

Web-Based Bioinformatics Tools and Databases for Functional and Comparative Genomics of Fabaceae Transcriptome Shotgun Assembly and Whole Genome Sequencing Datasets

Dr. V. K. Singh

Information Officer, Centre for Bioinformatics, School of Biotechnology, Institute of Science,
Banaras Hindu University, Varanasi, Uttar Pradesh 221005

Abstract - The Fabaceae family, encompassing major legumes such as *Ammopiptanthus mongolicus*, *Arachis duranensis*, *Arachis hypogaea*, *Arachis ipaensis*, *Arachis stenosperma*, *Glycine max*, *Glycyrrhiza uralensis*, *Indigofera argentea*, *Lathyrus oleraceus*, *Lathyrus sativus*, *Lens culinaris*, *Leucaena leucocephala*, *Lupinus angustifolius*, *Medicago sativa*, *Medicago truncatula*, *Pueraria montana*, *Senna alexandrina*, *Trigonella foenum-graecum*, *Vicia faba*, and *Zenia insignis*, is of significant agricultural and ecological importance. Advances in high-throughput sequencing technologies, including Transcriptome Shotgun Assembly (TSA) and Whole Genome Sequencing (WGS), have generated vast genomic and transcriptomic datasets for Fabaceae species. Efficient analysis and interpretation of these datasets require robust bioinformatics resources. Web-based bioinformatics tools, servers, and databases provide accessible, integrative, and interactive platforms for sequence retrieval, functional annotation, gene expression profiling, comparative genomics, and visualization. This manuscript reviews key web-based resources relevant to Fabaceae genomics, highlighting their roles in functional annotation, expression analysis, and comparative studies, emphasizing their importance in functional genomics, evolutionary studies, and crop improvement.

Keywords: Legumes, Fabaceae, Whole Genome Sequencing, Transcriptome Shotgun Assembly, Crop improvement

1. INTRODUCTION

Legumes are a diverse and nutritionally significant plant family with critical roles in nitrogen fixation, soil fertility, and human nutrition. Despite their importance, many legume species remain underutilized, with limited genomic and functional characterization. Comparative genomics and functional genomics approaches enable the identification of evolutionary relationships, gene families, and regulatory networks critical for adaptation, stress tolerance, and key agronomic traits.

The Fabaceae family, one of the largest plant families, includes species with high nutritional, ecological, and economic value. Legumes contribute to soil fertility via nitrogen fixation and serve as primary sources of protein and bioactive compounds. The development of high-throughput sequencing technologies such as TSA and WGS has generated enormous amounts of sequence data for Fabaceae species. However, the volume and complexity of these datasets necessitate computational tools for efficient data management, analysis, and interpretation. Web-based bioinformatics tools and databases provide user-friendly interfaces for accessing genomic and transcriptomic sequences, performing functional annotation, analyzing gene expression, exploring comparative genomics, and visualizing results without the need for extensive local computational infrastructure.

Recent advances in sequencing technologies and integrative bioinformatics resources, such as GenBank, SoyBase, and Phytozome, have facilitated large-scale analyses of legume genomes. Tools like OrthoVenn3, STRING, and PlantPAN3.0 allow exploration of orthologous gene families, protein-protein interaction networks, and transcriptional regulatory elements. In this study, we integrate genomic, transcriptomic, and phenomic data to uncover functional insights into legume genome evolution and regulatory mechanisms.

2. WEB-BASED GENOME AND TRANSCRIPTOME DATABASES

Online databases are foundational for storing, retrieving, and integrating TSA and WGS data in Fabaceae.

NCBI GenBank and TSA (<https://www.ncbi.nlm.nih.gov/genbank/>; <https://www.ncbi.nlm.nih.gov/Traces/wgs/>) host nucleotide and protein sequences from multiple legume species. Researchers can perform BLAST searches to identify homologous sequences, retrieve gene models, and download sequences for downstream analysis. Cross-links with Gene, BioProject, and PubMed facilitate integrated analysis.

Phytozome (<https://phytozome-next.jgi.doe.gov/>) provides fully sequenced plant genomes, including Fabaceae, along with gene annotations, protein families, and syntenic information. Comparative genomics tools allow identification of orthologs, gene family expansion, and evolutionary conservation.

Legume Information System (LIS) (<https://legumeinfo.org/>) integrates WGS, TSA, genetic maps, molecular markers, QTLs, and functional annotations. Its web interface supports genome visualization, sequence searches, and comparative analyses.

SoyBase (<https://soybase.org/>) focuses on *Glycine max*, offering genome browsers, gene models, expression data, and QTL mapping resources, aiding functional genomics and breeding applications.

