Research Perspective Open Access

# Biostatistical Challenges in High-Dimensional Data Analysis: Strategies and Innovations

Jianjun Wang

BGI Genomics Co., Ltd., Shenzhen, 518083, Guangdong, China

Corresponding email: jianjun.wang@jicat.org

Computational Molecular Biology, 2024, Vol.14, No.4 doi: 10.5376/cmb.2024.14.0019

Received: 09 Jun., 2024

Accepted: 28 Jul., 2024

Published: 12 Aug., 2024

**Preferred citation for this article**:

Wang J., 2024, Biostatistical challenges in high-dimensional data analysis: strategies and innovations, Computational Molecular Biology, 14(4): 163-172 (doi: 10.5376/cmb.2024.14.0019)

**Abstract** In contemporary biological research, the emergence of high-dimensional data has become the norm, especially in fields such as genomics, transcriptomics, and metabolomics. With the widespread application of high-dimensional data, researchers must adopt appropriate strategies to address issues of data sparsity, multicollinearity, and heterogeneity. This study not only summarizes existing dimensionality reduction, regularization, and ensemble learning methods, but also discusses innovative technologies such as machine learning, deep learning, and multi omics data integration to address high-dimensional problems in biological data, providing effective strategies and cutting-edge methods for researchers and data scientists.

**Keywords** High-dimensional data; Biostatistical challenges; Machine learning; Multi-omics data integration; Regularization methods

## 1 Introduction

The advent of high-throughput technologies has revolutionized biological research, enabling the generation of vast amounts of high-dimensional data across various omics layers, including genomics, transcriptomics, proteomics, and metabolomics. These technologies facilitate a comprehensive understanding of biological systems by allowing the simultaneous measurement of thousands of variables, thus providing a multi-faceted view of cellular processes and disease mechanisms. For instance, single-cell RNA sequencing has enabled the profiling of genome-wide features at the single-cell level, revealing unprecedented levels of biological variation (Amezquita etb al., 2019). Similarly, the integration of multi-omics data has become a cornerstone in systems biology, aiming to elucidate complex molecular interactions and pathways (Misra et al., 2019; Wörheide et al., 2021).

Despite the transformative potential of high-dimensional data, its analysis poses significant biostatistical challenges. The high dimensionality and heterogeneity of the data, coupled with issues such as batch effects, data sparsity, and the need for integration across different omics layers, complicate the extraction of meaningful biological insights (Davis-Turak et al., 2017; Juan and Huang, 2023). Traditional statistical methods often fall short in handling the complexity and scale of such data, necessitating the development of specialized techniques. For example, the least absolute shrinkage and selection operator (LASSO) and its derivatives have been employed to address issues of multicollinearity and feature selection in high-dimensional settings (Kim et al., 2019). Moreover, knowledge-guided statistical learning methods have shown promise in improving prediction accuracy and interpretability by incorporating biological knowledge into the analysis (Zhao et al., 2019).

This study provides a comprehensive overview of the biostatistics challenges related to high-dimensional data analysis in biological research, and discusses innovative strategies developed to address these challenges. We will explore various methods of data quantification, integration, and interpretation, with a focus on their applications and limitations. In order to guide future research directions and promote the development of robust analytical frameworks that can fully utilize the potential of high-dimensional biological data.

## 2 Characteristics of High-Dimensional Data

### 2.1 Dimensionality and complexity

High-dimensional data in the context of biological research often involves datasets with a vast number of features but relatively few samples. This imbalance, known as the high-dimension, low sample size (HDLSS) problem, poses significant computational and statistical challenges. For instance, traditional regression methods may become ineffective due to the non-identifiability of the optimization problem inherent in such data (Kim et al., 2019; Vinga, 2020). The complexity is further compounded by the need to integrate diverse types of data, such as genomics, transcriptomics, proteomics, and metabolomics, each contributing to the overall dimensionality and complexity (Misra et al., 2019; Leonavicius et al., 2019). Advanced techniques like structured sparsity regularization and high-dimensional LASSO (Hi-LASSO) have been developed to address these challenges by imposing additional constraints and improving feature selection and prediction performance.

