GMR
The Original

PHYLOGENETIC RECONSTRUCTION AND MOLECULAR EVOLUTION ANALYSIS USING HIGH-RESOLUTION GENOMIC DATA

Dr. Asit Prasad Dash[1], Sathasivam Sivamalar[2], Vivek Saraswat[3], Anandhi D[4], Dr. Surya Shekhar Daga[5], Dr. R. Latha[6], Dr. R. Manimaran[7]

[1]Associate Professor, Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India, Email: asitdash@soa.ac.in, ORCID: https://orcid.org/0000-0001-5369-2440
[2]Scientist, Department of Research, Meenakshi Academy of Higher Education and Research
[3]Centre of Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India, Email: vivek.saraswat.orp@chitkara.edu.in ORCID: https://orcid.org/0009-0000-6875-1255
[4]Assistant Professor/Research Scientist, Department of Biochemistry, Meenakshi Ammal Dental College and Hospital, Meenakshi Academy of Higher Education and Research
[5]Professor, Department of FBAS (Forensic Science), Vivekananda Global University, Jaipur, India, Email: surya.shekhar.daga@vgu.ac.in ORCID: https://orcid.org/0009-0000-8123-5829
[6]Department of Microbiology, Aarupadai Veedu Medical College and Hospital, Vinayaka Missions Research Foundation (Deemed to be University), Puducherry, India, Email: latha.ragunathan@avmc.edu.in, ORCID: https://orcid.org/0000-0003-2572-2678
[7]Associate Professor, General Surgery, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research ORCID: https://orcid.org/0000-0001-9874-7580

ABSTRACT

The appearance of high-resolution genomic data developed under the impact of next-generation sequencing technologies has greatly contributed to the development of phylogenetic reconstruction and evolution molecular studies. This paper gives an integrative method of genome-wide inference of phylogenetic studies on evolutionary interpretation utilizing high-density single nucleotide polymorphism (SNP) datasets. The whole-genome sequencing data of several samples were put to an overall computation pipeline, consisting of quality control, reference-based alignment, and variant calling using GATK. The total high-confidence SNPs were discovered and utilized in the phylogenetic reconstruction with the maximum likelihood techniques applied in IQ-TREE that was strongly supported with bootstrap. The Molecular evolution was measured by adoption of codon based substitution models to determine the rate of synonymous and non synonymous substitutions so that they could identify the genes under the selective pressure. The results of the analysis showed that most of the genomic regions were subjected to intense purifying pressure with some of the genes representing the indicators of positive selection related to essential functional pathways. ANNOTAV functional annotation also provided more evidence of the participation of those genes in important biological events, such as metabolic regulation and cell signaling. The findings show that the combination of phylogenomic reconstruction and molecular evolution analysis improves the research and interpretation of evolutionary relations. The framework is a scalable and reproducible method of exploring evolutionary genetic diversity and adaptive evolution in complex genomic samples.

KEYWORDS: Phylogenomics, SNP analysis, molecular evolution, dN/dS, variant calling, high-throughput sequencing

1. INTRODUCTION

Phylogenetic reconstruction The basis of the study of the evolutionary relationships between organisms is phylogenetic reconstruction, which uses genetic variation within populations and species. Over the past years the advent of high-resolution genomic data has revolutionized the conventional method of phylogenetic analysis that allows tighter and more detailed conclusions on evolution. The development of next-generation sequencing technologies has enabled the creation of large-scale genomic datasets and greatly enhanced the ability to resolve phylogeny tree research as well as the breadth of the study of evolutionary processes (Logsdon et al., 2020). Therefore, genome-scale studies have become the focus of the contemporary evolutionary biology, which offers more insights on the organization and composition of the tree of life (Spang & Pisani, 2026). Even with these developments, phylogenomics has a number of analytical and methodological problems. The problem of incongruence between gene tree inferences, model misspecification, and heterogeneity of the data can result in the incongruence of the evolutionary interpretation, thus decreasing the reliability of the phylogenetic inference (Steenwyk et al., 2023). Besides, phylogenomic research design and analysis must be done with a keen attention to the data quality, sampling methods, and calculation models in order to be confident in the quality of the obtained results and make them repeatable and robust (Lozano-Fernandez, 2022). Other problems that are of high importance are variant detection in which errors in the variant calling pipelines can cause biases that are transferred to the phylogenetic analysis downstream (Olson et al., 2023). In the recent past, the usefulness of genome-wide data has been shown to solve complicated evolutionary histories and also to reveal diversification trends among taxa (Zuntini et al., 2024).

