Feature Review Open Access

Building an Integrated Multi-Omics Database for Rare Diseases

Huixian Li, Jingqiang Wang

Institute of Life Science, Jiyang College of Zhejiang A&F University, Zhuji, 311800, China

Corresponding author: jingqiang.wang@jicat.org

Computational Molecular Biology, 2025, Vol.15, No.3 doi: 10.5376/cmb.2025.15.0012

Received: 11 Mar., 2025

Accepted: 22 Apr., 2025

Published: 14 May, 2025

Preferred citation for this article:

Li H.X., and Wang J.Q., 2025, Building an integrated multi-omics database for rare diseases, Computational Molecular Biology, 15(3): 122-130 (doi: 10.5376/cmb.2025.15.0012)

Abstract Rare diseases are diverse in types and have a small number of patients with each type, but they cumulatively affect hundreds of millions of patients worldwide. Current research on rare diseases is confronted with challenges such as scattered data, inconsistent standards and difficulties in sharing. This article reviews the characteristics of the existing major rare disease databases (such as Orphanet, RD-Connect, MONDO, etc.), discusses the progress and limitations of multi-omics data integration methods, and introduces the new trend of data-driven rare disease research in the era of precision medicine. The application prospects of this database in discovering disease markers and therapeutic targets, supporting clinical decision-making and patient stratification, integrating artificial intelligence prediction models and drug reuse, etc. were explored. The contributions and main findings of this study were summarized. The potential impact of this integrated database on rare disease research and clinical translation was emphasized, and ideas for future expansion and sustainable development were proposed.

Keywords Rare diseases; Multi-omics; Data integration; Database architecture; Duchenne muscular dystrophy

1 Introduction

Rare diseases refer to those that affect very few people. In the European Union, it is defined as a disease with a prevalence rate of less than 1 in 2000, while in the United States, it refers to a disease affecting fewer than 200 000 people. It is known that there are over 7,000 rare diseases. Although each disease has a small number of patients, the total number of patients affected by it amounts to 263 to 446 million, accounting for approximately 3.5% to 5.9% of the global population. Most rare diseases are genetic disorders, with about 70 to 80 percent having genetic causes, and they often occur in childhood. Due to the wide variety of diseases and complex and diverse symptoms, patients with rare diseases often encounter problems such as difficult diagnosis and delayed diagnosis (Casas-Alba et al., 2022). According to statistics, it takes an average of many years from the appearance of symptoms to a confirmed diagnosis, and one has to visit multiple departments. This is called "diagnostic roaming".

The number of patients with rare diseases is small and their distribution is scattered. A single center often finds it difficult to collect sufficient samples, resulting in severe data fragmentation and isolation. The problem of "information silos" is prominent. The patient registration systems, sample banks and databases established by different research institutions are independent of each other and lack unified standards, making it difficult to share data. As Marsh et al. pointed out, "Data silos are hindering drug development and harming patients with rare diseases". This fragmentation limits researchers' understanding of the full picture of the disease and also hinders large-sample studies across centers.

Meanwhile, the data types of rare diseases are diverse and highly heterogeneous, and there are technical obstacles in integrated analysis, including different omics data such as genomic variations, transcriptional expression, protein abundance, metabolite profiles, and epigenetic modifications (Hesterlee et al., 2021). These data have different measurement methods, inconsistent data formats and scales, and require complex normalization processing and coordination. Even within the same data category, the technical platforms and analysis processes adopted by different studies may vary. For instance, differences in sequencing depth, mass spectrometry instruments, and data preprocessing methods can lead to batch effects and noise. High-dimensional and high-noise data pose challenges to statistical analysis and machine learning. Multi-omics data often have problems such as missing values, small sample sizes but tens of thousands of variables (Liu et al., 2022). If not integrated properly, it may "increase complexity rather than improve performance". Therefore, how to maintain biological information while reducing noise and filling in the gaps is a major difficulty in data integration.

2 Current Situation of Rare Diseases

2.1 Overview of the existing rare disease databases

Before building a new multi-omics database for rare diseases, it is essential to review existing databases to learn from their strengths and identify remaining gaps. Orphanet is a comprehensive international portal for rare diseases, supported by the EU. It has developed the OFA numbering system and provides detailed clinical, genetic, and treatment information for over 6 000 rare diseases (Mitani and Haneuse, 2020). Its major strength lies in high-quality, expert-curated clinical data and standardized ontologies.