### 2.2 Data sparsity and multicollinearity

Data sparsity is a common characteristic of high-dimensional biological datasets, particularly those derived from single-cell RNA sequencing and other high-throughput technologies. This sparsity arises because many features (e.g., genes) may not be expressed in all samples, leading to a large number of zero values in the dataset (Amezquita et al., 2019). Multicollinearity, where features are highly correlated, further complicates the analysis. This issue is particularly problematic in omics data, where different types of measurements (e.g., gene expression, protein levels) are often interrelated (ShyamMohanJ, 2016; Alzubaidi, 2018). Techniques such as regularized optimization and the development of specialized algorithms like Hi-LASSO help mitigate these issues by refining importance scores and improving the robustness of the models.

### 2.3 Heterogeneity in biological datasets

Biological datasets are inherently heterogeneous, reflecting the complex and varied nature of biological systems. This heterogeneity can be seen across different levels, from the molecular (e.g., gene expression) to the cellular (e.g., single-cell measurements) and even the organismal level. The integration of multi-modal high-throughput data, such as combining genomic, transcriptomic, and proteomic data, is essential for a comprehensive understanding of biological processes but introduces significant challenges. Single-cell technologies, for example, have highlighted the variability between individual cells, necessitating advanced analytical methods to accurately interpret this heterogeneity (Palit et al., 2019). The development of computational principles and tools for data integration is crucial to address these challenges and to enable the extraction of meaningful insights from complex biological data (Argelaguet et al., 2021; Juan and Huang, 2023).

## 3 Biostatistical Challenges in High-Dimensional Data Analysis

### 3.1 Curse of dimensionality

The curse of dimensionality refers to the exponential increase in computational complexity and data sparsity as the number of dimensions in a dataset grows. This phenomenon poses significant challenges in various domains, including feature selection, clustering, and anomaly detection. For instance, in single-cell RNA sequencing (scRNA-seq) data, the high dimensionality combined with technical noise complicates downstream analyses. The RECODE method has been proposed to address this issue by reducing noise without dimension reduction, thereby improving the accuracy of cell clustering and gene expression recovery (Imoto et al., 2022). Similarly, data augmentation techniques have been employed to mitigate sparsity in industrial data, enhancing the robustness and interpretability of data-driven models (Jiang et al., 2023). Evolutionary algorithms like the variable-size cooperative coevolutionary particle swarm optimization (VS-CCPSO) have also shown promise in effectively selecting relevant features from high-dimensional datasets (Song et al., 2022).

### 3.2 Model overfitting

3.2.1 Challenges with model complexity

High-dimensional data often contain correlated and noisy predictors, making it difficult to fit empirical models without overfitting. For example, in hyperspectral remote sensing, the large number of narrow spectral bands can lead to overly complex models that fit the noise rather than the underlying biological processes (Rocha et al., 2017). Similarly, in biomedical datasets, the high dimensionality can obscure the true signal, making it challenging to develop reliable predictive models (Yan et al., 2018).

3.2.2 Strategies for reducing overfitting

Several strategies have been proposed to reduce overfitting in high-dimensional data analysis. One approach is the use of feature selection methods that identify and retain only the most relevant features. For instance, a Genetic Programming-based feature selection strategy has been shown to effectively reduce the data space while maintaining classification accuracy (Viegas et al., 2018). Another strategy involves the use of hybrid models that combine unsupervised feature extraction with supervised learning. A model combining deep belief networks (DBNs) and one-class support vector machines (SVMs) has demonstrated reduced training and testing times while maintaining high anomaly detection performance (Erfani et al., 2016).

3.2.3 Model selection criteria

Optimal model complexity can be determined using various model selection criteria. The Naive Overfitting Index Selection (NOIS) method, for example, quantifies relative model overfitting and selects an optimal model complexity supported by the data. This method has been shown to select less complex models that perform comparably to those selected by traditional cross-validation methods, thereby reducing the risk of overfitting (Rocha et al., 2017).

### 3.3 Statistical inference and multiple testing problems

High-dimensional data often require multiple statistical tests, increasing the risk of false positives. This is a significant challenge in fields like genomics, where thousands of genes may be tested simultaneously for associations with a particular trait. Effective strategies to address this issue include the use of correction methods such as the Bonferroni correction or the False Discovery Rate (FDR) approach. Additionally, advanced noise reduction techniques like RECODE can improve the reliability of statistical inferences by mitigating the impact of technical noise in high-dimensional datasets (Imoto et al., 2022).