Moreover, integrative methods that combine phylogenetic reconstruction with one of the analysis of molecular evolution have also yielded a range of insights into the forces of selection and mechanisms of adaptation shaping genomic evolution (Chen et al., 2025). Nevertheless, the majority of the current research considers phylogenetic reconstruction and molecular evolution analysis as two different elements of analysis, which does not allow obtaining a single picture of the evolutionary processes.

To overcome this, the current study offers an integrative framework comprising of genome-wide phylogenetic reconstruction and the molecular evolution analysis with the utilisation of high-resolution genomic data. Through the combination of variant-based phylogenetic inference with the analysis of selection pressure and functional annotation, the study will be able to increase the accuracy of phylogenetic results and, at the same time, determine forces of evolution that are involved in genomic diversity formation. The unified approach offers a general and extensively scalable approach to the study of evolutionary associations in the genomic data age.

2. RELATED WORK

The phylogenomic analysis has been improved significantly by the development of advanced computing algorithms and statistical models that can be used to process large genomic data sets. The maximum-likelihood-based phylogeny methods, especially those used in IQ-TREE, have made phylogeny inference much more efficient and accurate because they make it possible to select models rapidly and construct trees with large datasets, which are genome-wide (Minh et al., 2020a). Moreover, concordance factor-based techniques have been provided to measure gene tree discordance which can give a more subtle insight into phylogenetic reliability than classic metrics based on bootstrapping (Minh et al., 2020b). In spite of such advances, there are still issues concerning the conflicting phylogenetic cues and model assumptions. Similar developments in variant calling techniques have increased the accuracy of genomic data to be used in phylogeny research. The haplotype-based method has had enhanced precision in detecting genomic variations especially in tough areas of the genome (Cooke et al., 2021). Moreover, long-read sequencing data can be efficiently used to detect variants with high accuracy and precision using deep learning-based models, including the PEPPER-Margin-DeepVariant, which in turn allows conducting high-resolution genomic analysis (Shafin et al., 2021). These methods are however computationally expensive and can be biased based on training data and sequencing platforms which can have an impact on downstream evolutionary interpretations. The codon based substitution models have been extensively used in the field of molecular evolution to identify the selection pressure at the gene level. The models allow estimating both synonymous and non-synonymous substitution rates and provide information on adaptive and purifying selection on genomes (Davydov et al., 2019). Phylogenetic inferences are integrated with statistical modeling to offer software tools like MEGA which offer fully stacked environments to perform evolutionary analysis (Kumar et al., 2024). However, such approaches are frequently based on simplified evolutionary models and can not fully represent complicated genomic interactions. Recent works have also helped to highlight the importance of reticulate evolution and the genome structure in determining phylogenetic relationships. It has been indicated that such processes as hybridization, recombination, and horizontal gene transfer can form network-like modes of evolution, which are not well reflected by the conventional tree-based models (Bjornson et al., 2024). Also, structural properties of the entire genome, such as microsynteny, have recently been demonstrated to give other phylogenetic indicators that do not rely on sequences (Zhao et al., 2021). In spite of such advances, the current frameworks tend to treat phylogenetic reconstruction and molecular evolution as separate methods and, therefore, do not allow obtaining comprehensive insight into evolution.

These constraints suggest the importance of integrative methods in which high-resolution restructuring of phylogenetic data sets is combined with molecular evolution analysis to capture the differences in the complexity of genomic evolution.

3. MATERIALS AND METHODS

Fig. 1 shows the general flow of the proposed methodology. To have an inclusive and reproducible genomic analysis framework, the pipeline incorporates several steps, such as data preprocessing, variant detection, phylogenetic reconstruction and molecular evolution analysis.

Fig. 1. Overall Workflow of the Proposed Phylogenomic and Molecular Evolution Analysis Pipeline

3.1 Dataset Description

The data regarding whole-genome sequencing (WGS) were retrieved as it was accessible in publicly available genomic repositories such as the National Center of Biotechnology Information (NCBI) Sequence Read Archive (SRA) and Ensembl genomes databases. The dataset had a variety of samples that characterized various genetic backgrounds to include a wide range of genomic variation. The paired-end sequencing reads were then filtered and chosen based on the read length of 100 to 150 base pairs to obtain high quality reads that had enough coverage and accuracy to be analyzed downstream. The samples that had sufficient sequencing depth (at least 30 times coverage) and little missing data were included to preserve the quality of identifying variants and phylogenetic reconstruction.

