OXFORD

BIOLOGY Methods & Protocols

Biology Methods and Protocols, 2024, bpae040

https://doi.org/10.1093/biomethods/bpae040
Advance Access Publication Date: 29 May 2024
Methods Article

IntelliGenes: Interactive and user-friendly multimodal AI/ML application for biomarker discovery and predictive medicine

Rishabh Narayanan [1], William DeGroat[1], Dinesh Mendhe[1], Habiba Abdelhalim[1] and Zeeshan Ahmed [1,2,*]

[1]Rutgers Institute for Health, Health Care Policy and Aging Research, The State University of New Jersey, New Brunswick, 08901, NJ, United States
[2]Department of Medicine, Division of Cardiovascular Disease and Hypertension, Robert Wood Johnson Medical School, New Brunswick, NJ, 08901, United States

*Correspondence address. Rutgers Institute for Health, Health Care Policy and Aging Research, Rutgers University, 112 Paterson Street, New Brunswick, 08901, NJ, United States. E-mail: zahmed@ifh.rutgers.edu

Abstract

Artificial intelligence (AI) and machine learning (ML) have advanced in several areas and fields of life; however, its progress in the field of multi-omics is not matching the levels others have attained. Challenges include but are not limited to the handling and analysis of high volumes of complex multi-omics data, and the expertise needed to implement and execute AI/ML approaches. In this article, we present IntelliGenes, an interactive, customizable, cross-platform, and user-friendly AI/ML application for multi-omics data exploration to discover novel biomarkers and predict rare, common, and complex diseases. The implemented methodology is based on a nexus of conventional statistical techniques and cutting-edge ML algorithms, which outperforms single algorithms and result in enhanced accuracy. The interactive and cross-platform graphical user interface of IntelliGenes is divided into three main sections: (i) Data Manager, (ii) AI/ML Analysis, and (iii) Visualization. Data Manager supports the user in loading and customizing the input data and list of existing biomarkers. AI/ML Analysis allows the user to apply default combinations of statistical and ML algorithms, as well as customize and create new AI/ML pipelines. Visualization provides options to interpret a diverse set of produced results, including performance metrics, disease predictions, and various charts. The performance of IntelliGenes has been successfully tested at variable in-house and peer-reviewed studies, and was able to correctly classify individuals as patients and predict disease with high accuracy. It stands apart primarily in its simplicity in use for nontechnical users and its emphasis on generating interpretable visualizations. We have designed and implemented IntelliGenes in a way that a user with or without computational background can apply AI/ML approaches to discover novel biomarkers and predict diseases.

Keywords: artificial intelligence; machine learning; multi-omics; multimodal; biomarker discovery; predictive analysis

Introduction

Analyzing integrated multi-omics (e.g. genomics, transcriptomics, epigenomics, proteomics, and metabolomics) and phenotypic data, has the potential to deeply enhance our understanding of the underlying genetic basis of human diseases [1]. Recent developments in the fields of artificial intelligence (AI) and machine learning (ML) have improved our ability for analyzing RNA-seq-driven gene expression data, whole-genome sequencing (WGS) variant data, clinical and demographic-based patient data, and for discovering complex and nonlinear genotypic–phenotypic relationships [2]. The rightful application of AI/ML techniques can aid in unveiling associated biomarkers and help stratify patient populations based on disease risk to provide personalized treatment options [3]. However, despite the progress made by AI in various biomedical realms, numerous challenges include but are not limited to, high volume multi-omics data inaccessibility, collection, processing, and management; generation and dissemination of AI/ML-ready datasets for predictive analysis and deep phenotyping; data modeling with the establishment of correct associations between input variables and expected outcomes; overfitting and generalization; interpretability and explainability of AI-driven insights; handling inherent error rates and missing data imputations; and ineffective interrogation including delayed computational time and inability to reproduce results [1]. These can have a vital impact on the accuracy of predictions. Each model can make fundamentally different assumptions about the datasets, such as independence, even distribution, or absence of multicollinearity.