## 4 Strategies for Addressing Biostatistical Challenges

### 4.1 Dimensionality reduction techniques

Dimensionality reduction is a critical step in high-dimensional data analysis, particularly in biomedical research where datasets often contain a vast number of features. Techniques such as feature selection and feature extraction are commonly employed to reduce the complexity of the data while retaining essential information (Ashraf et al., 2023). For instance, MRMD3.0 is a tool that integrates various feature ranking algorithms using an ensemble strategy to ensure robustness and accuracy in dimensionality reduction (He et al., 2023). Additionally, hybrid approaches that combine filter and wrapper methods, optimized using algorithms like Genetic Algorithm (GA), have shown significant improvements in predictive accuracy by effectively reducing dimensionality (Gangavarapu and Patil, 2019). These methods help in identifying key features, thereby facilitating better data visualization and analysis.

### 4.2 Regularization methods

Regularization techniques are essential for managing the high dimensionality of omics data, which often leads to overfitting and non-identifiability issues in traditional regression models. Structured sparsity regularization, which includes methods like the elastic net that combines lasso and ridge penalizations, has been particularly effective in building parsimonious models. These models not only enhance prediction accuracy but also improve interpretability by identifying relevant molecular signatures (Vinga, 2020). Such regularization methods impose additional constraints on the solution parameter space, thereby addressing the ill-posed nature of high-dimensional data problems.

### 4.3 Ensemble learning approaches

Ensemble learning methods have proven to be highly effective in handling high-dimensional data by combining multiple models to improve classification performance. For example, an ensemble learning framework that integrates various algorithms has been shown to increase accuracy by up to 3% over the best individual model (Devine et al., 2023). Another study demonstrated the effectiveness of an adaptive classifier ensemble method based on spatial perception, which enhances both the performance and diversity of the ensemble members (Figure 1) (Xu et al., 2021). These approaches leverage the strengths of different models, making them robust against the challenges posed by high-dimensional datasets.

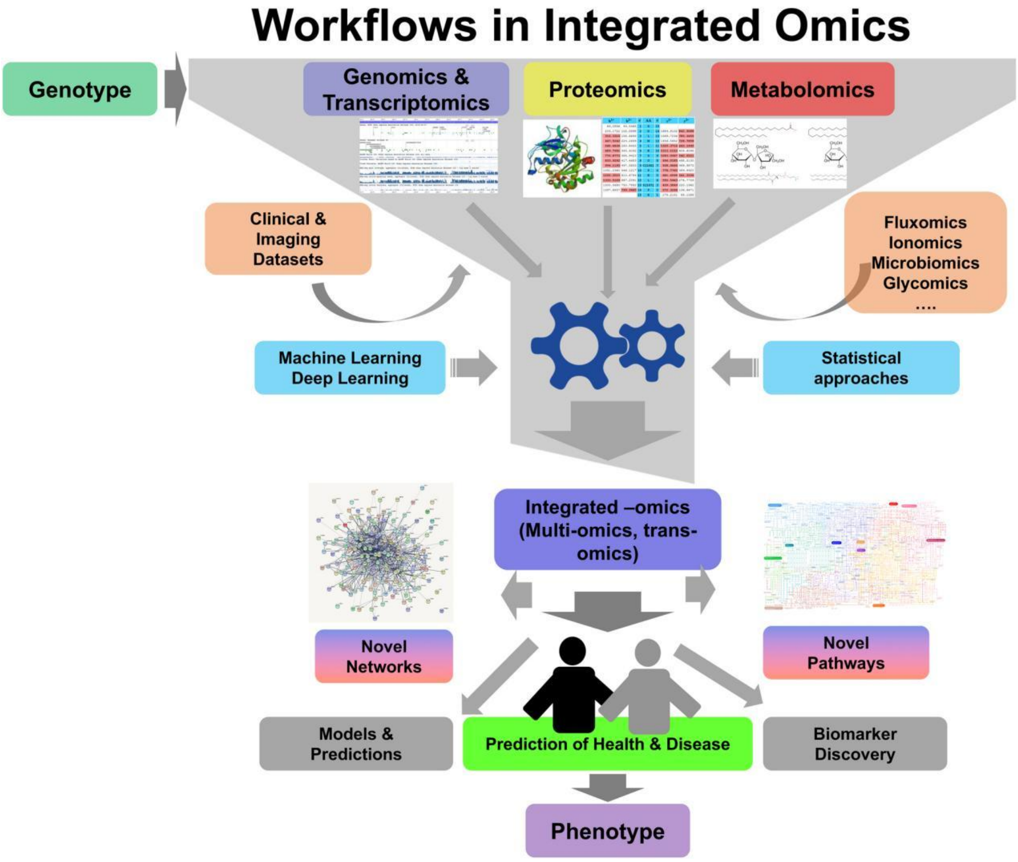
Workflows in Integrated Omics
Genotype
Genomics & Transcriptomics
Proteomics
Metabolomics
Clinical & Imaging Datasets
Fluxomics
Ionomics
Microbiomics
Glycomics
....
Machine Learning Deep Learning
Statistical approaches
Integrated –omics (Multi-omics, trans-omics)
Novel Networks
Novel Pathways
Models & Predictions
Prediction of Health & Disease
Biomarker Discovery
Phenotype