Ensembl Plants (<https://plants.ensembl.org/>) provides interactive genome browsing, regulatory element visualization, variant tracking, and comparative genomics across Fabaceae species.

These databases allow researchers to access curated genomic resources, perform cross-species comparisons, and analyze both structural and functional genomics in Fabaceae.

3. WEB-BASED FUNCTIONAL ANNOTATION AND SEQUENCE ANALYSIS TOOLS

Functional annotation provides biological meaning to sequences derived from TSA and WGS datasets. Web-based servers integrate multiple predictive algorithms and databases:

BLAST (NCBI BLAST) (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) identifies homologous genes, conserved domains, and orthologs in Fabaceae.

InterProScan Web (<https://www.ebi.ac.uk/interpro/search/sequence/>) predicts protein domains, families, and motifs using integrated databases such as Pfam, SMART, PROSITE, and TIGRFAMs.

KEGG Mapper (<https://www.genome.jp/kegg/mapper.html>) maps Fabaceae genes to metabolic pathways and functional networks, enabling reconstruction of processes like nitrogen fixation and secondary metabolite biosynthesis.

STRING (<https://string-db.org/>) predicts protein-protein interactions, allowing the construction of regulatory and signaling networks for legume genes involved in stress responses, nodulation, and metabolism.

UniProtKB/Swiss-Prot (<https://www.uniprot.org/>) provides curated protein sequences and functional annotation, enhancing the quality of TSA and WGS gene characterization.

4. WEB-BASED TRANSCRIPTOME AND EXPRESSION ANALYSIS SERVERS

Gene expression analysis is central to functional genomics. Several web-based platforms support TSA and RNA-Seq data analysis:

Expression Atlas (<https://www.ebi.ac.uk/gxa/home>) provides curated gene expression data across tissues, developmental stages, and stress conditions in Fabaceae.

Genevestigator (<https://genevestigator.com/>) allows web-based exploration and meta-analysis of expression data, aiding functional inference of uncharacterized genes.

PlantPAN3.0 (<http://plantpan.itps.ncku.edu.tw/>) predicts promoter regions, transcription factor binding sites, and regulatory networks for genes derived from TSA/WGS data.

5. WEB-BASED COMPARATIVE GENOMICS AND ORTHOLOGY TOOLS

Comparative genomics enables identification of orthologs, syntenic blocks, and evolutionary relationships:

CoGe (<https://genomeevolution.org/coge/>) supports whole-genome comparisons, synteny visualization, and gene collinearity analysis in Fabaceae.

OrthoVenn3 (<https://orthovenn3.bioinfotoolkits.net/>) identifies and visualizes orthologous gene clusters, useful for analyzing conserved and unique gene families across legumes.

These websites offers genome browsers and comparative genomics tools for transcript-to-genome alignments and gene model validation.

6. WEB-BASED VISUALIZATION AND DATA INTEGRATION PLATFORMS

Visualization enhances interpretation of genomic and transcriptomic datasets:

IGV-Web (<https://igv.org/app/>) allows interactive exploration of gene models, alignments, and variants.

Circos Online (https://circos.ca/intro/tabular_visualization/) provides circular visualizations of genome structure, synteny, and transcriptomic relationships.

Ensembl Plants Genome Browser integrates TSA and WGS data for gene, variant, and regulatory element visualization.

These platforms facilitate **integrated analysis of genome and transcriptome data**, enabling identification of candidate genes, regulatory elements, and syntenic regions in Fabaceae.

7. CONCLUSION

Web-based bioinformatics tools, servers, and databases play a pivotal role in Fabaceae genomic research, offering accessible, integrative, and interactive solutions for sequence analysis, functional annotation, expression profiling, comparative genomics, and visualization. By removing computational barriers, these web resources enable researchers to maximize the utility of TSA and WGS data, supporting functional genomics, evolutionary studies, and crop improvement. Continued development of legume-focused web-based platforms will accelerate discovery of trait-associated genes, regulatory networks, and molecular markers critical for sustainable agriculture.

