Computational Biology and Bioinformatics
2024, Vol. 12, No. 1, pp. 1-11
https://doi.org/10.11648/j.cbb.20241201.11

Review Article

From Data to Diagnosis Exploring AWS Cloud Solutions in Multi-Omics Breast Cancer Biomarker Research

Gnanam Subramanian*, Kavitha Ramamoorthy

Department of Biotechnology, Periyar University, Salem, India

Abstract

Breast cancer presents a profound global health challenge, compounded by unique intricacies within the Indian demographic, necessitating bespoke research methodologies. This abstract delineates the profound impact of Amazon Web Services (AWS) Cloud Solutions on advancing multi-omics breast cancer biomarker research, with a particular focus on Indian patient cohorts. It initiates with an exposition of the inherent challenges encountered during the transition from raw data acquisition to clinical diagnosis, emphasizing the indispensable role of cloud-based infrastructures in expediting this complex trajectory. Harnessing the comprehensive capabilities of AWS, this study elucidates how cloud solutions facilitate the seamless integration and analysis of multifaceted omics datasets, encompassing genomics, transcriptomics, proteomics, and metabolomics. Central to this endeavor is a meticulous exploration of region-specific molecular markers germane to breast cancer within the Indian populace, illuminating their diagnostic and therapeutic ramifications. By capitalizing on AWS Cloud's scalability and computational acumen, this research underscores notable efficiency enhancements in processing voluminous datasets and distilling salient patterns therein. Furthermore, the discourse extends to the broader ramifications of these technological advancements within the precision medicine landscape, emphasizing the potential for tailored therapeutic interventions. This research heralds a paradigmatic shift in the application of cloud-based infrastructures to unravel the intricate tapestry of breast cancer, transcending geographical confines. Through its provision of insights poised to augment diagnostic precision and therapeutic efficacy on a global scale, this study marks a seminal stride towards fully harnessing the potential of precision oncology in combating breast malignancies.

Keywords

Biomarker, Breast Cancer, Multi-Omics, Amazon Web Services

1. Introduction

Breast cancer stands as a formidable global health challenge, accounting for a significant burden of morbidity and mortality worldwide. Within this broader context, the Indian population presents unique epidemiological, genetic, and socio-cultural nuances that demand tailored research approaches. Amidst the evolving landscape of oncology, the integration of multi-omics data has emerged as a promising avenue for elucidating the intricate molecular mechanisms underpinning breast cancer pathogenesis and progression. Against this backdrop, this paper endeavors to explore the transformative potential of Amazon Web Services (AWS) Cloud Solutions in advancing multi-omics breast cancer biomarker research, with a specific focus on Indian patient cohorts. By leveraging the vast computational resources and

*Corresponding author: gnanams@periyaruniversity.ac.in (Gnanam Subramanian)

Received: 27 March 2024; **Accepted:** 12 April 2024; **Published:** 15 August 2024

analytical tools offered by AWS, this study seeks to streamline the integration and analysis of diverse omics datasets, encompassing genomics, transcriptomics, proteomics, and metabolomics. Through a meticulous examination of region-specific molecular signatures, the research aims to decipher the diagnostic and therapeutic implications thereof, thereby paving the way for personalized and targeted therapeutic interventions. The impetus for this research arises from the recognition of the burgeoning role played by cloud-based technologies in accelerating scientific discovery and innovation. However, while considerable strides have been made in leveraging cloud solutions for biomedical research, their application within the context of Indian breast cancer genomics remains relatively underexplored. Consequently, this study not only fills a critical gap in the existing literature but also holds profound implications for advancing our understanding of breast cancer biology in diverse populations. Furthermore, the significance of this research extends beyond academic realms, with tangible implications for clinical practice and public health policy. By elucidating region-specific molecular markers and their associations with disease outcomes, this study has the potential to inform the development of more effective screening modalities, prognostic tools, and therapeutic strategies tailored to the Indian population. Moreover, the insights gleaned from this research may contribute to the broader discourse on precision oncology, underscoring the imperative of accounting for ethnic and geographical diversity in cancer research and treatment paradigms. Thus, this paper not only contributes to the academic discourse but also holds promise for catalyzing real-world improvements in breast cancer care and outcomes [1, 2].

1.1. Breast Cancer Overview

Breast cancer originates in the milk ducts or lobules of the breast, with early forms often confined to these regions. However, as cancer cells invade nearby tissues, tumors may form, leading to lumps or thickening [3]. The gravity of the disease intensifies when cancer cells metastasize, spreading to lymph nodes or other organs, resulting in potentially fatal consequences. Treatment strategies encompass a combination of surgery, radiation therapy, and medications, tailored to the individual and the cancer type [4].

1.2. Scope of the Problem

Breast cancer is a global health concern, with 2.3 million women diagnosed and 685,000 deaths reported in 2020. It is the most prevalent cancer worldwide, affecting women of all ages. Mortality rates have historically improved since the 1990s, thanks to early detection programs and comprehensive treatments. Risk factors include age, gender, family history, and certain genetic mutations, with early detection proving essential for effective treatment [4-6].

2. Signs and Symptoms

Breast cancer symptoms vary but commonly include painless lumps, changes in breast size or appearance, skin abnormalities, and nipple changes. Early-stage cancer may be asymptomatic, underscoring the importance of regular screenings and prompt medical attention for any abnormal breast changes. Timely intervention significantly improves the chances of successful treatment [7].