Figure 1 A typical integrated omics workflow showing input datasets, output datasets and results (Adopted from Xu et al., 2021)

## 5 Innovations in High-Dimensional Data Analysis

### 5.1 Machine learning and AI-based methods

5.1.1 Supervised vs. unsupervised learning

Supervised learning involves training a model on labeled data, which is particularly useful for classification and regression tasks. In contrast, unsupervised learning does not require labeled data and is often used for clustering and dimensionality reduction. For instance, unsupervised learning techniques like clustering and dimensionality reduction have been applied to high-dimensional single-cell imaging data to reduce computational workload while maintaining accuracy (Peralta and Saeys, 2020). Additionally, unsupervised and semi-supervised learning methods are gaining traction in plant systems biology, where labeled data is scarce (Yan and Wang, 2022).

5.1.2 Deep learning for high-dimensional data

Deep learning has shown remarkable success in handling high-dimensional data, particularly in fields like neuroimaging and bioinformatics. For example, deep learning approaches have been used for the diagnostic classification of Alzheimer's disease using neuroimaging data, achieving high accuracy rates (Jo et al., 2019). Similarly, deep learning-based clustering methods have been employed in bioinformatics to analyze high-dimensional data such as gene expressions and biomedical texts (Figure 2), leading to improved clustering results (Karim et al., 2020).

5.1.3 Hybrid statistical and AI approaches

Hybrid approaches that combine traditional statistical methods with AI techniques are emerging as powerful tools for high-dimensional data analysis. A novel hybrid method using a weighted-chaotic salp swarm algorithm (WCSSA) combined with a kernel extreme learning machine (KELM) classifier has been proposed for microarray classification, demonstrating higher classification accuracy and significant gene reduction (Baliarsingh et al., 2019). Another example is the combination of deep belief networks (DBNs) with one-class support vector machines (SVMs) for anomaly detection in high-dimensional spaces, which has shown to be both scalable and computationally efficient (Erfani et al., 2016).