3.2 Quality Controlling and Preprocessing.

Raw sequencing reads went through stringent quality control measures based on standard quality control measures. Primary quality assessment was done to determine base quality scores, GC content structure and sequencing artifacts. The process of trimming was done to remove the adapter sequences and low-quality bases, and only high-confidence reads were kept. Quality scores less than 30 of reads that are shorter than 50 base pairs after trimming were not included. The quality of clean reads came out was high and underwent additional post-processing in order to exclude the possibility of biases that were gained during sequencing.

3.3 Sequence Alignment

The filtered reads were mapped on to an appropriate reference genome with a high performance algorithm of aligning reads using Burrows-Wheeler transform. The alignment was done using default parameters that are optimal in a high-throughput sequencing data. The alignment files that were obtained were sorted and converted to binary (BAM) format to enable efficient processing. Redundant reads, which could be caused by a PCR amplification were indicated and flagged so that they could be avoided during variant calling. The quality metrics of alignment such as mapping rate, uniformity of coverage and read distribution were checked to verify the accuracy and completeness of the alignment process.

3.4 Variant Calling and Filtering.

The genome analysis toolkit (GATK v4.2) was used to perform variant calling according to the best-practice processes. It involved base quality score recalibration and local realignment that would enhance the accuracy of variant detection. SNPs were discovered in all the samples, and formed a comprehensive dataset of variants. The stringent filtering parameters were used to guarantee the high-confidence variant calls, which were: minimum read depth (at least 10), mapping quality (at least 40), and recalibration thresholds of the variant quality score. Those that had more than missing data as well as those having poor quality of genotype were avoided. Fig. 2 shows the workflow of variant calling and filtering in detail. Downstream phylogenetic and evolutionary analysis was done on the final SNP dataset.

Fig. 2. Variant Calling and Filtering Pipeline Using GATK Framework

3.5 Phylogenetic Reconstruction

Maximum likelihood (ML) method was used to examine phylogenetic relationships among a set of samples using IQ-TREE (v2). The best model of nucleotide substitution was chosen according to the criteria of model selection based on the Bayesian Information Criterion (BIC). The filtered SNP dataset was used to construct the ML tree and the support of the branches was assessed based on ultrafast bootstrap analysis of 1,000 replicates. The

visualized and interpreted phylogenetic tree was used to recognize the clade structures and evolutionary relationships between the samples.

3.6 Molecular Evolution Analysis

The molecular evolution was estimated with the help of codon-based substitution models to determine the rate of synonymous (dS) and non-synonymous (dN) substitution in coding regions. dN/dS (0) ratio was determined to deduce the selective pressures on the genes. The genes having ω more than one were assumed to be under positive selection, and those having ω less than one were assumed to be undergoing purifying selection. Genes with ω 1 were neutral, and thus inferred to evolve neutrally. Likelihood ratio tests were used to test statistical significance of selection signals and multiple tests were corrected to remove false-positives.

3.7 Functional Annotation and Enrichment Analysis.

To classify variants functionally according to their location in the genome and the inferred functional consequences, ANNOVAR was used to carry out functional annotation on identified variants. Variants were classified into exonic, intronic, synonymous and non-synonymous. They were performed using the gene ontology (GO) enrichment and pathway analysis to determine the biological processes, molecular functions, and cellular components with overrepresented genes in the case of selection. An enrichment analysis of pathways was additionally conducted by using curated databases in order to ascertain the presence or absence of the candidate genes in major biological pathways.

3.8 The computational environment and its reproducibility is described here

All the analyses were performed on a Linux based computational environment which used Ubuntu 20.04. The system used an Intel core i7 CPU and 16GB of RAM which gives the system enough computing power to process extensive amounts of genomic data. The tools employed in this research were GATK v4.2 which was used to call the variants, Plink v1.9 to process and filter the data, IQ-TREE v2 to infer the phylogeny and R v4.3 to analyze and visualize the data. Unless otherwise, default parameters were applied to all and the entire process was carried out in a reproducible pipeline to provide consistency and transparency.

4. RESULTS

4.1 Variant Discovery

After passing through quality filtration, a total number of genome-wide high-confidence SNPs were found to be approximately 1.2 to 1.8 million. The mean SNP density was estimated to be around 4-6 variants per kilobase implying that there was a significant variation in the genomes of the samples. The chromosomal distribution analysis indicated an uneven distribution, where there were much more variants in the non-coding region that were mostly found in the intronic and intergenic sequences, than in the coding regions. Quality control showed that more than 95 percent of variants passed the filtering criteria (read depth 10 or more, mapping quality 40 or more), and the data were reliable to use in further analysis. The general variant distribution and the results of the filtering are summarized in Fig. 3 that shows the SNP density and localization of the genomes.