2.1. Treatment

Treatment approaches depend on cancer subtype and stage, encompassing surgery, radiation therapy, and medications. Surgical options range from lumpectomy to mastectomy, with sentinel node biopsy replacing complete axillary dissection for lymph node assessment. Medications, including hormonal therapies and chemotherapy, target specific cancer characteristics. Radiation therapy plays a crucial role in preventing recurrence and managing advanced stages [7].

2.2. Global Impact

Breast cancer mortality has seen a 40% reduction in high-income countries from the 1980s to 2020. Successful strategies involve strengthening health systems, emphasizing early detection, timely diagnosis, and comprehensive cancer management. Breast cancer serves as an "index" disease, providing pathways applicable to the management of other cancers [8].

2.3. WHO Response

The World Health Organization's Global Breast Cancer Initiative aims to reduce global breast cancer mortality by 2.5% annually, preventing 2.5 million deaths by 2040. This involves health promotion, timely diagnosis, and comprehensive management. Public and health worker education is crucial for early detection, with rapid diagnosis linked to specialized cancer care [9].

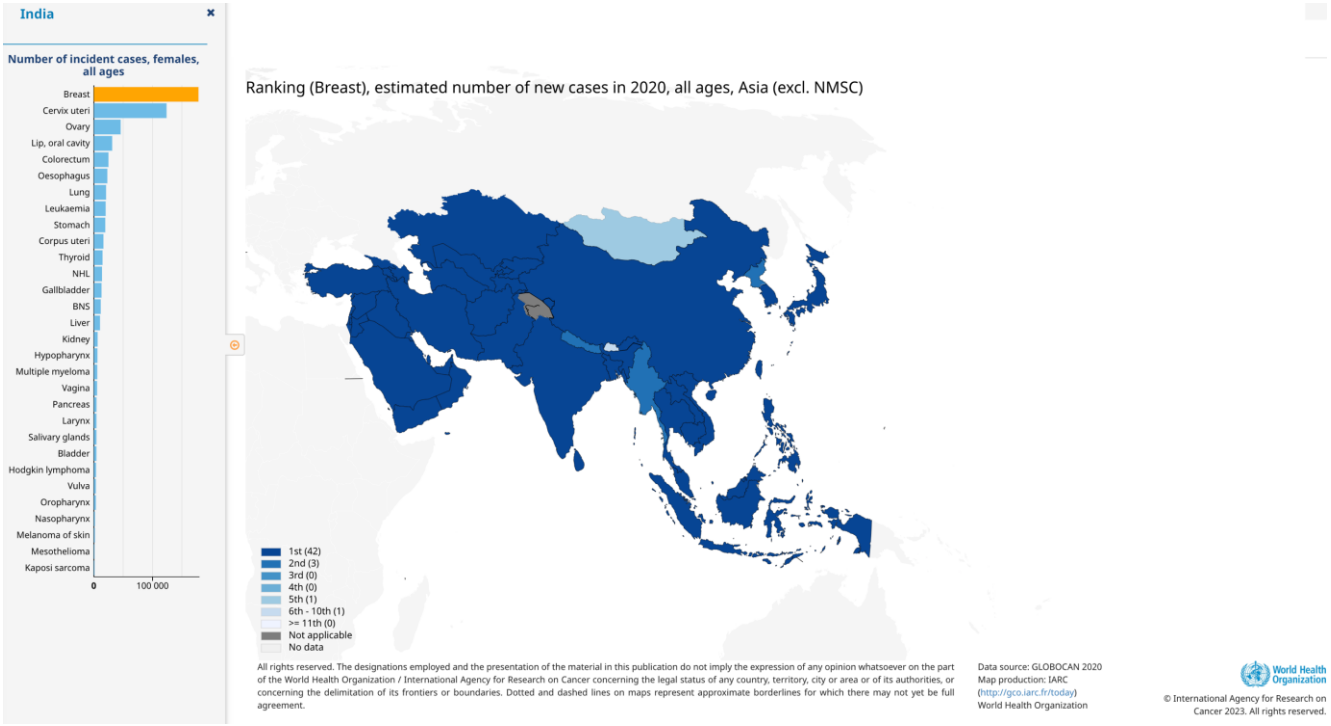

Figure 1. *Estimated Number of cases in 2020, Asia.*

3. Multiomics Overview

Multiomics represents a paradigm shift in biomedical research by concurrently examining genomics, epigenomics, transcriptomics, proteomics, and metabolomics. This integrated approach aims to unravel the intricate interplay between various molecular layers, offering a more comprehensive perspective on biological processes and disease mechanisms [10-14].