OMIM (online mendelian inheritance in man) focuses on Mendelian genetic disorders and serves as a gene-centric reference for researchers. It provides gene-disease associations, locus information, and mutation types, which are critical for identifying candidate genes for rare diseases. However, like Orphanet, OMIM does not host raw omics data and often needs to be integrated with external sequencing or expression databases. Other emerging platforms such as RareDDB and eRAM offer integrated views by combining disease annotations with SNPs, genes, phenotypes, and even drug links, offering promising resources for precision medicine applications (Jia et al., 2018).

2.2 Progress and limitations of multi-omics integration methods

Multi-omics integration has emerged as a powerful approach in rare disease research but presents multiple challenges, especially due to small sample sizes and high dimensionality. Rare disease datasets often have very few samples and tens of thousands of variables. This makes traditional machine learning algorithms prone to overfitting and limits generalizability. Batch effects from different platforms (e.g., mass spectrometry, sequencing depth) can introduce noise. Tools like Combat and Harmony can help reduce this noise in single-omics, but a unified framework for multi-omics correction remains lacking (Olexiouk, 2023).

Interpretability is another key limitation. Although complex models like deep neural networks or multi-omics graphs can achieve high accuracy, their outputs are not easily mapped to biological pathways or mechanisms without additional analysis (Braconi et al., 2021; Zaghlool and Attallah, 2022). This hinders clinical translation and undermines trust among medical practitioners.

2.3 New trends in precision medicine and data-driven rare disease research

With the rise of precision medicine, large-scale genomic projects like the UK’s 100,000 Genomes Project have incorporated rare diseases into public health systems, significantly improving diagnosis rates (Figure 1) (Kerr et al., 2020). Countries are increasingly promoting the interconnection of clinical and research data platforms, such as UDNI, which facilitates global case sharing.

Therapeutically, the emergence of platform-based technologies like AAV gene therapy and antisense oligonucleotides supports rapid adaptation for different single-gene disorders. However, their development still relies heavily on integrative databases capable of connecting genotype to phenotype and underlying biological pathways (Pahelkar et al., 2024).

Ultimately, building a unified, multi-omics rare disease knowledge platform is not just a research goal-it is a critical enabler of faster diagnosis, improved therapy development, and clinical decision-making in a data-driven healthcare ecosystem.

3 Data Sources and Collection Strategies

3.1 Data types and schemas

The data scope of this integrated database covers typical "five major" omics data types and related clinical phenotypic data. The following respectively introduces each data type and their significance in rare disease research, including DNA sequence variation information obtained from whole genome sequencing (WGS), whole exome sequencing (WES), etc. For single-gene hereditary diseases, genomic variations are often the root cause of the disease. This database collects confirmed pathogenic variants of patients, candidate variant lists, and raw sequencing data (such as FASTQ/BAM), etc. If the patient is from an existing database such as ClinVar, we also record their variant pathogenicity interpretation (Krawitz and Haack, 2023). Genomic data serves as the foundation for diagnosis. For instance, patients with Duchenne muscular dystrophy (DMD) often have large deletions or frameshift mutations in the *DMD* gene. Collecting such information is helpful for diagnosis and the design of gene therapy. We named variations using the HGVS standard and linked them to genomic coordinates (such as GRCh38) to facilitate cross-study comparisons (Denton et al., 2021).