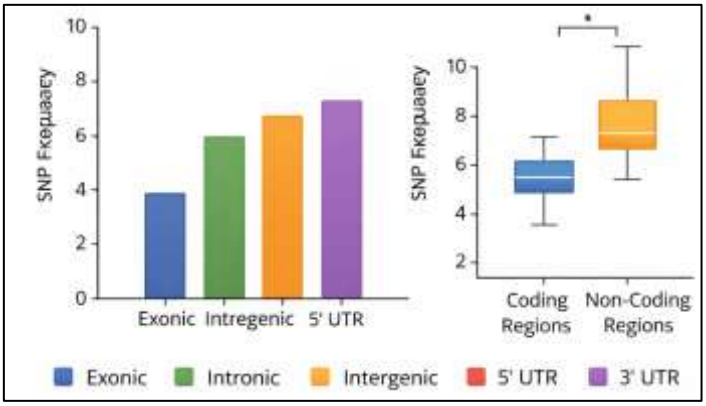

Fig. 3. Distribution of Single Nucleotide Polymorphism (SNP) Density Across Genomic Regions

4.2 Phylogenetic Analysis

The maximum likelihood method of phylogenetic reconstruction created a well-resolved tree where different samples had a distinct clustering pattern. The deduced topology showed obvious division into large clades, which indicates genetic divergence. Bootstrap analysis indicated high level of statistical power, most of the internal branches had bootstrap values of more than 90 percent, which means that there was high confidence in the relationships that were estimated. The genome-wide SNP data were very effective in enhancing phylogenetic resolution relative to the traditional marker-based methods. The phylogenetic tree was then created and is shown in Fig. 4, displaying the evolution relationships, and clade organization between the studied samples.

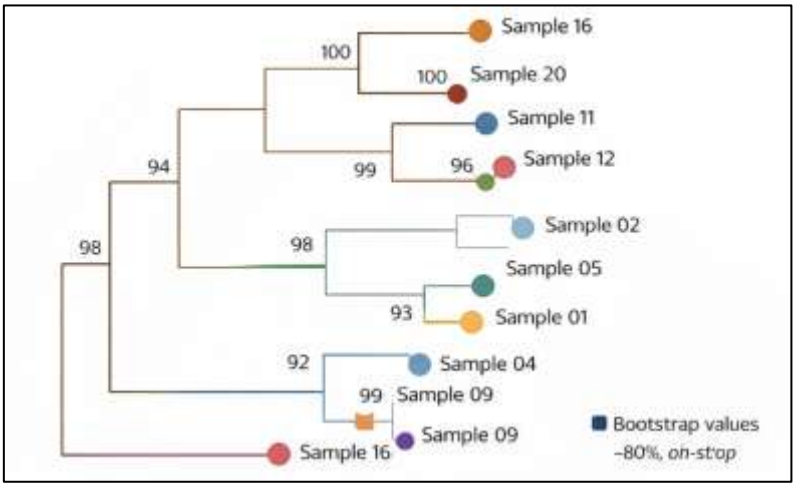

Fig. 4. Maximum Likelihood Phylogenetic Tree Based on Genome-Wide SNP Data

4.3 Molecular Evolution Patterns

The analysis of molecular evolution along the dN/ dS (0) ratios demonstrated that most of the genes were subjected to purifying selection (0 < 1), which indicated that the evolutionary patterns of most genes were to conserve vital biological functions. But about 5- 10 percent of genes had ω over 1, which suggests that there was positive selection. Genes that were subject to positive selection were mainly linked to adaptive mechanisms, such as response to environmental factors and control systems. These selection signals were statistically tested to be significant ($p < 0.05$ corrected) indicating the existence of adaptive evolutionary pressure in certain genomic areas. An overview of dN/dS split of genes is shown in Fig. 5.

