A Comprehensive Workflow for Genome-Wide Analysis: Tools, Techniques, and Applications

Akhila K Kumar, Mudita Prakash* and Shruti Banswal**

*Institut für Biologie, Otto-von-Guericke University, Leipziger Straße 44, 39120 Magdeburg, Germany

**Associate Professor, Department of Biomedical Science, Shaheed Rajguru College of Applied Sciences for Women, Vasundhara Enclave, University of Delhi-110096.

Abstract

Genome-wide analysis (GWA) is a cornerstone in modern genomics, allowing researchers to uncover gene function, regulation, and evolutionary relationships across various organisms. This review presents a detailed overview of the GWA workflow, including key bioinformatics tools, databases, and techniques. It covers gene identification and annotation, sequence alignment, phylogenetic analysis, motif and domain prediction, gene structure and chromosomal mapping, synteny analysis, cis-regulatory element identification, subcellular localization prediction, and gene expression profiling. The integration of these techniques facilitates a systems-level understanding of genomic functions and adaptations. This review is intended to serve as a comprehensive guide for students and researchers in the fields of functional genomics, evolutionary biology, and molecular plant biology.

Keywords

Genome-wide analysis, gene annotation, phylogenetic analysis, motif prediction, cis-regulatory elements, subcellular localization, gene expression profiling, synteny analysis

1. Introduction

The rapid advancement of genome sequencing technologies has made genome-wide analysis (GWA) a fundamental approach in modern biology. GWA encompasses the study of entire genomes to determine the structural, functional, and regulatory components of genes. It is instrumental in linking genetic variation to phenotypic traits, whether in plants, animals, or humans. For example, understanding how genes respond to stress in plants can help in breeding more resilient crops, while in humans, GWA can reveal genetic predispositions to diseases. The power of GWA lies in its ability to integrate vast amounts of genomic data and identify patterns that were previously unobservable with traditional gene-by-gene approaches.