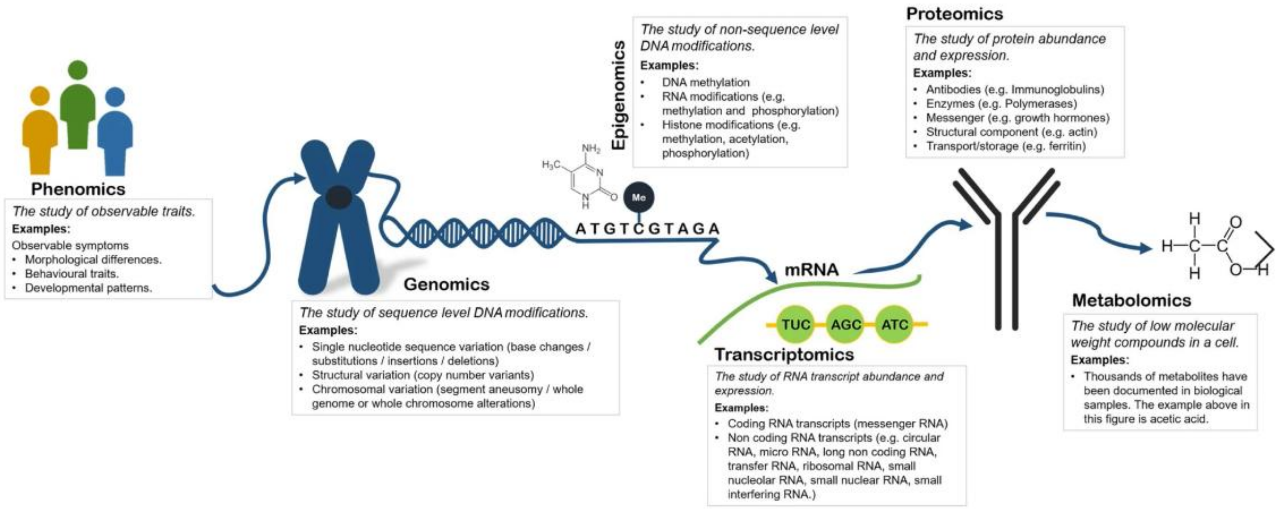
Phenomics
The study of observable traits.
Examples:
Observable symptoms
• Morphological differences.
• Behavioural traits.
• Developmental patterns.
Genomics
The study of sequence level DNA modifications.
Examples:
• Single nucleotide sequence variation (base changes / substitutions / insertions / deletions)
• Structural variation (copy number variants)
• Chromosomal variation (segment aneusomy / whole genome or whole chromosome alterations)
Epigenomics
The study of non-sequence level DNA modifications.
Examples:
• DNA methylation
• RNA modifications (e.g. methylation and phosphorylation)
• Histone modifications (e.g. methylation, acetylation, phosphorylation)
NH2
H3C
N
O
Me
ATGTCGTAGA
mRNA
TUC
AGC
ATC
Transcriptomics
The study of RNA transcript abundance and expression.
Examples:
• Coding RNA transcripts (messenger RNA)
• Non coding RNA transcripts (e.g. circular RNA, micro RNA, long non coding RNA, transfer RNA, ribosomal RNA, small nucleolar RNA, small nuclear RNA, small interfering RNA.)
Proteomics
The study of protein abundance and expression.
Examples:
• Antibodies (e.g. Immunoglobulins)
• Enzymes (e.g. Polymerases)
• Messenger (e.g. growth hormones)
• Structural component (e.g. actin)
• Transport/storage (e.g. ferritin)
H
H—C—C
O
O—H
H
Metabolomics
The study of low molecular weight compounds in a cell.
Examples:
• Thousands of metabolites have been documented in biological samples. The example above in this figure is acetic acid.

Figure 1 The diagram emphasises the potential of studies which, following careful phenotyping at study conception, utilise integrated multi-omic analysis to consider multiple components in the journey from DNA to expression (Adopted from Kerr et al., 2020)

Transcriptomics data mainly come from the results of RNA sequencing (RNA-SEq). The transcriptome reflects the activity level of genes in cells and is of great value for understanding the molecular mechanisms of diseases and discovering diagnostic markers. For instance, in the muscle tissues of DMD patients, a large number of genes related to inflammation and fibrosis are abnormally expressed (Lembo et al., 2024). This database will include the transcriptional expression profiles of patient tissues/cells, which may be stored in standardized forms such as FPKM and TPM. For key genes, we also store quantitative PCR verification data. If a patient has multiple samplings (such as before and after treatment), we classify their transcriptome by time or condition to support dynamic analysis.

Proteomics data refers to information on protein expression levels, post-translational modifications, and protein-protein interactions determined by methods such as mass spectrometry. Proteins are functional performers, and their states are often more directly related to phenotypes. Many diagnostic markers for rare diseases are proteins. For instance, amyotrophic lateral sclerosis has serum neurofilament light chain protein indicators, etc. Our database will integrate quantitative protein data generated from literature and experiments, such as a list of differential protein expression between patients with a certain disease and controls. In addition, it is planned to incorporate protein-protein interaction networks to demonstrate the position of pathogenic proteins in cellular pathways. For instance, the role of DMD protein (dystrophin) in the muscle cell membrane complex and the series of downstream protein changes triggered by its absence can be revealed through integrated proteome data. Proteomic data can provide a bridge for multi-omics associations: sometimes gene mutations do not change the mRNA level but affect protein stability, which can be reflected in the proteome (Lochmüller et al., 2018).