State-of-the-art literature has supported the implementation of various AI/ML methodologies and algorithms. Researchers in one multi-omics study developed a deep learning framework using integrated omics, demographic, and clinical data to extract insights from the high-dimensional gene and miRNA expression data, enabling prognosis prediction for patients with breast and ovarian cancer [4]. In another example, investigators integrated genetic variants with demographic and clinical data to predict survival rates for a dementia patient using an Xtreme gradient boosting algorithm [5]. A new multi-omics and multimodal analytic platform has been published (i.e. Molecular Twin) implementing an ensemble of variable ML algorithms to accurately predict disease survival using genomic, transcriptomic, proteomic, and lipid data from adenocarcinoma patients [6].

Received: 17 April 2024. **Revised:** 19 May 2024. **Editorial decision:** 20 May 2024. **Accepted:** 28 May 2024

Unsupervised approaches have also been reported, for example, in one study utilizing gut microbiome to reevaluate and reclassify acquired neonatal intestinal injuries [7]. A genomic language model (gLM) trained on millions of metagenomic scaffolds was also reported to uncover hidden functional and regulatory connections among genes [8]. Each existing tool provides valuable workflows for multi-omics analyses on different datasets, but most still require some level of programmatic proficiency that creates barriers for usage.

Addressing some of the current challenges, especially related to the AI/ML-ready data generation, optimized AI/ML modeling, and interpretable and reproducible AI/ML results production, we have developed and published a new multimodal solution, that is, *IntelliGenes*, a command-line AI/ML approach that employs a combination of classical statistical models for feature selection and cutting-edge ML classifiers for disease prediction [9]. It is based on Findable, Accessible, Intelligent, and Reproducible paradigm that outperforms single algorithms and results in enhanced accuracy, deeper insights, and more precise predictions essential for personalized early disease-risk detection in individuals. In this study, we present an advanced version of *IntelliGenes*, which empowers the user to perform interactive AI/ML-ready data preparation and management, customized and tailored AI/ML analysis, and automated and interpretable visualization of results.

Materials and methods

The overall methodology is divided into three sections: (i) AI/ML-ready data generation in Clinically Integrated Genomics and Transcriptomics (CIGT) format, (ii) ML and classical statistics analysis, and (iii) interactive and customizable user interface.

AI/ML-ready data generation in CIGT format

Multi-omics broadly refers to the types of biological data, such as genomics (e.g. variants), transcriptomics (e.g. gene expression), proteomics (e.g. protein abundances), metabolomics (e.g. metabolite levels), epigenomics (e.g. methylation profiles), etc. Handling multi-omics data is fraught with many challenges because of its heterogeneity [10] and large volumes [11] requiring costly high-performance computing (HPC) and storage [12]. Another persistent challenge in integrating multi-omics data is the incompatible data formats and unstandardized vocabularies used, such as the different systems for variant or protein IDs [10, 13]. Data missingness also presents a major challenge when managing multi-omics data as certain features may be partially omitted for reasons including poor quality [14], privacy and ethical challenges [15], or limited funding [16]. Considering these challenges, we proposed a new AI/ML-ready data format, that is, CIGT. It supports integrating multi-omics data for predictive analysis within our *IntelliGenes* framework. The CIGT format is specifically designed for extensibility, interpretability, and AI/ML readiness, being directly pluggable into multimodal AI/ML analyses. It employs a tabular structure encoding samples as rows and multi-omics/clinical features as columns. Namely, a dataset may consist of useful patient demographic information (e.g. race, age, and gender), and biomarkers such as transcriptomic (e.g. RNA-seq-based expression counts) and genomic (whole-genome/exome-based variants) features. The attached Supplementary material (User Guide) provides further details on the CIGT formatting and usage for AI/ML analysis.

ML and classical statistics analysis

IntelliGenes applies a dual-phase analysis to discover biomarkers and create efficient, generalizable models for single-disease prediction. The implemented methodology in *IntelliGenes* is based on a nexus of conventional statistical techniques and cutting-edge ML algorithms. By combining predictions across multiple selectors and ML classifiers, we strengthen the confidence in individual predictors and further diversify the types of data to which our methodology may be applied to. We have published a study in the *Briefings in Bioinformatics* (Oxford) outlining the advantages and disadvantages of ML algorithms and their application to various diseases [3]. We recommend researchers using *IntelliGenes* treat this publication as a companion piece to best guide customization.