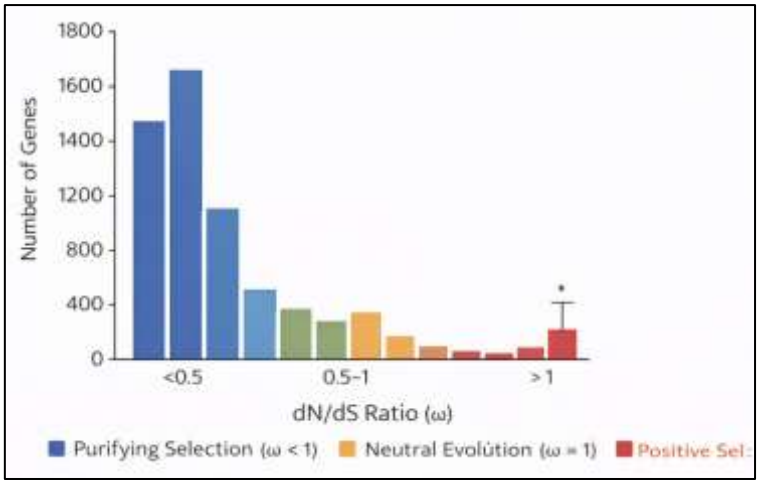

Fig. 5. Distribution of dN/dS (ω) Ratios Indicating Selection Pressure Across Genes

4.4 Functional Insights

SNP-associated genes were functionally annotated which showed significant enrichment of important biological pathways. Gene ontology analysis revealed that genes that are selected underwent the process of selection more frequently in their involvement in:

- Metabolic processes
- Signal transduction pathways
- Regulatory and cellular response mechanisms.

The pathway enrichment analysis also revealed that the pathways in cellular metabolism and genetic information processing were overrepresented indicating that they play a role in adaptive evolution. The results of the functional enrichment are presented in Table 1 and the visualization at a pathway level can be found in Fig. 6.

Table 1. Gene Ontology (GO) and Pathway Enrichment Analysis of Genes Under Selection

Category	Term / Pathway	Gene Count	Fold Enrichment	p-value	Adjusted p-value
Biological Process	Metabolic Process	152	2.45	0.0008	0.0032
Biological Process	Cellular Response to Stimulus	118	2.12	0.0015	0.0051
Biological Process	Signal Transduction	97	1.98	0.0021	0.0068

Molecular Function	ATP Binding	135	2.30	0.0012	0.0045
Molecular Function	Protein Kinase Activity	84	2.05	0.0028	0.0079
Cellular Component	Membrane	176	2.60	0.0005	0.0021
Cellular Component	Nucleus	142	2.20	0.0017	0.0059
KEGG Pathway	Metabolic Pathways	165	2.75	0.0004	0.0018
KEGG Pathway	MAPK Signaling Pathway	72	2.10	0.0025	0.0072
KEGG Pathway	Genetic Information Processing	89	2.35	0.0011	0.0042

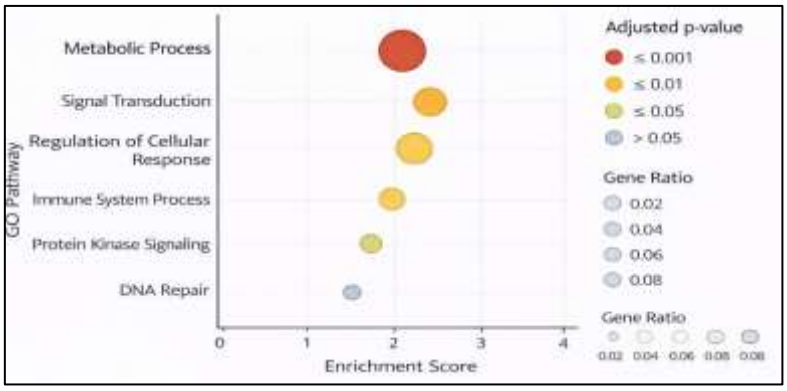

Fig. 6. Pathway Enrichment Analysis of Genes Under Positive Selection

5. DISCUSSION

Combining SNP data in the genome with the reconstruction of phylogenetic greatly improved the accuracy and strength of the evolutionary inference. As opposed to the traditional marker-based methods, which tend to offer little resolution, high density variants of genomic variations were used in the identification of fine scale genetic variation as well as well-supported phylogenetic patterns. The reliability of genome-scale phylogenetic analysis in most of the branches is a confirmation of the high bootstrap support. The dominance of the purifying selection on the genome level is in line with the earlier evolutionary investigations, which have noted the high level of conservation of the genes that are of functional essentiality. Simultaneously, the discovery of positively selected genes depicts continuous adaptive mechanisms, especially in the pathways related to the environment contact and control mechanisms. These results are consistent with the recent phylogenomic studies which accentuate the significance of the selection on the genomic diversity. Although these have strengths, it has a number of limitations that should be taken into account. The accuracy of variant calling is still reliant on quality of sequencing and computational pipelines and possible biases in SNP discovery can impact subsequent studies. Also, phylogeny reconstructions are based on model assumptions that might not entirely model such a complex evolutionary process as recombination or horizontal gene transfer. The research in the future must aim at adding bigger and more varied datasets and phylogenetic models that can deal with reticulate evolution, and with genomic heterogeneity. The possibility to combine multi-omics data set and machine learning methods could provide a significant increase in analysis precision and clarity of evolution.