3.2 Ethical considerations for rare disease data and data sharing framework

Rare disease data usually involves patients' genetic and health information. Ethical, legal and social factors must be fully considered in data sharing and use. We have formulated a series of policies and measures in the database construction to balance promoting scientific research and protecting patients' rights and interests. For the patient data we obtain from our partner hospitals/institutions, informed consent forms were signed by the patients themselves or their legal representatives at the time of collection, explicitly agreeing to the use of the data for scientific research and allowing it to be shared with the database anonymously. The informed consent form complies with local regulations and explains the purpose of the data, potential risks, and the rights of the subjects, including their right to withdraw the data at any time (Gainotti et al., 2016). We only use patient data with explicit consent and respect the patients' right to make decisions. This is particularly important in rare diseases, as the patient population is small and they are more likely to be identified due to data breaches. We fully inform and obtain consent to ensure ethical legitimacy (Koromina et al., 2021).

All patient data entering the database undergo strict de-identification processing before being stored. Direct identity information such as name, ID number and contact details will never be stored. Information that may indirectly identify an identity (such as rare geographical locations, biometric photos, etc.) is also not included in the database. For genomic sequence data, we follow the commonly used methods internationally, encoding and storing individual data, and using random ids to replace real identities. When users obtain data, we also take measures to prevent re-identification caused by cross-comparison. For example, limit the amount of fine-grained data returned for each query to prevent malicious users from piecing together identities through step-by-step queries. For extremely rare data (such as data of only individual patients worldwide), we may only provide aggregated information or must access it in a specially controlled environment to minimize the risk of exposure (Hansson et al., 2016; Takashima et al., 2018).

4 Data Integration and Computing Methods

4.1 Multi-omics data normalization and coordination techniques

Before integrating and analyzing multi-level data such as genomic, transcriptomic and proteomic data, it is necessary to standardize and coordinate the data of different omics to make them comparable and compatible. Since multi-omics data often come from different batches, platforms, and even different laboratories, batch effects and system biases are inevitable. If not corrected, batch differences will be mistaken for biological differences during integrated analysis. For this purpose, we conducted batch correction on the omics data before integration. For the transcription and protein quantification data, we normalized them using the Combat algorithm, which adjusts the mean and variance of different batches through a linear model to make the data distribution more consistent. In practice, Combat significantly reduced the differences in the distribution of expression values among different studies, making the comparison between the patient group and the control group more reliable.

For metabolite data, we used normalization based on quality markers to proportionally adjust the data of each batch according to the common internal standard (Ali et al., 2025). For the methylated chip data, methods such as COMBAT-Seq were used for correction. In addition, we conduct QC after calibration to ensure that the calibration does not overly weaken the real biological signal. For example, principal component analysis (PCA) is used to check the clustering between groups before and after correction. If it is found that the correction affects the differences between groups, we will retain certain batch variables or reevaluate using a mixed-effects model. In conclusion, batch effect correction ensures that multi-source data can be compared on the same scale, laying the foundation for subsequent integration.

To reduce the complexity of the analysis, we carried out appropriate dimensionality reduction processing on the data before integration. Common methods include Principal Component Analysis (PCA) and automatic feature selection. For instance, for high-dimensional transcriptional data, before integrating it into the network, we first use PCA to extract the first few principal components, capture the major variations, and discard the components with a low proportion to reduce noise interference.

4.2 Data integration algorithm

After completing the data preprocessing and normalization, we apply multiple algorithms to conduct in-depth integrated analysis of multi-omics data. We constructed multiple interconnected network layers: gene regulation/co-expression networks, protein-protein interaction networks, metabolite pathway networks, etc., and then mapped patient data onto these networks to identify active modules and key nodes of the networks. Specifically, we use the Similarity Network Fusion (SNF) algorithm to fuse the sample similarity networks of different omics. SNF first calculates the similarity network between samples based on gene expression, protein, and metabolic data respectively, and then iteratively fuses them to obtain the comprehensive similarity. This fusion network is used for patient classification and prediction, and through community testing, subgroups of patients with similar multi-omics characteristics are identified. A review indicates that network fusion has achieved success in scenarios such as drug discovery. We applied it to patient clustering and initially discovered some interesting subtype phenomena.

Bayesian methods can naturally integrate multi-source uncertain information through probabilistic graphical models. They construct directed acyclic graphs of genes and phenotypes through Bayesian networks, with edges representing the probabilities of causal relationships. By learning from patient data, we have obtained some optimal network structures. For instance, for a set of genetic metabolic disease data, our model automatically learns the causal chain of a certain enzyme gene -> key metabolites -> corresponding phenotypes, verifying the existing knowledge and suggesting new relationships (Ibrahim, 2023).