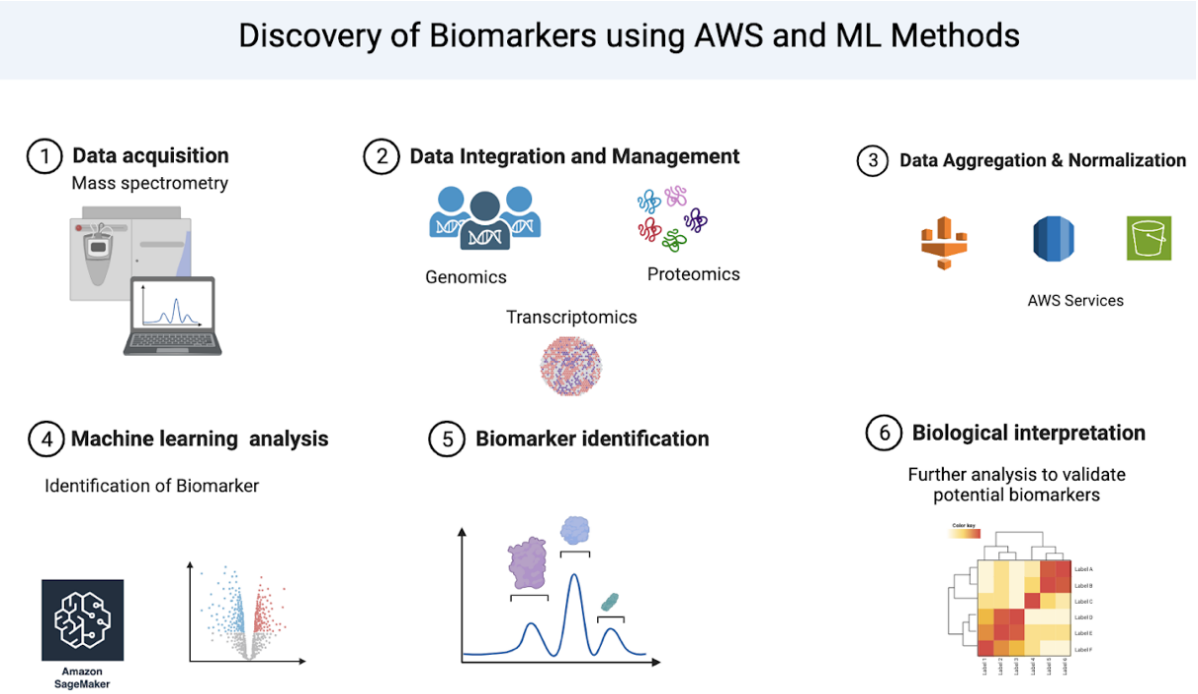

Figure 2. *Discovery of Biomarkers Using AWS and Machine Learning.*

3.1. Genomics in Multiomics

Genomics, the foundational layer of multiomics, investigates whole genome sequences and DNA variations. Next-generation sequencing (NGS) technologies have propelled genomic analysis, enabling rapid, cost-effective sequencing. The integration of genomics in multiomics provides a foundation for understanding genetic contributions to complex biological phenomena, disease etiology, and personalized medicine [15-16].

Table 1. *Summary of Genomics data sources [15-17].*

Data source	Aspect	Description
NCBI GenBank	Genomic Sequences	A comprehensive repository of publicly available nucleotide sequences, providing a vast collection for genomic studies.
Ensembl	Genomic Annotations	Offers genome annotations, gene sequences, and functional information, facilitating the interpretation of genomic data.
1000 Genomes Project	Genomic Variation	A global initiative cataloging human genetic variation, aiding in understanding population diversity and disease genetics.
UCSC Genome Browser	Genomic Visualization	A user-friendly platform for visualizing and exploring genomic data, supporting researchers in data interpretation.
dbSNP	Single Nucleotide Polymorphisms (SNPs)	A database of common and rare SNPs, assisting in the identification of genetic variations across populations.
ClinVar	Clinical Genomics	Aggregates information on clinically relevant genomic variations, facilitating the interpretation of genetic variants in a clinical context.
Genomic Data Commons (GDC)	Cancer Genomics	A resource providing access to genomic and clinical data from cancer studies, promoting collaborative research in cancer genomics.
Encode	Functional Genomics	Focuses on functional elements in the genome, providing data on gene regulation, chromatin structure, and epigenetic modifications.
RefSeq	Reference Genomes	A curated collection of reference sequences, serving as a benchmark for genomic analysis and annotation.
HapMap	Population Genetics	A project mapping common genetic variations in human populations, aiding in the understanding of population genetics and disease susceptibility.

3.2. Epigenomics in Multiomics

Epigenomics explores the chemical modifications of DNA and histones, influencing gene regulation. Techniques such as bisulfite sequencing and ChIP-seq offer precise mapping of genome-wide methylation patterns and chromatin modifications. In the multiomics context, epigenomics contributes to a nuanced understanding of how epigenetic factors interact with genomic information, shaping cellular phenotypes and disease states [18].

Table 2. *Summary of Epigenomics Data Sources [19].*

Data sources	Aspect	Descriptions
ENCODE (Encyclopedia of DNA Elements)	Genomic DNA Methylation	Comprehensive resource providing maps of DNA methylation patterns, aiding in understanding epigenetic regulation at a genome-wide level.
Roadmap Epigenomics Project	Histone Modifications	Large-scale initiative mapping histone modifications across various cell types and tissues, facilitating the exploration of