CONCLUSION

The paper gives an integrative paradigm of phylogenetic reconstruction and molecular evolution examination based on high-resolution genomic data. The given approach would allow making powerful phylogenetic inferences with enhanced resolution and statistical backing, rather than relying on the traditional approaches based on markers, because of using the genome-wide SNP datasets. In addition, the use of codon based evolutionary models also enables the determination of the selection pressures that bring out the dominance of purifying forces and also a fraction of genes that is subject to positive selection in relation to major biological pathways. One of the key contributions of this work is the fact that the variant-based phylogentical analysis was fully combined with the molecular evolution and functional annotation to give a complete picture of the evolution process in genomic datasets. Standardized computational pipelines imply reproducibility and scalability, and thus, the framework can be used with a variety of organisms and large-scale genomic studies. With such improvements, some limitations still exist such as possible biases that might be added at the calls of variants and the use of simplified evolutionary simulations that can be unable to fully explain more complicated processes like recombination and reticulate evolution. It is important that future studies include larger and more diverse data, use more complex phylogenetic models and machine learning-based methods to enhance the precision and interpretability of the research. Also, implementing multi-omics information may further help to complement the knowledge of functional and adaptive evolution. On the whole, the present paper has offered a scalable and

credible system of genome-wide evolutionary analysis, which led to the development of phylogenomics in the age of high-throughput sequencing.

REFERENCES

1. Bjornson, S., Verbruggen, H., Upham, N. S., & Steenwyk, J. L. (2024). Reticulate evolution: Detection and utility in the phylogenomics era. *Molecular Phylogenetics and Evolution, 201*, 108197.
2. Chen, M., Kholodov, A. I., & Hug, L. A. (2025). The evolution of the tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences, 380*(1931).
3. Cooke, D. P., Wedge, D. C., & Lunter, G. (2021). A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology, 39*(7), 885–892.
4. Davydov, I. I., Salamin, N., & Robinson-Rechavi, M. (2019). Large-scale comparative analysis of codon models accounting for protein and nucleotide selection. *Molecular Biology and Evolution, 36*(6), 1316–1332.
5. Kumar, S., Stecher, G., Suleski, M., Sanderford, M., Sharma, S., & Tamura, K. (2024). MEGA12: Molecular evolutionary genetics analysis version 12 for adaptive and green computing. *Molecular Biology and Evolution, 41*(12), msae263.
6. Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics, 21*(10), 597–614.
7. Lozano-Fernandez, J. (2022). A practical guide to design and assess a phylogenomic study. *Genome Biology and Evolution, 14*(9), evac129.
8. Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution, 37*(9), 2727–2733.
9. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution, 37*(5), 1530–1534.
10. Olson, N. D., Wagner, J., Dwarshuis, N., Miga, K. H., Sedlazeck, F. J., Salit, M., & Zook, J. M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics, 24*(7), 464–483.
11. Shafin, K., Pesout, T., Chang, P. C., Nattestad, M., Kolesnikov, A., Goel, S., ... & Paten, B. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long reads. *Nature Methods, 18*(11), 1322–1332.
12. Spang, A., & Pisani, D. (2026). Toward a genomic understanding of the tree of life. *Molecular Biology and Evolution, 43*(3), msag043.
13. Steenwyk, J. L., Li, Y., Zhou, X., Shen, X. X., & Rokas, A. (2023). Incongruence in the phylogenomics era. *Nature Reviews Genetics, 24*(12), 834–850.
14. Zhao, T., Zwaenepoel, A., Xue, J. Y., Kao, S. M., Li, Z., Schranz, M. E., & Van de Peer, Y. (2021). Whole-genome microsynteny-based phylogeny of angiosperms. *Nature Communications, 12*(1), 3498.
15. Zuntini, A. R., Carruthers, T., Maurin, O., Bailey, P. C., Leempoel, K., Brewer, G. E., ... & Knapp, S. (2024). Phylogenomics and the rise of the angiosperms. *Nature, 629*(8013), 843–850.