4.3 Function annotation, feature extraction and path enrichment analysis

One of the key goals of multi-omics integrated analysis is to transform massive amounts of data into biological knowledge that can be explained by humans. This requires functional annotation and exploration of the biological significance of the discovered patterns. Therefore, in the analysis process, we incorporated steps such as functional annotation, feature extraction, and path/network enrichment to assist in result interpretation and new discoveries.

When the integrated analysis yields a list of important genes or molecules, we immediately annotate and explain their biological functions. For example, if the network algorithm identifies a group of Gene nodes highly associated with diseases, we invoke the gene ontology (GO) database to annotate the known functions of these genes (such as the biological processes they are involved in, molecular functions, etc.). If most of these genes are concentrated in a certain process (such as muscle structure development), this suggests that the disease mechanism is related to this process (Bottini et al., 2022). For instance, the Bayesian model selects several candidate metabolites. We query databases such as HMDB to obtain their biochemical pathways and disease-associated annotations, in order to determine which ones are worth in-depth verification.

When obtaining a list of significantly different genes, proteins or metabolites, we often use pathway enrichment to condense biological topics. The specific approach is to input the list of differences into the KEGG pathway, Reactome pathway or GO biological process database, and use hypergeometric tests or gene set enrichment analysis (GSEA) to calculate which pathways significantly enrich the differentially expressed molecules. For metabolites, we conduct metabolic pathway enrichment, such as using the algorithm in MetaboAnalyst. Abnormalities such as "purine metabolism" and "arginine and proline metabolism" were found. Through pathway enrichment, scattered molecular lists are elevated to the level of pathway networks, facilitating the understanding of their functional significance (Paczkowska et al., 2020).

5 Case Study

5.1 Overview of DMD and available multi-omics datasets

Duchenne muscular dystrophy (DMD) is an X-linked recessive disorder caused by mutations in the DMD gene, leading to absence of dystrophin protein and progressive muscle degeneration. It is one of the most prevalent genetic muscle diseases in childhood. Currently, standard diagnostics rely on genetic tests and muscle biopsy. In our integrated database, we identified several DMD-relevant datasets: genomic variants from patient exomes (including copy-number deletions in DMD), and transcriptomic profiles from muscle biopsies of patients and controls. For example, we imported RNA-seq data from recent single-nucleus studies of DMD muscle tissue. Additional resources include proteomic profiles and limited metabolomic data from DMD cohorts (Dowling et al., 2024). By aggregating these, we have a comprehensive multi-omics view of DMD.

5.2 Implementation of integration pipeline and visualization tools

We ran a custom pipeline on the DMD case: first, we normalized gene expression and identified differentially expressed genes (DEGs) between DMD and control muscle. Variants in the DMD gene and other muscle-related genes were annotated. Using our multi-omics integration framework, we constructed a network linking mutated genes to downstream expression changes via known muscle pathway interactions (Lu et al., 2019). We also applied a Bayesian factor model to jointly analyze the genomic and transcriptomic data, revealing latent factors correlated with disease severity. Results are visualized in the web portal: for instance, a genome browser highlights the large deletions in DMD for each patient, while an interactive heatmap shows upregulated muscle regeneration genes in DMD samples (Schneegans et al., 2023). We also implemented Circos plots to display connections between genetic loci and transcriptional changes. These tools allow a user to explore how a variant (e.g., an exon deletion in DMD) propagates to altered gene networks and pathways.

5.3 Insights gained: molecular biomarkers and potential therapeutic targets

The integrated analysis yielded new insights into DMD pathology. Consistent with prior studies, we observed upregulation of genes involved in muscle regeneration (e.g., *MYOG*, *PAX7*) and fibrosis (e.g., collagen genes) in DMD patients. Our model identified fibro-adipogenic progenitor (FAP) cells as key regulators: their aberrant signaling (through PDGF and TGF-β pathways) likely drives excess extracellular matrix deposition (Figure 2). Importantly, by cross-referencing DEGs with drug databases, we propose candidate molecular targets. For example, the gene *SPP1* (osteopontin) was highly overexpressed and is known to modulate muscle inflammation; it emerges as a potential drug target or biomarker (Vera et al., 2022). Also, our protein network analysis suggests that upregulated myogenesis genes (e.g., *MYH8*, *ACTA1*) could serve as blood biomarkers for disease progression. These findings align with and extend published DMD research. This case study demonstrates the platform’s ability to generate clinically relevant hypotheses from integrated omics data.