IntelliGenes extracts predicted disease-associated biomarkers from a CIGT-formatted dataset using a robust, user-driven ensemble of selectors. Currently, *IntelliGenes* implements Pearson's correlation (PC), the chi-square test (χ^2), analysis of variance (ANOVA), and recursive feature elimination (RFE). Selectors in each user-built model run independently, and biomarkers that are present in *IntelliGenes*' initial output must be found statistically significant ($P = .05$) in all included statistical tests (PC, χ^2, or ANOVA) and within the 90th quantile in RFE ranks. Subsequently, *IntelliGenes* can utilize these discovered biomarkers to create, optimize, and test ML models for patient single-disease prediction. Specifically, the discovered biomarkers are used directly as input features along with patient diagnosis information to train a configurable suite of supervised ML classifiers for single-disease prediction. Our default multimodal pipeline currently supports five classifiers: random forest, support vector machine, Xtreme Gradient Boosting (XGBoost), k-nearest neighbors, and multi-layer perceptron. Users can customize need-specific ML models with uncontaminated training datasets and tune hyperparameters using GridSearchCV. *IntelliGenes* captures diverse predictions from multiple classifiers which are later compiled downstream with a voting algorithm, and our novel metric, I-Genes Score. These scores profile a genomic and/or transcriptomic feature's usefulness in predicting disease and its regulation in patients (i.e. upregulation or downregulation). Alongside these metrics, *IntelliGenes* reports accuracy, AUC, and F1 scores for each model.

Interactive and customizable user interface

Many clinicians and bench scientists are not trained to incorporate AI/ML techniques into their research and clinical practices [17, 18]. In this study, we present an advanced version of *IntelliGenes* to place ML techniques into the hands of noncomputational experts with an interactive sandbox. The graphical user interface (GUI) of software comprises three sections: Data Manager, AI/ML Analysis, and Visualization (Fig. 1). Data Manager allows the user to shape the pipeline's input. *IntelliGenes* expects a CIGT-formatted dataset. AI/ML Analysis offers users complete customization for *IntelliGenes*' methodology. Here, selectors and classifiers are chosen. Additional options, such as dataset normalization, training size, and random state reproducibility, are available in this section. Users can bypass the selection process, providing a curated list of biomarkers for training models.

Visualization provides users with a detailed and customizable view of the statistical results and visual outputs. The importance of diverse visualizations lies in the different interpretations offered into the underlying datasets, and in the effective visual approach to assimilating multifarious results. The current