REFERENCES

- [1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- [2] Bolser, D. M., Staines, D. M., Perry, E., & Kersey, P. J. (2017). Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomic Data. *Methods in molecular biology (Clifton, N.J.)*, 1533, 1–31. https://doi.org/10.1007/978-1-4939-6658-5_1
- [3] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., Poux, S., Bougueleret, L., & Xenarios, I. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods in molecular biology (Clifton, N.J.)*, 1374, 23–54. https://doi.org/10.1007/978-1-4939-3167-5_2
- [4] Brown, A. V., Connors, S. I., Huang, W., Wilkey, A. P., Grant, D., Weeks, N. T., Cannon, S. B., Graham, M. A., & Nelson, R. T. (2021). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic acids research*, 49(D1), D1496–D1501. <https://doi.org/10.1093/nar/gkaa1107>
- [5] Chow, C. N., Lee, T. Y., Hung, Y. C., Li, G. Z., Tseng, K. C., Liu, Y. H., Kuo, P. L., Zheng, H. Q., & Chang, W. C. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic acids research*, 47(D1), D1155–D1163. <https://doi.org/10.1093/nar/gky1081>
- [6] Gonzales, M. D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W. D., & Waugh, M. E. (2005). The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic acids research*, 33(Database issue), D660–D665. <https://doi.org/10.1093/nar/gki128>
- [7] Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, 40(Database issue), D1178–D1186. <https://doi.org/10.1093/nar/gkr944>
- [8] Grover, J. W., Bomhoff, M., Davey, S., Gregory, B. D., Mosher, R. A., & Lyons, E. (2017). CoGe LoadExp+: A web-based suite that integrates next-generation sequencing data analysis workflows and visualization. *Plant direct*, 1(2), 10.1002/pld3.8. <https://doi.org/10.1002/pld3.8>

- [9] Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., & Zimmermann, P. (2008). Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Advances in bioinformatics*, 2008, 420747. <https://doi.org/10.1155/2008/420747>
- [10] Kanehisa, M., & Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein science : a publication of the Protein Society*, 29(1), 28–35. <https://doi.org/10.1002/pro.3711>
- [11] Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9), 1639–1645. <https://doi.org/10.1101/gr.092759.109>
- [12] Lu, Y., Li, M., Gao, Z., Ma, H., Chong, Y., Hong, J., Wu, J., Wu, D., Xi, D., & Deng, W. (2025). Advances in Whole Genome Sequencing: Methods, Tools, and Applications in Population Genomics. *International journal of molecular sciences*, 26(1), 372. <https://doi.org/10.3390/ijms26010372>
- [13] Ojuederie, O. B., Akpojotor, U. L., Adeniji, A. A., Ojuederie, T. C., Popoola, J. O., & Babalola, O. O. (2025). Comparative genomic analysis of underutilized legumes: insights into evolutionary relationships, genome evolution and stress tolerance. *Biotechnology reports (Amsterdam, Netherlands)*, 48, e00918. <https://doi.org/10.1016/j.btre.2025.e00918>
- [14] Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A. F., Jupp, S., Marioni, J., Meyer, K., Petryszak, R., ... Brazma, A. (2020). Expression Atlas update: from tissues to single cells. *Nucleic acids research*, 48(D1), D77–D83. <https://doi.org/10.1093/nar/gkz947>
- [15] Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic acids research*, 33(Web Server issue), W116–W120. <https://doi.org/10.1093/nar/gki442>
- [16] Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., Feldgarden, M., Fine, A. M., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T. D., O'Sullivan, C., Schmieder, E., Skripchenko, Y., ... Pruitt, K. D. (2025). Database resources of the National Center for Biotechnology Information in 2025. *Nucleic acids research*, 53(D1), D20–D29. <https://doi.org/10.1093/nar/gkae979>
- [17] Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. *Nucleic acids research*, 47(D1), D94–D99. <https://doi.org/10.1093/nar/gky989>
- [18] Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., & Wang, Y. (2023). OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic acids research*, 51(W1), W397–W403. <https://doi.org/10.1093/nar/gkad313>
- [19] Szklarczyk, D., Nastou, K., Koutrouli, M., Kirsch, R., Mehryar, F., Hachilif, R., Hu, D., Peluso, M. E., Huang, Q., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & von Mering, C. (2025). The STRING database in 2025: protein networks with directionality of regulation. *Nucleic acids research*, 53(D1), D730–D737. <https://doi.org/10.1093/nar/gkae1113>
- [20] Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2), 178–192. <https://doi.org/10.1093/bib/bbs017>
- [21] Wang, H., Wu, Z., Zuo, Y., Yan, X., Zou, B., Chen, Y., Yuan, Z., & Du, Z. (2025). Phenomics, RNA sequencing and weighted gene co-expression network analysis reveals key regulatory networks and genes involved in the determination of seed hardness in vicia sativa. *BMC genomics*, 26(1), 950. <https://doi.org/10.1186/s12864-025-12138-z>
- [22] Wang, J., Chen, Y., & Zou, Q. (2023). Comparative Genomics and Functional Genomics Analysis in Plants. *International journal of molecular sciences*, 24(7), 6539. <https://doi.org/10.3390/ijms24076539>
- [23] Zhang, H., Yasmin, F., & Song, B. H. (2019). Neglected treasures in the wild - legume wild relatives in food security and human health. *Current opinion in plant biology*, 49, 17–26. <https://doi.org/10.1016/j.pbi.2019.04.004>