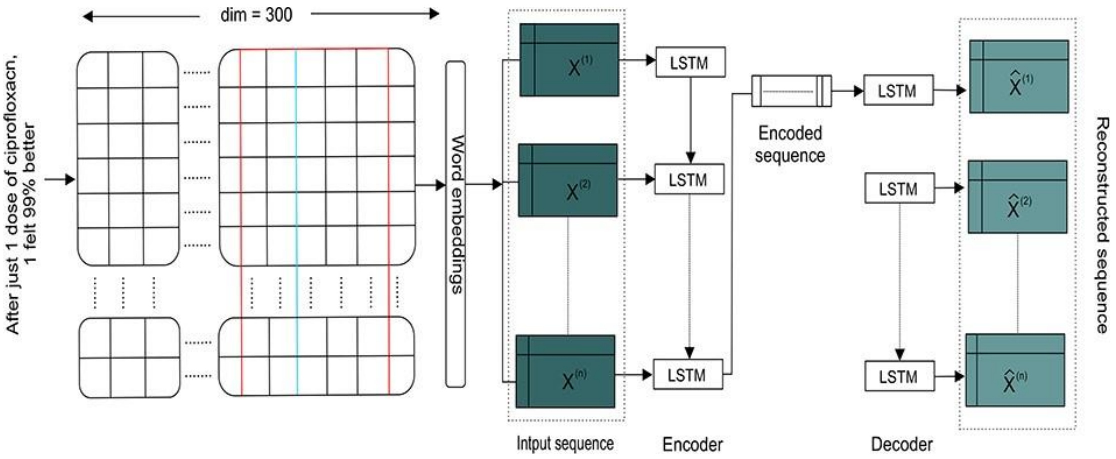

Figure 2 Schematic representation of the LSTM-AE, used for biomedical text clustering, where individual drug review texts are embedded using word2vec before feeding as a sequence (Adopted from Karim et al., 2020)

### 5.2 Advanced visualization techniques

Advanced visualization techniques are essential for interpreting high-dimensional data. Techniques such as tensor-based representations and hierarchical clustering have been used to visualize and organize large-scale molecular biophysics data, providing insights into the intrinsic multiscale structure of the data (Ramanathan et al., 2015). These visualization methods help in understanding complex datasets and identifying patterns that may not be apparent through traditional analysis methods.

### 5.3 Integration of multi-omics data

The integration of multi-omics data is crucial for a comprehensive understanding of complex biological systems. Knowledge-guided statistical learning methods that incorporate biological knowledge, such as functional genomics and proteomics, have been developed to improve prediction and classification accuracy in precision oncology (Zhao et al., 2019). These methods enable the analysis of multifactorial diseases like cancer by aggregating weak signals from individual genes into stronger pathway-level signals, making it easier to detect significant changes.

## 6 Case Studies of Biostatistical Applications

### 6.1 Genomics and transcriptomics data analysis

The analysis of high-dimensional genomics and transcriptomics data presents unique challenges and opportunities in the field of biostatistics. High-throughput techniques have enabled the rapid generation of vast amounts of data, necessitating advanced methods for effective analysis and integration. One significant approach is the quantitative analysis and integration of multi-omics data, which includes genomics, transcriptomics, proteomics, and metabolomics (Ding et al., 2024). This multi-omics approach provides a comprehensive perspective on biological systems, but it also introduces challenges related to data heterogeneity and batch effects. Methods such as network analysis and biological contextualization are employed to address these challenges and enhance the understanding of complex biological relationships (Misra et al., 2019; Wörheide et al., 2021).

In precision oncology, knowledge-guided statistical learning methods have been developed to improve the analysis of high-dimensional -omics data. These methods incorporate biological knowledge, such as functional genomics and proteomics, to enhance prediction and classification accuracy. This approach is particularly useful in identifying weak signals in important pathways, which can be aggregated to detect stronger signals and yield biologically interpretable results (Kaur et al., 2021).

Another innovative method is the multi-objective chaotic emperor penguin optimization (MOCEPO) algorithm, which is designed for feature selection and cancer classification in high-dimensional genomics data. This algorithm aims to minimize the number of selected genes while maximizing classification accuracy, demonstrating superior performance compared to existing methods (Zhao et al., 2019).

Bias correction and data integration are also critical in high-dimensional genomic data analysis. The MANCIE method, for example, uses a Bayesian-supported principal component analysis-based approach to improve consistency between sample-wise distances in different genomic profiles. This method has shown effectiveness in various applications, including tissue-specific clustering and prognostic prediction (Kalina, 2014).

### 6.2 Imaging and multi-modal biological data

The integration of imaging data with other biological data types, such as genomics and transcriptomics, is another area where biostatistical methods are crucial. The rapid increase in data dimension and acquisition rate from technologies like genomics and imaging challenges conventional analysis strategies. Modern machine learning methods, such as deep learning, have shown promise in leveraging large datasets to uncover hidden structures and make accurate predictions. These methods are particularly useful in regulatory genomics and cellular imaging, providing new insights into biological processes and diseases (Mirza et al., 2019).