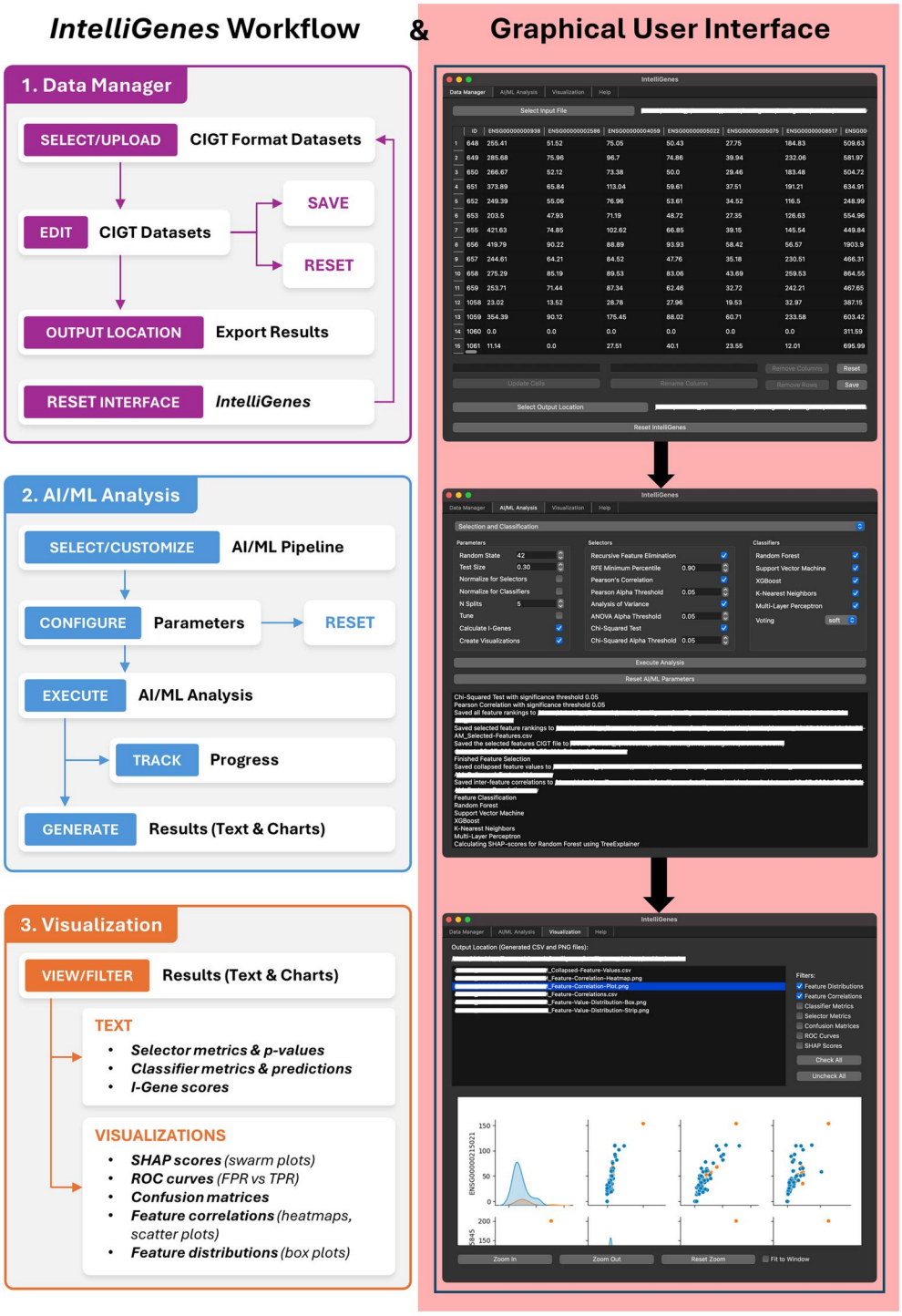

Figure 1. *IntelliGenes* workflow (left) and graphical user interface (right): (1) Data Manager, (2) AI/ML Analysis, and (3) Visualization. Data Manager supports the user in loading and editing input dataset. AI/ML Analysis allows the user to combine statistical and ML algorithms into custom AI/ML pipelines. Visualization provides options to interpret a diverse set of produced results (e.g. performance metrics, disease predictions, and various charts).

interactive version of *IntelliGenes* generates seven primary plot types: (i) confusion matrices, (ii) feature correlation heatmaps, (iii) pairwise inter-feature scatter plots, (iv) intra-feature kernel density estimation (KDE) plots, (v) feature distribution box plots and swarm plots, (vi) reporter-operating curves (ROC), and (vii) feature importance swarm plots using SHapley Additive exPlanations (SHAP). When analyzing the accuracy of each model, especially on imbalanced datasets that are biased toward specific predictions, it is important to understand not only the overall accuracy but also the specificity of the model. To distill a model's ability to accurately discriminate diagnosis types, *IntelliGenes* generates confusion matrices for each model. A confusion matrix depicts each of the true positive, false positive (Type I error), true negative, and false negative (Type II error) rates. Analyzing these plots may also crucially aid in identifying sub-optimal parameters and results. For example, a confusion

matrix with large values across the main diagonal and comparatively low values elsewhere is indicative of a good (highly discriminative) classifier. With multi-dimensional datasets, it is useful to analyze pairwise feature correlations to assess inter-feature relationships and any multi-collinearity. Analyzing collinearity is important with AI/ML models as a highly collinear feature space may create volatile predictions if feature independence is assumed. *IntelliGenes* produces correlation heatmaps which organize pairwise Pearson correlation coefficients in a square grid and provide insight on pairwise collinearity. Moreover, individual pairwise scatter plots offer more detailed insight into correlation heatmaps by enabling the user to visually gauge clustering, linear, and nonlinear relationships. They are particularly useful when analyzing the centering and variance of the data under each diagnosis/response type. KDE plots, represented in the diagonal of the pairwise plot, are also useful in estimating the distribution of each feature for each response type.