2. Overview of the GWA workflow

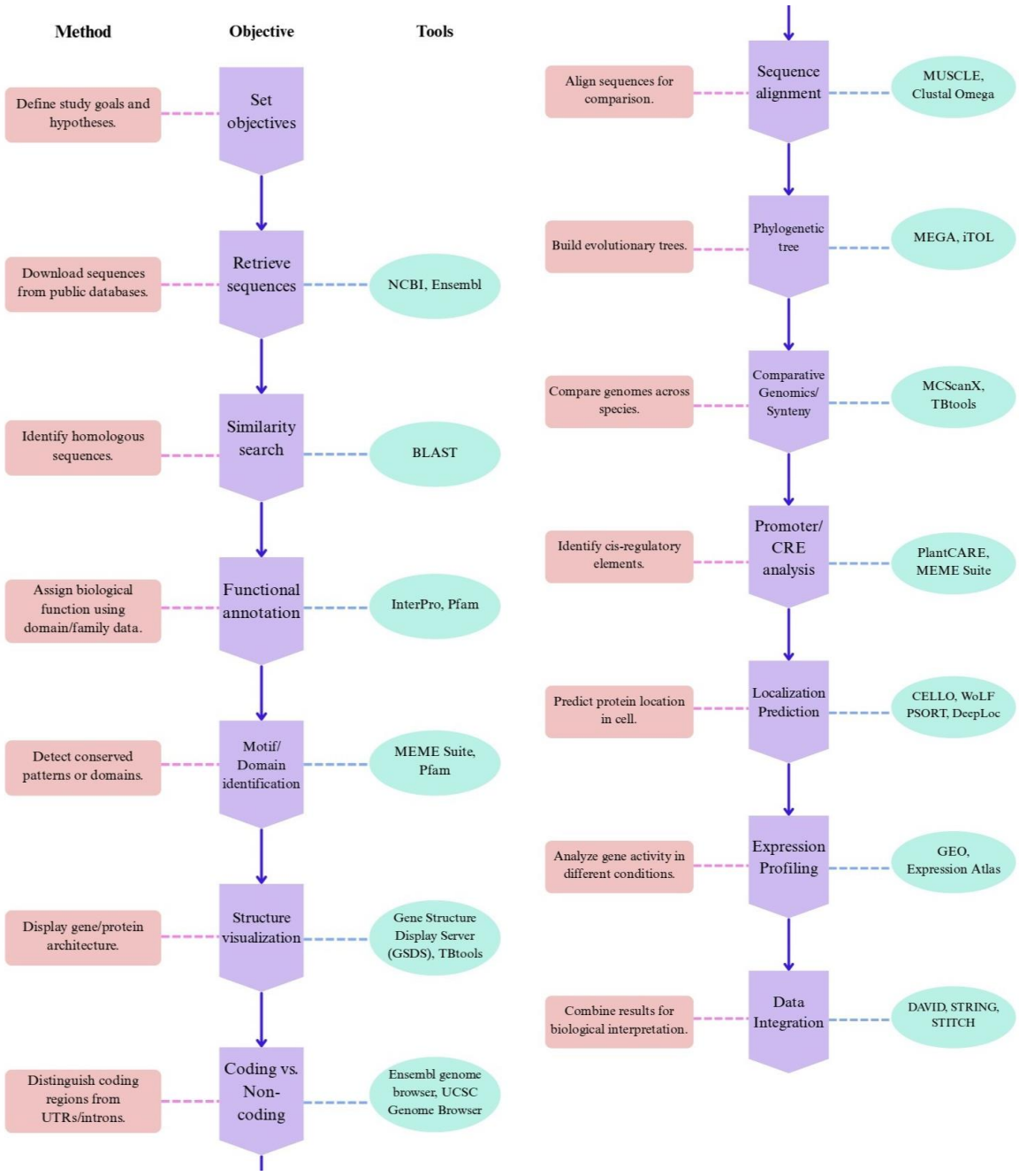

Figure 1: Overview of the GWA workflow

3. Tools and databases

A wide range of bioinformatics tools and databases support genome-wide analysis workflows. Sequence and protein databases such as NCBI, Ensembl, and UniProt serve as primary resources for retrieving gene and protein information. Tools like BLAST, MUSCLE, and CLUSTAL OMEGA facilitate similarity searches and multiple sequence alignments, while phylogenetic tools like MEGA and iTOL enable evolutionary analysis. Visualization platforms, including Geneious and Jalview, streamline data interpretation. Motif and domain identification is handled by MEME Suite and InterProScan, whereas TMHMM and SignalP assist in protein structure and trafficking prediction. Enrichment and interaction analysis are

performed using DAVID, STRING, and STITCH. Lastly, tools like CELLO, DeepLoc, and TargetP support subcellular localization prediction.

See Appendix: Tables of tools and software used

4. Gene Identification and Annotation

The first step in GWA involves retrieving and identifying genes of interest from genomic databases. BLAST is commonly used to align sequences and identify homologs (Altschul et al., 1990). BLASTn aligns nucleotide sequences, while BLASTp works with protein sequences. Ensembl (Yates et al., 2020), NCBI Gene (Sayers et al., 2025), and UniProt (Bateman et al., 2024) provide detailed annotations, including gene structure, function, and known variants. Functional annotation is refined using Pfam (Paysan-Lafosse et al., 2025), SMART (Letunic et al., 2020), and InterPro (Blum et al., 2024), which predict conserved domains and motifs. These annotations allow for distinguishing orthologous genes, understanding evolutionary conservation, and inferring gene functions based on known homologs.

5. Sequence Alignment and Phylogenetic Analysis

Multiple sequence alignment (MSA) helps identify conserved regions across genes or species. Tools like MUSCLE (Edgar, 2004) and CLUSTAL Omega (Sievers et al., 2011) provide accurate alignments. MUSCLE is favored for moderate-size datasets, while CLUSTAL Omega handles large, diverse sequences efficiently. Phylogenetic trees constructed using MEGA (Tamura, Stecher & Kumar, 2021) reveal evolutionary relationships. Methods like Neighbor-Joining (NJ) (Saitou & Nei, 1987), Maximum Likelihood (ML) (Felsenstein, 1981), and Bayesian Inference (Huelsenbeck et al., 2001) vary in complexity and accuracy. NJ is fast but less accurate, ML is more precise with model-based assumptions, and Bayesian analysis provides statistical rigor at higher computational cost. These tools aid in tracing gene evolution and predicting gene functions based on evolutionary context.

6. Motif and Domain Analysis

Motifs are short conserved sequences within DNA or protein that signify functional regions such as binding sites or enzymatic activity. MEME Suite identifies these motifs without prior knowledge of sequence structure (Li et al., 2022). For protein domains, Pfam and InterProScan use HMMs to classify sequences into families (Paysan-Lafosse et al., 2025; Blum et al., 2024). Understanding domain architecture helps predict the molecular function and interaction capability of proteins. For example, kinase domains indicate signaling roles, while transmembrane domains suggest localization to membranes. Domain arrangements also inform on possible gene duplications or functional innovations (Blum et al., 2024).