Integrative analysis of multi-modal biological data, including gene expression data, is essential for identifying biomarkers and understanding complex biological systems. Methods like the Multi-View based Integrative Analysis of microarray data (MVIAm) address challenges such as high noise, small sample size, and batch effects. MVIAm applies cross-platform normalization and robust learning mechanisms to integrate multiple datasets, enhancing the identification of significant biomarkers in cancer classification problems (Ma and Dai, 2011).

## 7 Challenges in Implementing High-Dimensional Analysis

### 7.1 Computational and data storage constraints

High-dimensional data analysis often involves handling massive datasets that can reach tera- to peta-byte sizes, especially in fields like genomics, transcriptomics, proteomics, and metabolomics. This sheer volume of data presents significant computational and storage challenges. Traditional data storage solutions and computational frameworks may not be sufficient to manage such large datasets efficiently. For instance, the integration of multi-omics data requires substantial computational power and advanced data storage solutions to handle the diverse and voluminous data types (Figure 3) (Misra et al., 2019). Additionally, the scalability of computational methods is a critical issue, as the complexity of data increases exponentially with the number of dimensions (Fan et al., 2013).

### 7.2 Scalability of analytical methods

The scalability of analytical methods is another major challenge in high-dimensional data analysis. Standard multivariate statistical methods often fail when applied to high-dimensional datasets due to the curse of dimensionality. This issue necessitates the development of new classification and dimension reduction methods that can handle the increased complexity and size of the data (Kalina, 2014). Machine learning techniques, such as those used in integrative analysis of multi-omics data, must be specifically designed to address scalability issues, ensuring that they can process large datasets efficiently without compromising accuracy (Mirza et al., 2019). Moreover, the need for scalable visualization tools that can intuitively represent high-dimensional data structures is crucial for deriving meaningful insights (Moon et al., 2019).

### 7.3 Data privacy and security concerns

Data privacy and security are paramount when dealing with high-dimensional datasets, particularly in sensitive fields like healthcare and genomics. The integration and sharing of large-scale biomedical data pose significant risks to patient confidentiality and data integrity. Ensuring robust data privacy measures and secure data-sharing infrastructures is essential to protect sensitive information from unauthorized access and breaches. Additionally, the development of standardized benchmarking metrics and data-sharing protocols can help mitigate these concerns by providing a secure framework for data exchange and analysis (Atta and Fan, 2021).

## 8 Concluding Remarks

High-dimensional data analysis presents numerous biostatistical challenges, particularly in the context of computational biology and genomics. Strategies to address these challenges include the development of advanced LASSO methods such as Hi-LASSO, which improves prediction and feature selection by addressing multicollinearity and providing statistical significance tests. Additionally, causal discovery algorithms have been enhanced through empirical Bayes approaches to manage high dimensionality and multicollinearity, enabling the identification of latent factors and biologically meaningful pathways. Sparse estimation strategies in linear mixed models and the integration of random forest techniques have also been proposed to handle multicollinearity and improve regression parameter estimation in high-dimensional settings.