To better analyze centering, variability, summary statistics, and individual feature expression levels, *IntelliGenes* generates both box plots and swarm plots for each feature partitioned by response. These plots provide a good visual indicator of strong predictors since features with lower variation within responses as compared to between responses are more easily discriminated. This idea is foundational to the ANOVA algorithm supported by *IntelliGenes*. It is important to note that we generate the plots with a logarithmic axis to ensure that both highly and lowly expressed genes are easily visible. The scatter plots, box plots, and swarm plots each offer different views into the distribution of data. Analyzing these datasets may help identify outliers or strong clusters of interest warranting further investigation. ROC graphs also provide insight into the model's ability to accurately differentiate between diagnosis types, measuring the tradeoff between sensitivity (true-positive rate) and specificity (false-positive rate). Inspecting the ROC curve enables scientists to determine whether predictions are more systematic or random by analyzing how close the curve is to the upper left corner (100% true positive, 0% false positive). An accurate model would have an ROC curve very close to the upper left column, whereas a model with seemingly random predictions would have a proportional amount of true- and false-positive rates (i.e. a straight diagonal curve). Last, the ability to interpret and understand the underlying biological pathways of otherwise black-box results is crucial in genomic analysis that employs AI/ML techniques. SHAP scores are a commonly used technique to identify both the magnitude and directional impact of each feature in local predictions. *IntelliGenes* generates swarm plots of SHAP scores to enable scientists to better explain and understand the contributions of individual biomarkers on prediction.

Our methodology is developed with an extensible and modular architecture that enables the easy addition of impactful visualizations to further aid in the explanation and interpretation of analytical results. *IntelliGenes* is an open-source application, and the source code is freely available on GitHub. The software was developed using Python and relies on existing libraries. *Pandas* and *numpy* are used for data manipulation and transformation of our CIGT format. The statistical methods and ML algorithms at the center of the methodology are implemented using *scikit-learn, SciPy,* and *XGBoost*. *SHAP* is used to crucially interpret feature importance for otherwise black-box ML methods. The diverse visualizations are generated using *Matplotlib* and *seaborn*, and the rendering of cross-platform GUI components is achieved through the extensible *Qt* framework with *PySide6*. To package the executable across major operating systems (i.e. MacOS and Windows), we used *PyInstaller*. The *IntelliGenes* was designed with a focus on modularity and extensibility, simplifying the addition of new methods and analyses. Generation of customized AI/ML pipeline is represented internally as a simple executable function with a series of configurable parameters that may be adjusted visually through the GUI. *IntelliGenes* is capable of making the best utilization of available hardware resources by using multi-processing, allowing the user interface to remain responsive even with intensive computation. *IntelliGenes* has been tested on the major operating systems, including Windows, MacOS, and Linux. Supplementary material 1 attached includes an extensive manual detailing installation, CIGT data formation, examples of all generated visualizations, and step-by-step usage.

Results

IntelliGenes has been utilized to discover novel biomarkers for cardiovascular disease (CVD) and create models that made high-accuracy CVD predictions (0.97 AUC) [14]. Here, we pre-processed a CIGT-formatted dataset to include well-represented transcriptomic biomarkers (<20% NaN) in a diverse 61-patient cohort. Transcripts per million (TPM) were used to quantify RNA-seq-driven expression, although normalized counts are also compatible with *IntelliGenes*. We discovered 18 significant transcriptomic biomarkers, of which 14 have been known to be associated with CVD. In addition, we identified three significant biomarkers (*RN7SL593P, AP003419.11,* and *CTA-363E6.6*), which would require further analysis to understand their role in disease etiology [14]. Demonstrating the potential to investigate the correlations between multifactorial and interrelated diseases, in another peer-reviewed study, we reported four genes (*GAS5, GPX1, HLA-B,* and *SNHG6*) associated with CVDs having a strong correlation with periodontitis [19]. All the statistical tests and ML algorithms currently available in *IntelliGenes* were tested. We suggest that user replicate the dataset preprocessing detailed before running *IntelliGenes*. The datasets presented in the Supplementary materials and detailed in this section use the Ensembl ID system to identify genes [20]. Although gene symbols are more easily recognizable, Ensembl IDs typically provide better clarity. The CIGT format, and by extension IntelliGenes, allows user to encode their dataset using any system of their choice. In the future, plan to provide support to convert between Ensembl IDs and gene symbols for further convenience and interpretability.