Data sources	Aspect	Descriptions
		chromatin dynamics and gene regulation.
GEO (Gene Expression Omnibus)	Epigenomic Data Sets	A public repository hosting a diverse range of epigenomic datasets, including ChIP-seq, bisulfite sequencing, and other assays, fostering data sharing and collaboration.
Blueprint Epigenome Project	DNA Methylation and Histone Modifications	Focuses on profiling epigenetic marks across diverse human cell types, contributing valuable insights into the regulatory landscape of the human genome.
UCSC Genome Browser	Epigenetic Annotations	Integrates epigenomic data with genomic annotations, allowing users to visualize and analyze epigenetic features in the context of the entire genome.
NIH Epigenomics Data Analysis and Coordination Center (EDACC)	Data Coordination	Central hub for storing and distributing epigenomic data generated by various consortia, ensuring standardized formats and accessibility for researchers.
BLUEPRINT Data Portal	DNA Methylation, Histone Modifications	Provides access to datasets generated by the BLUEPRINT project, offering insights into the epigenetic variation in different hematopoietic cell types.
Cistrome Data Browser	Transcription Factor Binding	A platform hosting a collection of ChIP-seq datasets, facilitating the exploration of transcription factor binding sites and their regulatory roles.
Epigenome Browser	Interactive Visualization	An online tool for exploring and visualizing epigenomic data, enabling researchers to interactively analyze and interpret epigenetic information.
IHEC Data Portal	International Epigenome Consortium (IHEC) Data	Hosts data from the IHEC, promoting global collaboration and standardization in epigenomic data generation and analysis.

3.3. Transcriptomics in Multiomics

Transcriptomics, capturing the complete set of RNA transcripts, plays a pivotal role in multiomics. RNA sequencing (RNA-Seq) technology provides insights into gene expression patterns, alternative splicing, and novel transcript discovery. Integrating transcriptomics into multiomics elucidates the dynamic relationship between genomic information and the functional output of the cell [20].

Table 3. *Summary of Transcriptomics Data Sources [20].*

Data sources	Aspect	Descriptions
Microarray	Technology	Utilizes hybridization-based methods for transcript quantification, providing a snapshot of gene expression in a sample.
RNA-Seq	Technology	Employs high-throughput sequencing to quantify RNA transcripts, allowing for accurate measurement and detection of novel transcripts.
SAGE	Technology	Serial Analysis of Gene Expression provides a quantitative assessment of gene expression patterns by sequencing short tags.
MPSS	Technology	Massively Parallel Signature Sequencing enables digital quantification of transcripts, providing a comprehensive view of the transcriptome.
RNA-ISH	Technology	RNA In Situ Hybridization allows for the visualization and localization of specific RNA transcripts within cells and tissues.

Data sources	Aspect	Descriptions
Databases	Resource	Publicly available databases such as NCBI Gene Expression Omnibus (GEO) and European Bioinformatics Institute (EBI) house vast transcriptomic datasets for diverse biological samples.
Single-Cell RNA-Seq	Technology	Facilitates transcriptomic analysis at the single-cell level, enabling the study of cellular heterogeneity and identification of rare cell types.
Long-Read Sequencing	Technology	Utilizes sequencing technologies with extended read lengths, providing a more comprehensive view of complex transcript structures and isoforms.
Pathway Analysis Tools	Analysis Tool	Various tools like Gene Set Enrichment Analysis (GSEA) and Ingenuity Pathway Analysis (IPA) interpret transcriptomic data in the context of biological pathways, aiding in functional interpretation.
Differential Expression Analysis Tools	Analysis Tool	Tools such as DESeq2 and edgeR identify genes showing significant expression changes between different experimental conditions, facilitating the discovery of key regulatory elements.
Data Integration Platforms	Analysis Tool	Platforms like Seurat and Scanpy integrate transcriptomic data with other omics layers, enhancing the understanding of the interplay between genes, proteins, and metabolites in complex biological systems.

3.4. Proteomics in Multiomics

Proteomics, quantifying protein identities and abundances, enhances multiomics by bridging the gap between genomic information and cellular function. Advances in mass spectrometry (MS) technologies enable the comprehensive analysis of the proteome. Multiomics applications of proteomics offer a deeper understanding of how proteins, as effectors of cellular processes, contribute to complex biological responses and disease phenotypes [21].

Table 4. *Summary of Proteomics Data Sources [21].*

Data sources	Aspect	Descriptions
MassIVE	Data Repository	MassIVE is a community resource developed by the National Center for Integrative Proteomics (NCIP) to promote the global, free exchange of mass spectrometry data.
ProteomeXchange	Data Exchange	ProteomeXchange serves as a comprehensive and centralized repository for proteomics data. It facilitates data sharing and retrieval across multiple proteomics repositories, promoting open data practices.
PeptideAtlas	Protein Identification	PeptideAtlas is a repository that houses high-quality mass spectrometry-based shotgun proteomics data, providing a valuable resource for protein identification and quantification.
Human Proteome Map	Protein Expression Atlas	The Human Proteome Map (HPM) project provides a comprehensive resource for understanding tissue-specific protein expression patterns, aiding in the exploration of the human proteome across various tissues.
PRIDE (PRoteomics IDEntifications)	Dataset Archive	PRIDE is a database for storing and disseminating mass spectrometry-based proteomics data. It allows researchers to submit, browse, and download proteomics datasets, fostering data accessibility and reuse.
MaxQB	Quantitative Proteomics	MaxQB is a database that specializes in quantitative information about the human proteome. It includes data on protein expression levels, modifications, and interactions, supporting research in quantitative proteomics.
MassBank	Mass Spectrometry Data	MassBank is an open-access database that focuses on mass spectrometry data for small molecules, including metabolites and peptides. It provides a platform for sharing and retrieving mass spectra information.

