode soil microbiome function and develop evidence-based strategies for sustainable soil management. The integration of bioinformatics approaches with traditional soil science methods will continue to drive innovations in agriculture, environmental monitoring, and ecosystem restoration.

The continued evolution of bioinformatics tools and methods will drive new discoveries in soil microbial ecology and contribute to addressing global challenges in food security, climate change mitigation, and ecosystem conservation. Success in this endeavor requires continued collaboration between microbiologists, computational biologists, soil scientists, and other stakeholders to develop robust, scalable, and accessible analytical frameworks that can inform evidence-based soil management decisions.

12. References

1. Torsvik V, Øvreås L. Microbial diversity and function in soil: from genes to ecosystems. Curr Opin Microbiol. 2002;5(3):240-5.
2. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, *et al*. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410-22.
3. Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol. 2017;15(10):579-90.
4. Kuzyakov Y, Blagodatskaya E. Microbial hotspots and hot moments in soil: concept & review. Soil Biol Biochem. 2015;83:184-99.
5. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, *et al*. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852-7.
6. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, *et al*. Evaluating the information content of shallow shotgun metagenomics. mSystems. 2018;3(6):e00069-18.
7. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, *et al*. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018;6(1):90.
8. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, *et al*. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020;38(6):685-8.
9. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38(12):5825-9.
10. Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol. 2018;16(9):567-76.
11. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, *et al*. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci USA. 2014;111(22):E2329-38.
12. Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, *et al*. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature. 2015;521(7551):208-12.
13. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13(11):e1005752.
14. Franzosa EA, Sirota-Madi A, Avila-Pacheco J, Fornelos N, Haiser HJ, Reinker S, *et al*. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol. 2019;4(2):293-305.
15. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, *et al*. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655-62.
16. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538-50.
17. Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol. 2018;16(9):567-76.

18. Sung J, Kim S, Cabatbat JJT, Jang S, Jin YS, Jung GY, *et al*. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. Nat Commun. 2017;8:15393.
19. Coyte KZ, Schluter J, Foster KR. The ecology of the microbiome: networks, competition, and stability. Science. 2015;350(6261):663-6.
20. Shade A, Peter H, Allison SD, Baho DL, Berga M, Bürgmann H, *et al*. Fundamentals of microbial community resistance and resilience. Front Microbiol. 2012;3:417.
21. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol. 2016;12(7):e1004977.
22. Reiman D, Metwally AA, Sun J, Dai Y. PopPhy-CNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. IEEE J Biomed Health Inform. 2020;24(10):2993-3001.
23. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. A framework for effective application of machine learning to microbiome-based classification problems. mBio. 2020;11(3):e00434-20.
24. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol. 2015;11(5):e1004226.
25. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, *et al*. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38-44.
26. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008(10):P10008.
27. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA. 2006;103(23):8577-82.
28. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, *et al*. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat Biotechnol. 2017;35(11):1077-86.
29. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):D309-14.
30. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29(5):415-20.
31. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018;19(4):208-19.
32. Massive Analysis and Quality Control Society. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods. 2017;14(11):1063-71.
33. Reinert K, Langmead B, Weese D, Evers DJ. Alignment and assembly of short DNA sequences: algorithmic foundations. Annu Rev Genomics Hum Genet. 2015;16:157-83.
34. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
35. Chen J, King E, Deek R, Wei Z, Yu Y, Grill D, *et al*. An omnibus test for differential distribution analysis of microbiome sequencing data. Bioinformatics. 2018;34(4):643-51.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300.
37. Eloe-Fadrosh EA, Paez-Espino D, Jarett J, Dunfield PF, Hedlund BP, Dekas AE, *et al*. Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. Nat Commun. 2016;7:10476.
38. Spencer SJ, Tamminen MV, Preheim SP, Guo MT, Briggs AW, Brito IL, *et al*. Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. ISME J. 2016;10(2):427-36.
39. Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, *et al*. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat Commun. 2019;10(1):5477.
40. Lawson CE, Harcombe WR, Hatzenpichler R, Lindemann SR, Löffler FE, O'Malley MA, *et al*. Common principles and best practices for engineering microbiomes. Nat Rev Microbiol. 2019;17(12):725-41.
41. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: a standalone metatranscriptome analysis pipeline. BMC Bioinformatics. 2018;19(1):175.
42. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. Nat Biotechnol. 2018;36(2):190-5.
43. Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. Plant-microbiome interactions: from community assembly to plant health. Nat Rev Microbiol. 2020;18(11):607-21.
44. Biamonte J, Wittek P, Pancotti N, Rebentrost P, Wiebe N, Lloyd S. Quantum machine learning. Nature. 2017;549(7671):195-202.
45. Schuld M, Sinayskiy I, Petruccione F. An introduction to quantum machine learning. Contemp Phys. 2015;56(2):172-85.
46. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, *et al*. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37(8):852-7.
47. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al*. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537-41.
48. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13(7):581-3.
49. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, *et al*. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol. 2023;41(11):1633-44.

50. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257.
51. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, *et al*. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. 2021;10:e65088.
52. Tessler M, Neumann JS, Afshinnekoo E, Pineda M, Hersch R, Velho LFM, *et al*. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. Sci Rep. 2017;7(1):6589.
53. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, *et al*. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207-14.
54. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science. 2016;353(6305):1272-7.
55. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12(5):e0177459.
56. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29(5):415-20.
57. Katz LS, Griswold T, Morrison SS, Caravas JA, Zhang S, den Bakker HC, *et al*. Mashtree: a rapid comparison of whole genome sequence files. J Open Source Softw. 2019;4(44):1762.
58. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. 2016:081257.
59. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, *et al*. Function and functional redundancy in microbial systems. Nat Ecol Evol. 2018;2(6):936-43.
60. Kelly BJ, Gross R, Bittinger K, Sherrill-Mix S, Lewis JD, Collman RG, *et al*. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics. 2015;31(15):2461-8.
61. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, *et al*. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci USA. 2012;109(52):21390-5.
62. Van der Heijden MGA, Bardgett RD, Van Straalen NM. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol Lett. 2008;11(3):296-310.
63. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, *et al*. A global atlas of the dominant bacteria found in soil. Science. 2018;359(6373):320-5.
64. Philippot L, Raaijmakers JM, Lemanceau P, Van der Putten WH. Going back to the roots: the microbial ecology of the rhizosphere. Nat Rev Microbiol. 2013;11(11):789-99.
65. Crowther TW, Todd-Brown KEO, Rowe CW, Wieder WR, Carey JC, Machmuller MB, *et al*. Quantifying global soil carbon losses in response to warming. Nature. 2016;540(7631):104-8.
66. Wieder WR, Bonan GB, Allison SD. Global soil carbon projections are improved by modelling microbial processes. Nat Clim Chang. 2013;3(10):909-12.
67. Kardol P, Wardle DA. How understanding aboveground-belowground linkages can assist restoration ecology. Trends Ecol Evol. 2010;25(11):670-9.
68. Harris J. Soil microbial communities and restoration ecology: facilitators or followers? Science. 2009;325(5940):573-4.
69. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, *et al*. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457-63.
70. Maestre FT, Delgado-Baquerizo M, Jeffries TC, Eldridge DJ, Ochoa V, Gozalo B, *et al*. Increasing aridity reduces soil microbial diversity and abundance in global drylands. Proc Natl Acad Sci USA. 2015;112(51):15684-9.
71. Jansson JK, Hofmockel KS. Soil microbiomes and climate change. Nat Rev Microbiol. 2020;18(1):35-46.
72. Mendes R, Garbeva P, Raaijmakers JM. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS Microbiol Rev. 2013;37(5):634-63.