Discussion

Integrating ML predictive engines with multi-omics data (e.g. RNA-seq-driven prediction of disease risk, WGS-driven disease-associated variant screenings) has the potential to revolutionize early detection and personalized treatment for common and rare diseases [21, 22]. We need multimodal intelligent solutions, which can support in solving the current challenges and future developments. Most importantly related to the high-volume complex multi-omics data handling and integration [23], and implementation of novel AI/ML applications for predictive analysis and biomarkers discovery at large [24]. In this study, we tried to address both challenges by introducing CIGT, an extensible and interpretable data format supporting the encoding of patient diagnosis information, gene expression data, clinical demographic attributes, and genomic variant data into an AI/ML-ready dataset. We proposed a new nexus of ML algorithms to support significant biomarker discovery and disease predictions. In addition, the diversity of visualizations generated by *IntelliGenes* may

render results more accessible, interpretable, and understandable to a wider audience of scientists. Our methodology was developed with accessibility and ease of use, paying special attention to the impact of each visualization on the types of analyses supported to ensure that the nuance of the dataset is represented in a more appropriate and interpretable ways.

In the future, we plan on continuing to add support for different types of visualizations and to enhance the global interactivity of our Visualization interface. Specifically, support for 3D graphics to visualize higher dimensional structures, options to generate user-customizable visualizations on-the-fly, and a more interactive image viewport may be added to streamline the interpretation phase of complex multi-omics data analysis. *IntelliGenes* has the potential to incentivize AI/ML research that pairs disease experts with analysts passionate about unraveling high-dimensional multi-omics data. It may empower its user to easily design intricate frameworks for examining and predicting diseases tailored to their research needs. We demonstrate that portable predictive models like *IntelliGenes* that are accessible to clinicians and non-computational scientists are feasible for scientific research and impactful clinical environments.

Acknowledgements

We appreciate great support from the Department of Medicine, Division of Cardiovascular Disease and Hypertension, Robert Wood Johnson Medical School; Rutgers Institute for Health, Health Care Policy, and Aging Research; and Rutgers Health, at Rutgers, The State University of New Jersey. The authors acknowledge the Rutgers Office of Advanced Research Computing (OARC) for providing access to the Amarel cluster and associated research computing resources.

Author contributions

Rishabh Narayanan (Software [equal], Methodology [equal], Writing—original draft [equal]), William DeGroat (Methodology [equal]), Dinesh Mendhe (Validation [equal], Writing—original draft [equal]), Habiba Abdelhalim (Project administration [equal]), and Zeeshan Ahmed (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Funding acquisition [lead], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead])

Supplementary data

Supplementary data is available at *Biology Methods and Protocols* online.

Conflict of interest statement. None declared.

Funding

This work has been supported by the Department of Medicine, Robert Wood Johnson Medical School, and Rutgers Institute for Health, Health Care Policy, and Aging Research at Rutgers, The State University of New Jersey.

Data availability

The source code of *IntelliGenes* is available on GitHub <https://github.com/drzeeshanahmed/IntelliGenes-GUI> and Code Ocean < https://codeocean.com/capsule/2636977/tree/v1 >

Biographical note

R.N., W.D., and H.A. are the Research Assistant at the Ahmed lab, Rutgers IFH/RWJMS. D.M. is the Lead Software Engineer at Rutgers IFH.

Z.A. is the Assistant Professor at the Department of Medicine/Division of Cardiovascular Diseases and Hypertension, Rutgers Robert Wood Johnson Medical School, and Rutgers Health. Z.A. is a Core Faculty Member at the Rutgers Institute for Health, Health Care Policy and Aging Research, at Rutgers, The State University of New Jersey. Furthermore, Z.A. is the Adjunct Assistant Professor at the Department of Genetics and Genome Sciences, School of Medicine, UConn Health, CT.