Data sources	Aspect	Descriptions
jPOST (Japan Proteome Standard Repository)	Proteome Standardization	jPOST is a repository dedicated to standardizing and archiving proteome data generated by Japanese researchers. It aims to enhance the quality and reproducibility of proteomics experiments through data sharing and standardization.
iProX (Integrated Proteome Resources)	Integrated Proteomics	iProX serves as an integrated platform for storing and sharing proteomics data along with related multiomics data. It supports the integration of proteomics with genomics, transcriptomics, and other molecular datasets.
PRIN (Proteomics Integrated)	Cross-Omics Integration	PRIN is a platform that integrates proteomics data with other omics data, enabling cross-omics analysis. It promotes a holistic understanding of biological systems by combining proteomics information with genomic, transcriptomic, and metabolomic data.

3.5. Metabolomics in Multiomics

Metabolomics, studying small molecules in the body, enriches multiomics by providing insights into metabolic pathways and cellular responses. Targeted and untargeted metabolomics, fluxomics, and metabolite imaging contribute to a holistic understanding of how metabolites, influenced by genetic and environmental factors, shape cellular function and disease progression [22].

Table 5. *Summary of Metabolomics Data Sources [22].*

Data sources	Aspect	Descriptions
HMDB (Human Metabolome Database)	Comprehensive Metabolite Information	HMDB provides a vast collection of metabolite data, including chemical structures, spectral information, and biological roles, offering a comprehensive resource for metabolomics research.
MetaboLights	Metabolomics Study Metadata	MetaboLights serves as a repository for metabolomics studies, housing metadata such as experimental protocols, sample information, and analytical data, facilitating data sharing and collaboration.
MassBank	Mass Spectrometry Data	MassBank offers a repository of mass spectrometry data, including spectral information and metabolite identification, supporting metabolomics researchers in the annotation and validation of compounds.
Lipid Maps	Lipid Metabolism Information	Lipid Maps focuses on lipid metabolism, providing a curated resource of lipid structures, pathways, and associated data, aiding in the exploration of lipidomics and its implications in health and disease.
GNPS (Global Natural Products Social Molecular Networking)	Molecular Networking and Dereplication	GNPS facilitates the sharing and analysis of mass spectrometry data, enabling molecular networking for the identification of known and novel metabolites, contributing to metabolomics research and discovery.
Metabolomics Workbench	Diverse Metabolomics Datasets	Metabolomics Workbench hosts a variety of metabolomics datasets, spanning different organisms and experimental conditions, offering researchers access to a broad range of data for comparative analyses and exploration.
NMRShiftDB	Nuclear Magnetic Resonance (NMR) Data	NMRShiftDB compiles nuclear magnetic resonance (NMR) data for metabolites, including chemical shifts and coupling constants, supporting metabolomics investigations leveraging NMR spectroscopy techniques.
KEGG (Kyoto Encyclopedia of Genes and Genomes)	Metabolic Pathways and Annotations	KEGG provides a comprehensive resource for metabolic pathways, offering annotated information on metabolites, enzymes, and their interactions, aiding in the contextualization of metabolomics data within biological pathways.

4. AWS Overview

AWS HealthOmics is a multifaceted service with three primary components—HealthOmics Storage, HealthOmics Analytics, and HealthOmics Workflows. Each component addresses key aspects of genomics and multiomics research, contributing to a seamless and efficient workflow for researchers and clinicians [23].