6 Conclusion

This study constructed an integrated multi-omics database for rare diseases and conducted case analyses of Duchenne muscular dystrophy and others based on this platform. We have integrated multi-level data such as genomics, transcriptomics, proteomics, metabolomics and phenotypes in accordance with a unified architecture, and developed standardized data models and interoperability frameworks. Through strict quality control and standard annotation, the database ensures the compatibility of data from different sources. Users can access it through a friendly Web interface and API for interactive analysis and visualization. This provides rare disease researchers with a pioneering tool, which is different from the previous single data resource and realizes the organic integration of knowledge and data.

We have integrated multiple algorithms such as network analysis, Bayesian models, and machine learning into the platform to mine the biological significance of multi-omics data. These methods complement each other and reveal the key pathways, molecular modules and modifying factors of rare diseases from different perspectives. We utilize methods such as path enrichment and knowledge graphs to transform complex results into interpretable knowledge. This analytical system can be extended and applied to a variety of rare diseases, accelerating the process from data to discovery.

This study demonstrates that by integrating multi-omics big data, biomarkers and drug targets of rare diseases can be systematically identified, thereby guiding patient stratification and individualized intervention. This has overturned the traditional model that relies on experience and fragmented research, pioneering a data-driven and AI-assisted research paradigm for rare diseases. Our platform has demonstrated potential in areas such as auxiliary diagnosis, drug reuse, and prognosis prediction, and will drive the development of precision medicine for rare diseases.

Although there are still challenges ahead, we firmly believe that as long as we adhere to the concepts of open cooperation, technological innovation and patient-centeredness, the platform will surely continue to grow and thrive. From a current academic research tool, it has grown into one of the key cornerstones supporting precision medicine for rare diseases and the development of new drugs. We look forward to seeing the actual changes it has brought about when we review this research in the near future: more rare diseases have been cracked and more patients have been cured, and our platform is an important driver of this. This will be the best reward for our work.