References

1. Ahmed Z, Wan S, Zhang F, Zhong W. Artificial intelligence for omics data analysis. *BMC Methods* 2024;**1**:4.
2. Ahmed Z, Mohamed K, Zeeshan S *et al.* Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* 2020; **2020**:baaa010.
3. Vadapalli S, Abdelhalim H, Zeeshan S *et al.* Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Briefings in Bioinformatics* 2022;**23**:bbac191.
4. Jiang L, Xu C, Bai Y *et al.* Autosurv: interpretable deep learning framework for cancer survival analysis incorporating clinical and multi-omics data. *NPJ Precis Oncol* 2024;**8**:4.
5. Zhang J, Song L, Miller Z *et al.* Machine learning models identify predictive features of patient mortality across dementia types. *Commun Med (Lond)* 2024;**4**:23.
6. Osipov A, Nikolic O, Gertych A *et al.* The molecular twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. *Nat Cancer* 2024;**5**:299–314.
7. Gipson DR, Chang AL, Lure AC *et al.* Reassessing acquired neonatal intestinal diseases using unsupervised machine learning. *Pediatr Res* 2024. https://pubmed.ncbi.nlm.nih.gov/38413766/.
8. Hwang Y, Cornman AL, Kellogg EH *et al.* Genomic language model predicts protein co-regulation and function. *Nat Commun* 2024;**15**:2880.
9. DeGroat W, Mendhe D, Bhusari A *et al.* IntelliGenes: a novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles. *Bioinformatics (Oxford, England)* 2023;**39**:btad755.
10. Manzoni C, Kia DA, Vandrovcova J *et al.* Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform* 2018;**19**:286–302.
11. Stephens ZD, Lee SY, Faghri F *et al.* Big Data: astronomical or Genomical? *PLoS Biol* 2015;**13**:e1002195.
12. Alvarez RV, Mariño-Ramírez L, Landsman D. Transcriptome annotation in the cloud: complexity, best practices, and cost. *GigaScience* 2021;**10**:giaa163.
13. García SA, Costa M, Leon A *et al.* The challenge of managing the evolution of genomics data over time: a conceptual model-based approach. *BMC Bioinformatics* 2022;**23**:472.

14. Claerhout B, DeMoor GJ. Privacy protection for clinical and genomic data. The use of privacy-enhancing techniques in medicine. *Int J Med Inform* 2005;**74**:257–65.
15. Hu M, Zhu J, Peng G *et al.* IMOVNN: incomplete multi-omics data integration variational neural networks for gut microbiome disease prediction and biomarker identification. *Brief Bioinformatics* 2023;**24**:bbad394.
16. Recht MP, Dewey M, Dreyer K *et al.* Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol* 2020;**30**:3576–84.
17. Romero-Brufau S, Wyatt KD, Boyum P *et al.* A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. *Int J Med Inform* 2020;**137**:104072.
18. DeGroat W, Abdelhalim H, Patel K *et al.* Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. *Sci Rep* 2024;**14**:1.
19. Ahmed Z, Degroat W, Abdelhalim H *et al.* Deciphering genomic signatures associating human dental oral craniofacial diseases with cardiovascular diseases using machine learning approaches. *Clin Oral Investig* 2024;**28**:52.
20. Hubbard T, Barker D, Birney E *et al.* The Ensembl genome database project. *Nucleic Acids Res* 2002;**30**:38–41.
21. Perkins BA, Caskey CT, Brar P *et al.* Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proc Natl Acad Sci USA* 2018;**115**:3686–91.
22. Ahmed Z. Precision medicine with multi-omics strategies, deep phenotyping, and predictive analysis. *Prog Mol Biol Trans Sci* 2022;**190**:101–25.
23. Chen C, Wang J, Pan D *et al.* Applications of multi-omics analysis in human diseases. *Medcomm (2020)* 2023;**4**:e315.
24. Reel PS, Reel S, Pearson E *et al.* Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol Adv* 2021;**49**:107739.

Biology Methods and Protocols, 2024, 00, 1–6
https://doi.org/10.1093/biomethods/bpae040
Methods Article