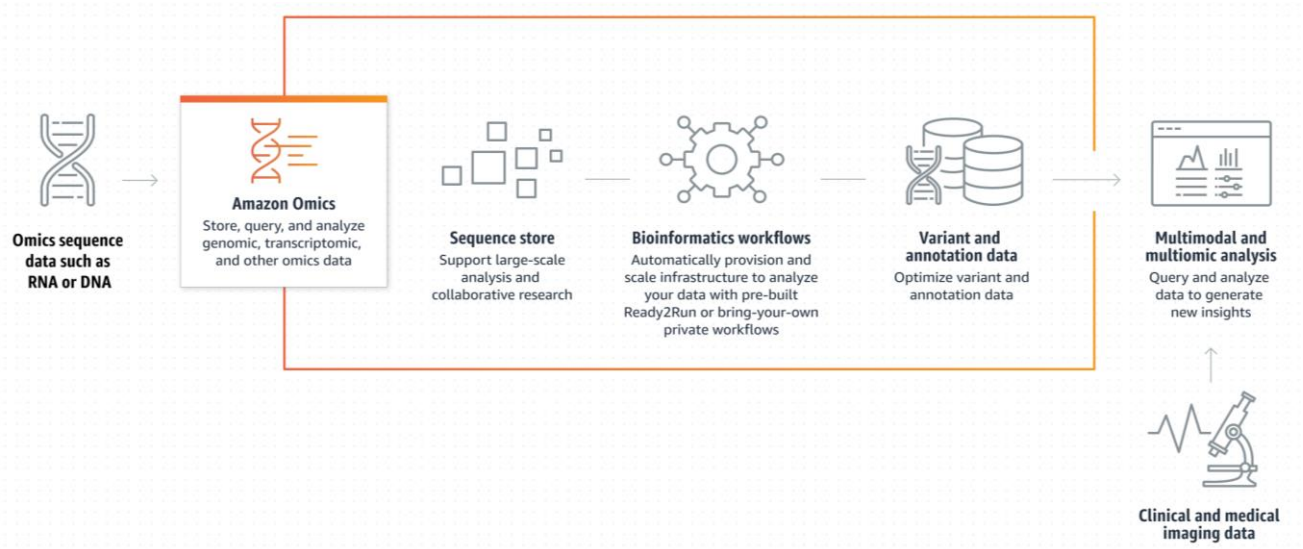

Figure 3. *AWS Bioinformatics Pipeline for Multiomics Analysis [25].*

4.1. HealthOmics Storage

HealthOmics Storage provides a purpose-built storage solution compatible with bioinformatics file formats such as FASTQ, BAM, and CRAM. The storage system allows for efficient, low-cost storage, and sharing of petabytes of genomics data. It supports read-set objects within a sequence store and enables the storage of reference genomes in FASTA format. The use of unique identifiers and attribute-based access controls ensures data provenance and security. AWS HealthOmics Storage is designed to reduce long-term storage costs by automatically archiving objects not accessed within 30 days, with the ability to reactivate archived objects at any time [23, 24].

4.2. Bioinformatics Workflows

AWS HealthOmics offers flexibility in running bioinformatics workflows at scale through two options—Ready2Run workflows and private workflows. Ready2Run workflows, developed by industry leaders and incorporating common open-source pipelines, allow users to process biological data without managing underlying infrastructure. Private workflows enable users to bring their own Workflow Description Language (WDL) or Nextflow scripts, providing a customizable solution with a pay-per-run pricing model [23, 24].

4.3. Analysis at Scale

The platform supports the quick ingestion and transformation of genomics data formats into Apache Iceberg tables, making the data accessible through analytics services like Amazon Athena. It allows for the transformation of both variant and annotation data, enabling comprehensive analyses. AWS HealthOmics ensures data security and access control through integration with AWS Lake Formation, facilitating queries across diverse data sources [25, 26].

4.4. Machine Learning

Machine learning (ML) models have emerged as powerful tools in the realm of healthomics, and Amazon Web Services (AWS) offers a robust platform for their development and deployment. In the context of healthomics, which integrates various omics technologies like genomics, epigenomics, transcriptomics, proteomics, and metabolomics, ML models on AWS play a pivotal role in extracting meaningful insights from vast and complex datasets [27].

AWS provides a scalable and secure infrastructure for implementing ML models in healthomics research. Through services like Amazon SageMaker, researchers can build, train, and deploy ML models efficiently. SageMaker supports a wide range of ML frameworks, allowing flexibility in choosing the most suitable algorithms for the specific healthomic analysis. This adaptability is crucial in handling

the diverse and dynamic nature of omics data. One of the key advantages of leveraging AWS for healthomics ML models is the availability of pre-built machine learning algorithms and models through Amazon SageMaker algorithms [27]. These algorithms are designed to handle common tasks such as feature extraction, classification, and regression, streamlining the model development process. This accelerates the pace of healthomics research, enabling scientists to focus more on the biological interpretation of results rather than the intricacies of model development [28]. Moreover, AWS supports the integration of ML models with other AWS services, enhancing the overall analytical capabilities in healthomics. For instance, Amazon Comprehend Medical can be seamlessly integrated to extract valuable medical insights from unstructured clinical text data. This integration fosters a holistic approach to healthomic analysis, combining molecular-level omics data with clinical information for a more comprehensive understanding of diseases and their underlying mechanisms [29].

4.5. Data Collaboration and Provenance

AWS HealthOmics streamlines collaboration among researchers by simplifying tagging of collaborators, setting up permissions, and securely sharing data. The incorporation of domain-specific metadata enhances data findability, accessibility, interoperability, and reusability (FAIR). This facilitates multiomics and multimodal analyses by linking AWS HealthOmics data stores with other omics and healthcare datasets [30].

4.6. Security, Privacy, and Compliance

AWS HealthOmics is HIPAA eligible, ensuring that it meets stringent security and privacy standards. Attribute-based controls allow for fine-grained data access and governance. The platform includes comprehensive logging and provenance capture features, providing transparency into data access and usage [24].

5. Pros and Cons of Using AWS

Advantages and disadvantages of using Amazon Web Services (AWS) can significantly affect your cloud computing strategy. Understanding these aspects is crucial for informed decision-making in the digital age [23].

Table 6. *Summary of AWS Pros and Cons.*

Pros	Cons
Scalability: Easily scale resources as needed for genomics projects.	Cost: Costs can accumulate quickly, and pricing can be complex. (ref)
Security: Benefit from AWS's robust security features and compliance standards.	Learning Curve: It may take time to learn and configure the services.
Data Management: Efficiently store, process, and manage large genomic datasets.	Connectivity: Reliance on an internet connection may lead to downtime.
Flexibility: Choose from various AWS services tailored to genomics research.	Data Transfer: Uploading and downloading large datasets can be time-consuming.
Integration: Easily integrate with other AWS services and third-party tools.	Compliance: Ensuring data compliance with regulatory standards can be complex.
Collaboration: Eases collaboration among research teams and organizations.	Dependency: Your project's success may depend on AWS services.
Automation: Streamline analysis and workflows with AWS automation tools.	Lock-In vendor: Migrating away from AWS can be challenging and costly.
Analytics: Access advanced analytics and machine learning capabilities.	Support: The quality and responsiveness of support can vary.
Disaster Recovery: Benefit from AWS's disaster recovery and backup solutions.	Data Privacy: Concerns may arise about data privacy and control.
Global Reach: AWS has data centres worldwide, ensuring global accessibility.	Resource Limits: There may be resource limits that affect your projects.