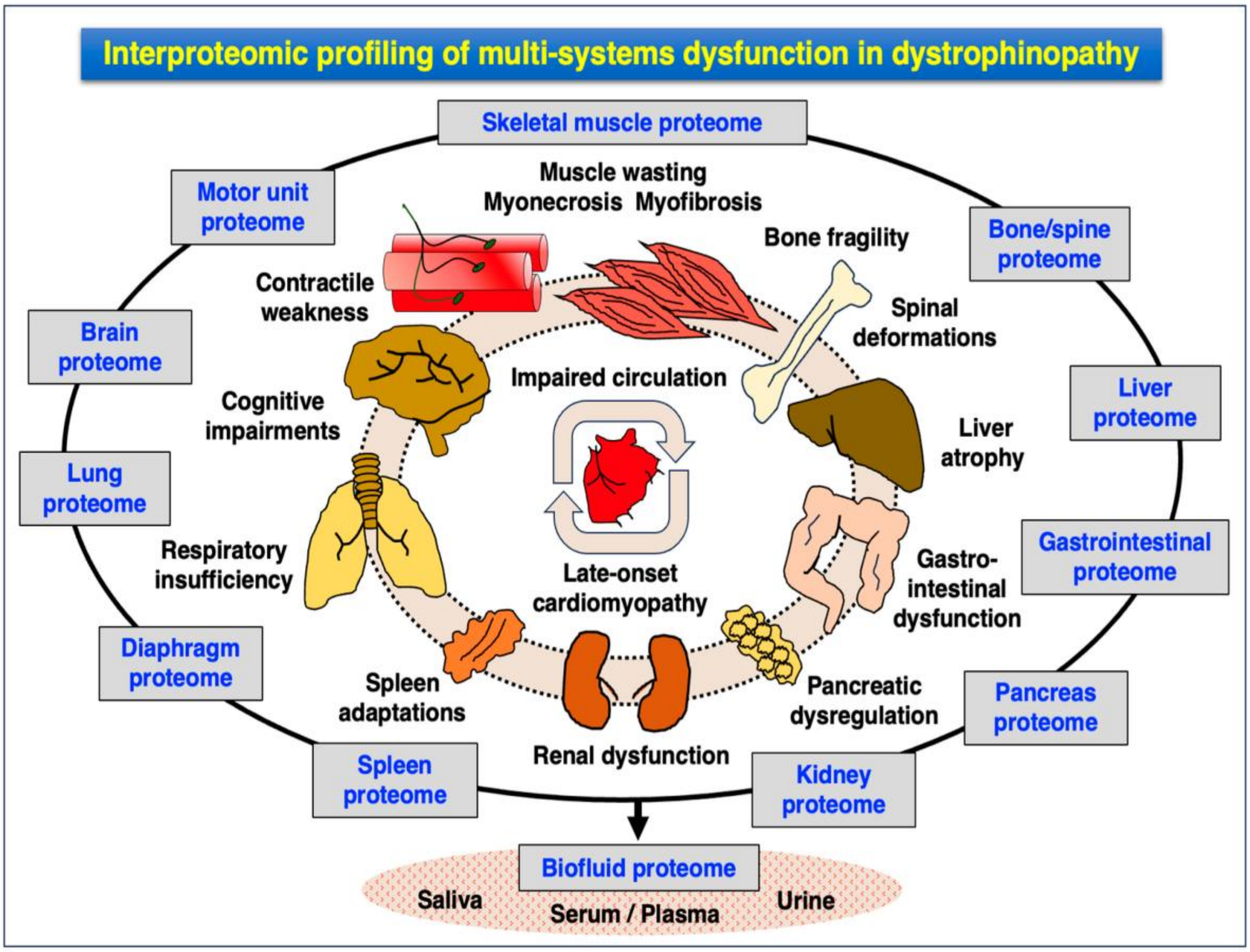

Figure 2 The pathoproteomic profile of multi-system changes in Duchenne muscular dystrophy. The diagram outlines the complexity of body-wide alterations due to dystrophin deficiency and illustrates how the systematic application of a comprehensive interproteomic profiling approach could help us better understand the multi-system dysfunction in dystrophinopathy (Adopted from Dowling et al., 2024)

Acknowledgments

The authors extend sincere thanks to two anonymous peer reviewers for their invaluable feedback on the manuscript.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Ali S., Li Q., and Agrawal P., 2025, Implementation of multi-omics in diagnosis of pediatric rare diseases, Pediatric Research, 97(4): 1337-1344.
https://doi.org/10.1038/s41390-024-03728-w

Braconi D., Bernardini G., Spiga O., and Santucci A., 2021, Leveraging proteomics in orphan disease research: pitfalls and potential, Expert Review of Proteomics, 18(4): 315-327.
https://doi.org/10.1080/14789450.2025.2468300

Casas-Alba D., Hoenicka J., Vilanova-Adell A., Vega-Hanna L., Pijuan J., and Palau F., 2022, Diagnostic strategies in patients with undiagnosed and rare diseases, Journal of Translational Genetics and Genomics, 6(3): 322-332.
https://doi.org/10.20517/jtgg.2022.03

Denton N., Molloy M., Charleston S., Lipset C.H., Hirsch J., Mulberg A., Howard P., and Marsh E., 2021, Data silos are undermining drug development and failing rare disease patients, Orphanet Journal of Rare Diseases, 16(1): 161.
https://doi.org/10.1186/s13023-021-01806-4

Dowling P., Trollet C., Negroni E., Swandulla D., and Ohlendieck K., 2024, How can proteomics help to elucidate the pathophysiological crosstalk in muscular dystrophy? Proteomes, 12(1): 4.
https://doi.org/10.3390/proteomes12010004

Gainotti S., Turner C., Woods S., Kole A., McCormack P., Lochmüller H., Riess O., Straub V., Posada M., Taruscio D., and Mascalzoni D., 2016, Improving the informed consent process in international collaborative rare disease research, European Journal of Human Genetics, 24(9): 1248-1254.
https://doi.org/10.1038/ejhg.2016.2

Hansson M., Lochmüller H., Riess O., Schaefer F., Orth M., Rubinstein Y., Molster C., Dawkins H., Taruscio D., Posada M., and Woods S., 2016, The risk of re-identification versus the need to identify individuals in rare disease research, European Journal of Human Genetics, 24(11): 1553-1558.
https://doi.org/10.1038/ejhg.2016.52

Hesterlee S., 2021, Optimizing rare disease registries and natural history studies, In: Rare Disease Drug Development: Clinical, Scientific, Patient, and Caregiver Perspectives, Springer International Publishing, pp.109-125.
https://doi.org/10.1007/978-3-030-78605-2_8

Ibrahim N., 2023, Navigating the complexity of rare diseases: challenges, innovations, and future directions, Global Journal of Medical Therapeutics, 5(4): 12-22.
https://doi.org/10.46982/gjmt.2023.108