7. Gene Structure and Chromosomal Mapping

Gene structure analysis distinguishes between exons, introns, UTRs, and regulatory elements. Ensembl and UCSC Genome Browser provide visualization tools to explore these features (Yates et al., 2020; Kent et al., 2002). IGV (Integrative Genomics Viewer) is another tool for visualizing gene expression and structure in genomic context (Thorvaldsdottir et al., 2012). Chromosomal mapping assigns genes to specific loci, facilitating QTL mapping and candidate gene identification. Genome browsers display interactive chromosome maps with annotated features, allowing users to explore the genetic architecture and its correlation to phenotypes (Kent et al., 2002).

8. Cis- Regulatory Element and Synteny Analysis

Cis-regulatory elements are key non-coding sequences that modulate gene expression. These include promoters, enhancers, silencers, and response elements. Tools like MEME Suite are used to discover motifs de novo (Li et al., 2022), while databases such as PlantCARE are utilized to annotate known regulatory elements. These elements are critical in controlling gene behavior under various physiological and stress conditions, particularly in plants. Synteny analysis compares the order and orientation of genes across species. It helps in identifying conserved genomic blocks, providing insight into gene conservation, duplication, or divergence. Tools like TBtools and MCScanX allow visualization of syntenic relationships, helping researchers trace gene evolution, chromosomal rearrangements, and predict gene function across species.

9. Subcellular Localization Prediction

Proteins perform specific functions depending on their cellular location. Predicting subcellular localization helps infer protein function. Tools like CELLO (Yu et al., 2004), WoLF PSORT (Horton et al., 2007), and DeepLoc (Thumuluri et al., 2022) use machine learning to predict localization to the nucleus, mitochondria, chloroplasts, membranes, or other compartments. SignalP is used to predict secretory signal peptides (Teufel et al., 2022), while TargetP predicts targeting to organelles. Understanding protein localization supports functional annotation, especially when dealing with unknown or novel proteins, and enhances the understanding of cellular processes in health, development, or stress conditions.

10. Gene Expression Profiling

Gene expression profiling reveals how genes respond to different conditions such as drought, pathogens, or development stages. Tools like GEO (Sayers et al., 2025) and Expression Atlas provide experimental datasets across species and conditions. Differential expression analysis helps identify upregulated or downregulated genes. Heatmaps are commonly used to visualize expression patterns. This analysis links genotype to phenotype and is invaluable for discovering

stress-related or disease-associated genes. In agricultural research, expression profiling helps identify targets for breeding or genetic engineering to improve stress resistance or yield.

11. Integration and Data Interpretation

Genome-wide analysis integrates data from sequence alignment, domain analysis, regulatory mapping, localization prediction, and expression profiling. The combined interpretation provides holistic insights into gene function, regulation, and evolution. Tools like DAVID (Sherman et al., 2022), STRING (Szklarczyk et al., 2022), and STITCH (Szklarczyk et al., 2015) further aid in contextualizing data through enrichment and interaction networks. These insights are essential for targeted crop improvement, functional genomics, disease diagnostics, and evolutionary biology.

12. Conclusion

Genome-wide analysis is a robust framework for understanding the intricate details of gene structure, function, and regulation. The tools and techniques discussed enable researchers to decode genomic complexity, identify trait-associated genes, and explore evolutionary trajectories. From fundamental research to applied agriculture and biomedicine, GWA provides the foundation for advancing knowledge and innovation. The integration of machine learning, omics data, and large-scale analytics is set to further enhance the precision and utility of genome-wide approaches, ushering in an era of data-driven biology.