6. Conclusion

In the dynamic landscape of multi-omics breast cancer biomarker research, the integration of AWS cloud solutions stands as a transformative force, catalyzing the journey from raw data to meaningful diagnoses. This exploration has underscored the remarkable versatility and efficiency of AWS in managing the complexities inherent in diverse omics data types, ranging from genomics and epigenomics to transcriptomics, proteomics, and metabolomics. The amalgamation of Amazon Web Services, equipped with its potent computational resources and scalable infrastructure, has not only facilitated but revolutionized the analytical pipelines, empowering researchers to navigate the intricate web of multi-omics datasets with unparalleled speed and precision.

Abbreviations

AWS	Amazon Web Services
BAM	Binary Alignment Map
BCSC	Breast Cancer Surveillance Consortium
BMC	BioMed Central
CA	Cancer
CRAM	Compressed Read Archive in Minutes
DB	Database
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
EDACC	Early Detection Research Network Data Analysis Center
ENCODE	Encyclopedia of DNA Elements
FAIR	Findable, Accessible, Interoperable, Reusable
FASTA	Fast-All Sequence Search Tool
FASTQ	Fast Quality Control
GAIT	Genome Analysis Information Tool
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
GM	Genetically Modified
GNPS	Global Natural Product Social Molecular Networking
GSEA	Gene Set Enrichment Analysis
HIPAA	Health Insurance Portability and Accountability Act
HMDB	Human Metabolome Database
HPM	Human Proteome Map
HUL	Human Unidentified LncRNA
HUMANA	Human Microbiome Analysis
IDE	Integrated Development Environment
IHEC	International Human Epigenome Consortium
IP	Intellectual Property
IPA	Ingenuity Pathway Analysis
ISH	In Situ Hybridization
IVE	Image Visualization Environment
KEGG	Kyoto Encyclopedia of Genes and Genomes
ML	Machine Learning
MOBC	Methylated DNA Immunoprecipitation with Bisulfite Conversion
MPSS	Massively Parallel Signature Sequencing
MS	Mass Spectrometry
NCBI	National Center for Biotechnology Information
NCIP	National Cancer Informatics Program
NGS	Next-Generation Sequencing
NIH	National Institutes of Health
NMR	Nuclear Magnetic Resonance
NMRS	Nuclear Magnetic Resonance Spectroscopy
PL	Pipeline
POST	Power of Statistical Tests
PR	Public Relations
PRESS	Public Repository for Electronically Stored Sequences
PRIDE	PRoteomics IDEntifications database
PRIN	Pipeline of RNA-Sequencing
QB	Quality Base
RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
SNP	Single Nucleotide Polymorphism
TM	Text Mining
UCSC	University of California, Santa Cruz
USA300	United States of America 300
WDL	Workflow Description Language
WHO	World Health Organization

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] A. N. Giaquinto *et al.*, "Breast Cancer Statistics, 2022," *CA. Cancer J. Clin.*, vol. 72, no. 6, pp. 524–541, Nov. 2022, https://doi.org/10.3322/caac.21754

[2] A. Vikram Pawar, "CLASSIFICATION OF BREAST CANCER CELL LINES INTO SUBTYPES BASED ON GENETIC PROFILES," 2015.

[3] F. Tian, Y. Wang, M. Seiler, and Z. Hu, "Functional characterization of breast cancer using pathway profiles," *BMC Med. Genomics*, vol. 7, no. 1, Jul. 2014, https://doi.org/10.1186/1755-8794-7-45

[4] K. Sathishkumar, M. Chaturvedi, P. Das, S. Stephen, and P. Mathur, "Cancer incidence estimates for 2022 & projection for 2025: Result from National Cancer Registry Programme, India," *Indian J. Med. Res.*, vol. 156, no. 4, pp. 598–607, Oct. 2022, https://doi.org/10.4103/ijmr.ijmr_1821_22

[5] S. Malvia *et al.*, "Study of Gene Expression Profiles of Breast Cancers in Indian Women," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, https://doi.org/10.1038/s41598-019-46261-1

[6] J. Trubek and W. Dissertation, "Cancer Bioinformatics for Biomarker Discovery," 2017.

[7] N. Cancer Institute, "DIVISION OF CANCER TREATMENT AND DIAGNOSIS," 2018.

[8] N. S. Fox, "Molecular Cancer Subtypes and Their Associations," 2021.

[9] J. P. Rennhack *et al.*, "Integrated sequence and gene expression analysis of mouse models of breast cancer reveals critical events with human parallels", https://doi.org/10.1101/375154

[10] A. Yazdanparast, "INTEGRATIVE ANALYSIS FOR IDENTIFYING MULTI-LAYER MODULES IN PRECISION MEDICINE," 2020.