Jia J., An Z., Ming Y., Guo Y., Li W., Liang Y., Guo D., Li X., Tai J., Chen G., Jin Y., Liu Z., Ni M., and Shi T., 2018, eRAM: encyclopedia of rare disease annotations for precision medicine, Nucleic Acids Research, 46(D1): D937-D943.
https://doi.org/10.1093/nar/gkx1062

Kerr K., McAneney H., Smyth L., Bailie C., McKee S., and McKnight A., 2020, A scoping review and proposed workflow for multi-omic rare disease research, Orphanet Journal of Rare Diseases, 15(1): 107.
https://doi.org/10.1186/s13023-020-01376-x

Koromina M., Fanaras V., Baynam G., Mitropoulou C., and Patrinos G., 2021, Ethics and equity in rare disease research and healthcare, Personalized Medicine, 18(4): 407-416.
https://doi.org/10.2217/pme-2020-0144

Krawitz P., and Haack T., 2023, Editorial-Diagnostic genome sequencing in rare disorders, Medizinische Genetik, 35(2): 89.
https://doi.org/10.1515/medgen-2023-2029

Lembo S., Barra P., Dash S., and Di Biasi L., 2024, Challenges and opportunities of symbiotic AI in rare disease diagnosis, In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp.6820-6825.
https://doi.org/10.1109/BIBM62325.2024.10822548

Liu J., Barrett J., Leonardi E.T., Lee L., Roychoudhury S., Chen Y., and Trifillis P., 2022, Natural history and real-world data in rare diseases: applications, limitations, and future perspectives, The Journal of Clinical Pharmacology, 62: S38-S55.
https://doi.org/10.1002/jcph.2134

Lochmüller H., Badowska D., Thompson R., Knoers N., Aartsma-Rus A., Gut I., Wood L., Harmuth T., Durudas A., Graessner H., Schaefer F., Riess O., RD-Connect consortium, NeurOmics consortium and EURenOmics Consortium, 2018, RD-Connect, NeurOmics and EURenOmics: collaborative European initiative for rare diseases, European Journal of Human Genetics, 26(6): 778-785.
https://doi.org/10.1038/s41431-018-0115-5

Lu Y., Chang Y., Hoffman E., Yu G., Herrington D., Clarke R., Wu C.T., Chen L., and Wang Y., 2019, Integrated identification of disease-specific pathways using multi-omics data, bioRxiv, 2019: 666065.
https://doi.org/10.1101/666065

Mitani A., and Haneuse S., 2020, Small data challenges of studying rare diseases, JAMA Network Open, 3(3): e201965.
https://doi.org/10.1001/jamanetworkopen.2020.1965

Olexiouk V., 2023, Challenges and opportunities with multi-omics integration in precision medicine, In: The 2nd International Conference on Systems Medicine AI & Drug Repurposing, REPO4EU, 2023:2.
https://doi.org/10.58647/REXPO.23030

Paczkowska M., Barenboim J., Sintupisut N., Fox N.S., Zhu H., Abd-Rabbo D., Mee M., Boutros P., PCAWG Drivers and Functional Interpretation Working Group, Reimand J., and PCAWG Consortium, 2020, Integrative pathway enrichment analysis of multivariate omics data, Nature Communications, 11(1): 735.

Pahelkar A., Sharma D., Vohra P., and Sawant S., 2024, Leveraging multi-omics approaches and advanced technologies for hemoglobin H disease, European Journal of Haematology, 113(6): 738-744.
https://doi.org/10.1111/ejh.14319

Schneegans E., Fancy N., Thomas M., Willumsen N., Matthews P., and Jackson J., 2023, Omix: A multi-omics integration pipeline, bioRxiv, 30: 555486.
https://doi.org/10.1101/2023.08.30.555486

Takashima K., Maru Y., Mori S., Mano H., Noda T., and Muto K., 2018, Ethical concerns on sharing genomic data including patients' family members, BMC Medical Ethics, 19(1): 61.
https://doi.org/10.1186/s12910-018-0310-5

Vera C.D., Zhang A., Pang P., and Wu J.C., 2022, Treating Duchenne muscular dystrophy: the promise of stem cells, artificial intelligence, and multi-omics, Frontiers in Cardiovascular Medicine, 9: 851491.
https://doi.org/10.3389/fcvm.2022.851491

Zaghlool S., and Attallah O., 2022, A review of deep learning methods for multi-omics integration in precision medicine, In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp.2208-2215. https://doi.org/10.1109/BIBM55620.2022.9995099

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.