References

1. Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bourexis, D., Brister, J. R., Bryant, S. H., Canese, K., Cavanaugh, M., Charowhas, C., Clark, K., Dondoshansky, I., Feolo, M., Fitzpatrick, L., Funk, K., Geer, L. Y., Gorelenkov, V., Zbicz, K. (2017). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 46(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095
2. 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
3. Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. Life Science Alliance, 2(5), e201900429. https://doi.org/10.26508/lsa.201900429
4. Ashburner, M., Ball, C., Blake, J. et al. Gene Ontology: tool for the unification of biology. Nat Genet 25, 25–29 (2000). https://doi.org/10.1038/75556
5. Bateman, A., Martin, M., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett, E. H., Bye-A-Jee, H., Carpentier, D., Denny, P., Fan, J., Garmiri, P., Da Costa

Gonzales, L. J., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Zhang, J. (2024). UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Research, 53(D1), D609–D617. https://doi.org/10.1093/nar/gkae1010

6. Bateman, A., Martin, M., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Zhang, J. (2022). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052
7. Blum, M., Andreeva, A., Florentino, L. C., Chuguransky, S. R., Grego, T., Hobbs, E., Pinto, B. L., Orr, A., Paysan-Lafosse, T., Ponamareva, I., Salazar, G. A., Bordin, N., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunic, I., Llinares-López, F., Bateman, A. (2024). InterPro: the protein sequence classification resource in 2025. Nucleic Acids Research, 53(D1), D444–D456. https://doi.org/10.1093/nar/gkae1082
8. Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Flicek, P. (2021). Ensembl 2022. Nucleic Acids Research, 50(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049
9. Eddy, S. R. (2011). Accelerated Profile HMM searches. PLoS Computational Biology, 7(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195
10. Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340
11. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2018). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427–D432. https://doi.org/10.1093/nar/gky995
12. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution, 17(6), 368–376. https://doi.org/10.1007/bf01734359
13. Felsenstein, J. (1985). Confidence Limits on Phylogenies: An approach using the Bootstrap. Evolution, 39(4), 783. https://doi.org/10.2307/2408678
14. Horton, P., Park, K., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., & Nakai, K. (2007). WoLF PSORT: protein localization predictor. Nucleic Acids Research, 35(Web Server), W585–W587. https://doi.org/10.1093/nar/gkm259
15. Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., Houdaigui, B. E., Fatima, R., Flicek, P. (2020). Ensembl 2021. Nucleic Acids Research, 49(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

16. Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian Inference of Phylogeny and its impact on Evolutionary Biology. Science, 294(5550), 2310–2314. https://doi.org/10.1126/science.1065889
17. Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27
18. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199
19. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, A. D. (2002). The Human Genome Browser at UCSC. Genome Research, 12(6), 996–1006. https://doi.org/10.1101/gr.229102
20. Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. Journal of Molecular Biology, 305(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315
21. Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research, 49(W1), W293–W296. https://doi.org/10.1093/nar/gkab301
22. Li, X., Ma, L., Mei, X., Liu, Y., & Huang, H. (2022). ggmotif: An R Package for the extraction and visualization of motifs from MEME software. PLoS ONE, 17(11), e0276979. https://doi.org/10.1371/journal.pone.0276979
23. Matthias Blum, Antonina Andreeva, Laise Cavalcanti Florentino, Sara Rocio Chuguransky, Tiago Grego, Emma Hobbs, Beatriz Lazaro Pinto, Ailsa Orr, Typhaine Paysan-Lafosse, Irina Ponamareva, Gustavo A Salazar, Nicola Bordin, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunic, Felipe Llinares-López, Aron Marchler-Bauer, Laetitia Meng-Papaxanthos, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Damiano Piovesan, Catherine Rivoire, Christian J A Sigrist, Narmada Thanki, Françoise Thibaud-Nissen, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, InterPro: the protein sequence classification resource in 2025, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D444–D456, https://doi.org/10.1093/nar/gkae1082
24. Pierleoni, A., Martelli, P. L., Fariselli, P., & Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. Bioinformatics, 22(14), e408–e416. https://doi.org/10.1093/bioinformatics/btl222
25. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nature Biotechnology, 29(1), 24–26. https://doi.org/10.1038/nbt.1754
26. Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. https://doi.org/10.1093/oxfordjournals.molbev.a040454
27. 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

28. Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research, 50(W1), W216–W221. https://doi.org/10.1093/nar/gkac194
29. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7(1). https://doi.org/10.1038/msb.2011.75
30. Sigrist, C. J. A. (2002). PROSITE: A documented database using patterns and profiles as motif descriptors. Briefings in Bioinformatics, 3(3), 265–274. https://doi.org/10.1093/bib/3.3.265
31. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A. L., Fang, T., Doncheva, N. T., Pyysalo, S., Bork, P., Jensen, L. J., & Von Mering, C. (2022). The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research, 51(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000
32. Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2015). STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Research, 44(D1), D380–D384. https://doi.org/10.1093/nar/gkv1277
33. Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution, 38(7), 3022–3027. https://doi.org/10.1093/molbev/msab120
34. Teufel, F., Armenteros, J. J. A., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., Von Heijne, G., & Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. Nature Biotechnology, 40(7), 1023–1025. https://doi.org/10.1038/s41587-021-01156-3
35. Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2012). 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
36. Thumuluri, V., Armenteros, J. J. A., Johansen, A. R., Nielsen, H., & Winther, O. (2022). DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Research, 50(W1), W228–W234. https://doi.org/10.1093/nar/gkac278
37. Typhaine Paysan-Lafosse, Antonina Andreeva, Matthias Blum, Sara Rocio Chuguransky, Tiago Grego, Beatriz Lazaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Felipe Llinares-López, Laetitia Meng-Papaxanthos, Lucy J Colwell, Nick V Grishin, R Dustin Schaeffer, Damiano Clementel, Silvio C E Tosatto, Erik Sonnhammer, Valerie Wood, Alex Bateman, The Pfam protein families database:

embracing AI/ML, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D523–D534, https://doi.org/10.1093/nar/gkae997

38. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033
39. Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C. G., Flicek, P. (2020). Ensembl 2020. Nucleic acids research, 48(D1), D682–D688. https://doi.org/10.1093/nar/gkz966
40. Yu, C., Lin, C., & Hwang, J. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science, 13(5), 1402–1406. https://doi.org/10.1110/ps.03479604

Appendix: Tables of tools and software used

Table 1: Databases

Database	Function	Use Case	Reference
NCBI	Repository of genomic and protein data	For sequence retrieval and variant analysis	(Sayers et al., 2024)
Ensembl	Genome annotations and ortholog data	For comparative genomics and gene structure	(Yates et al., 2019)
UniProt	Protein sequence, function, structure	For protein characterization	(Bateman et al., 2024)
Pfam	Protein families/domains using HMMs	For domain detection	(Paysan-Lafosse et al., 2024)
InterPro	Integrates domain/family databases	For complete domain/motif overview	(Blum et al., 2024)
GO	Functional classification of genes	For enrichment analysis	(Ashburner et al., 2000)
KEGG	Pathway mapping for genes/proteins	For pathway and interaction analysis	(Kanehisa, 2000)

Table 2: Sequence alignment tools

Tool	Function	Use Case	Reference
BLAST	Finds sequence similarity	For finding homologs and variants	(Altschul et al., 1990)
MUSCLE	High-accuracy MSA	For conserved motifs and evolution	(Edgar, 2004)
Clustal Omega	Scalable MSA for large datasets	For diverse species alignment	(Sievers et al., 2011)

Table 3: Phylogenetic analysis tools

Tool	Function	Use Case	Reference
MEGA	Builds phylogenetic trees	For visualizing evolutionary history	(Tamura et al., 2021)

Table 4: Visualization tools

Tool	Function	Use Case	Reference
iTOL	Annotated phylogenetic trees	For interactive and custom tree views	(Letunic & Bork, 2021)

Table 5: Motif/domain analysis tools

Tool	Function	Use Case	Reference
MEME Suite	Finds unknown sequence motifs	For regulatory or functional motifs	(Li et al., 2022)
InterProScan	Identifies domains/families	For comprehensive protein annotation	(Blum et al., 2024b)

Table 6: Protein analysis tools

Tool	Function	Use Case	Reference
TMHMM	Predicts membrane-spanning regions	For studying membrane proteins	(Krogh et al., 2001)

Table 7: Enrichment analysis platforms

Tool	Function	Use Case	Reference
DAVID	Enrichment and annotation analysis	For interpreting large gene lists	(Sherman et al., 2022)
STITCH	Protein-chemical interactions	For drug/pathway interaction studies	(Szklarczyk et al., 2015)
STRING	Protein-protein interactions	For functional network analysis	(Szklarczyk et al., 2022)

Table 8: Tools for predicting localization

Tool	Function	Use Case	Reference
CELLO	ML-based localization predictor	For broad subcellular prediction	(Yu et al., 2004)
WoLF PSORT	Predicts localization in eukaryotes	For organelle-specific prediction	(Horton et al., 2007)
DeepLoc	Deep learning localization predictor	For complex localization cases	(Thumuluri et al., 2022)