[11] A. Dhillon, A. Singh, and V. K. Bhalla, "A Systematic Review on Biomarker Identification for Cancer Diagnosis and Prognosis in Multi-omics: From Computational Needs to Machine Learning and Deep Learning," *Arch. Comput. Methods Eng.*, vol. 30, no. 2, pp. 917–949, Mar. 2023, https://doi.org/10.1007/s11831-022-09821-9

[12] "Multi-Omic Biomarker Identification and Characterization for Posttraumatic Stress Disorder Citation Terms of Use Share Your Story." [Online]. Available: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029495

[13] S. Rahman and A. K. Das, "Integrated Multi-omics, Virtual Screening and Molecular Docking Analysis of Methicillin-Resistant Staphylococcus aureus USA300 for the Identification of Potential Therapeutic Targets: An In-Silico Approach," *Int. J. Pept. Res. Ther.*, vol. 27, no. 4, pp. 2735–2755, Dec. 2021, https://doi.org/10.1007/s10989-021-10287-9

[14] N. Gómez-Cebrián, I. Domingo-Ortí J. L. Poveda, M. J. Vicent, L. Puchades-Carrasco, and A. Pineda-Lucena, "Multi-omic approaches to breast cancer metabolic phenotyping: Applications in diagnosis, prognosis, and the development of novel treatments," *Cancers*, vol. 13, no. 18, Sep. 2021, https://doi.org/10.3390/cancers13184544

[15] S. Firdous, A. Ghosh, and S. Saha, "BCSCdb: A database of biomarkers of cancer stem cells," *Database*, vol. 2022, 2022, https://doi.org/10.1093/database/baac082

[16] Y. Shen *et al.*, "Identification of Potential Biomarkers for Thyroid Cancer Using Bioinformatics Strategy: A Study Based on GEO Datasets," *BioMed Res. Int.*, vol. 2020, 2020, https://doi.org/10.1155/2020/9710421

[17] B. Xie, Z. Yuan, Y. Yang, Z. Sun, S. Zhou, and X. Fang, "MOBCdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine," *Breast Cancer Res. Treat.*, vol. 169, no. 3, pp. 625–632, Jun. 2018, https://doi.org/10.1007/s10549-018-4708-z

[18] L. M. Mcintyre *et al.*, "GAIT-GM: Galaxy tools for modeling metabolite changes as a function of gene expression", https://doi.org/10.1101/2020.12.25.424407

[19] M. Leclercq *et al.*, "Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data," *Front. Genet.*, vol. 10, no. MAY, 2019, https://doi.org/10.3389/fgene.2019.00452

[20] J. Thaiparambil *et al.*, "Integrative metabolomics and transcriptomics analysis reveals novel therapeutic vulnerabilities in lung cancer," *Cancer Med.*, vol. 12, no. 1, pp. 584–596, Jan. 2023, https://doi.org/10.1002/cam4.4933

[21] A. Awasthi, "Applications of Quantitative Proteomics and Phosphoproteomics to Study the Development of Resistance to Targeted Therapy in Cancer Item Type dissertation," 2018. [Online]. Available: http://hdl.handle.net/10713/7927

[22] S. Misener, S. A. Krawetz, and D. D. Womble, "Bioinformatics Methods and Protocols Methods in Molecular Biology Methods in Molecular Biology TM TM VOLUME 132 HUMANA PRESS HUMANA PRESS Bioinformatics Methods and Protocols The Wisconsin Package of Sequence Analysis Programs." [Online]. Available: http://www.gcg.com

[23] Koppad S, B A, Gkoutos GV, Acharjee A. Cloud Computing Enabled Big Multi-Omics Data Analytics. Bioinform Biol Insights. 2021 Jul 28; 15: 11779322211035921. https://doi.org/10.1177/11779322211035921 PMID: 34376975; PMCID: PMC8323418.

[24] AWS. (2024, January). HealthOmics: Transform genomic, transcriptomic, and other omics data into insights. Retrieved from https://docs.aws.amazon.com/omics/latest/dev/what-is-service.html

[25] N. G. Alharbi, "Interactive Visualization of Molecular Dynamics Simulation Data," 2020.

[26] X. Zhang, "Learning from Multi-Omics Data of Cancer," 2021.

[27] R. Diaz-Uriarte *et al.*, "Ten quick tips for biomarker discovery and validation analyses using machine learning," *PLoS Comput. Biol.*, vol. 18, no. 8, Aug. 2022, https://doi.org/10.1371/journal.pcbi.1010357

[28] X. Zeng, G. Shi, Q. He, and P. Zhu, "Screening and predicted value of potential biomarkers for breast cancer using bioinformatics analysis," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, https://doi.org/10.1038/s41598-021-00268-9

[29] A. Al-Fatlawi *et al.*, "Netrank: network-based approach for biomarker discovery," *BMC Bioinformatics*, vol. 24, no. 1, Dec. 2023, https://doi.org/10.1186/s12859-023-05418-6

[30] Z. Z. Hu *et al.*, "Omics-Based Molecular Target and Biomarker Identification," in *Methods in Molecular Biology*, vol. 719, Humana Press Inc., 2011, pp. 547–571. https://doi.org/10.1007/978-1-61779-027-0_26