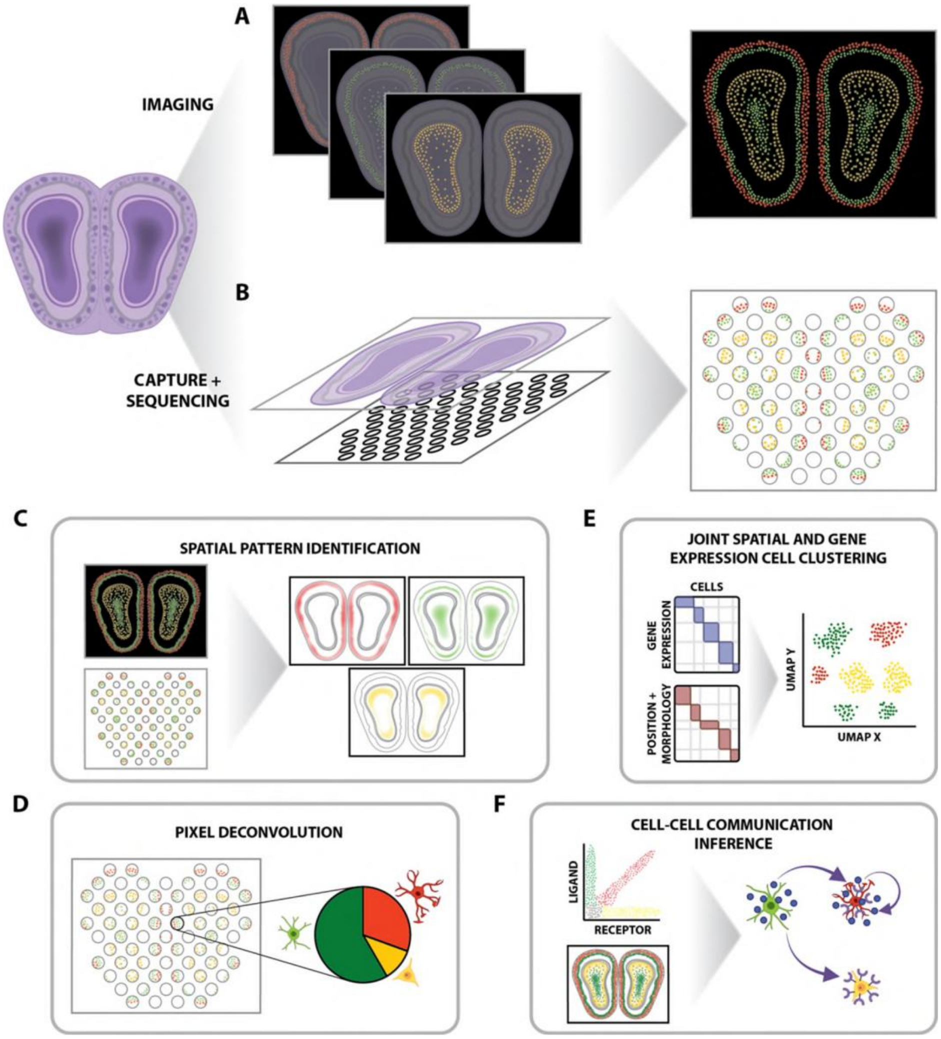

Figure 3 High-throughput spatially resolved transcriptomics data acquisition and analysis (Adopted from Misra et al., 2019)
Image caption: A Imaging-based, targeted, in situ transcriptomic profiling at molecular and single-cell resolution or B non-targeted, RNA capture, and sequencing at pixel resolution is used to measure RNA in tissues in a spatially resolved manner. Computational methods can be used to C identify genes with significantly spatially variable expression patterns, D deconvolve multi-cellular pixel-resolution data to determine pixel cell-type composition, E combine gene expression, position, and morphological information to cluster cell populations, or F identify spatially informed putative cell-cell communication networks (Adopted from Misra et al., 2019)

Future research in high-dimensional data analysis is likely to be driven by several key innovations. The Hi-LASSO method, with its refined importance scores and global oracle property, sets a new standard for feature selection and prediction accuracy in extremely high-dimensional datasets. The application of non-orthogonal empirical Bayes approaches to causal discovery algorithms represents a significant advancement, allowing for the extraction of interpretable latent factors from complex datasets. Additionally, the integration of machine learning techniques such as random forests with traditional regression models offers a promising avenue for improving model estimation in the presence of multicollinearity. The development of new data integration strategies for single-cell multimodal assays will also be crucial, as these methods enable the association of chromatin accessibility and genetic variation with transcription, providing deeper insights into cellular heterogeneity.

Researchers and data scientists working with high-dimensional data should prioritize the use of advanced LASSO methods, such as Hi-LASSO, to enhance feature selection and prediction accuracy while addressing multicollinearity issues. It is also recommended to explore empirical Bayes approaches for causal discovery to manage high dimensionality and extract meaningful biological insights. Employing sparse estimation strategies and integrating machine learning techniques like random forests can further improve model performance in high-dimensional settings. Additionally, stability selection methods should be considered to control false discoveries and ensure the reliability of variable selection. Finally, researchers should stay abreast of developments in single-cell data integration techniques to leverage the full potential of multimodal assays in understanding cellular heterogeneity.

## Acknowledgments

I would like to thank Professor Jin for his guidance and support throughout the entire research process. I also thank the anonymous reviewer for their feedback.

## Conflict of Interest Disclosure

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

## References

Alzubaidi A., 2018, Challenges in developing prediction models for multi-modal high-throughput biomedical data, Springer International Publishing, 2019: 1056-1069.
https://doi.org/10.1007/978-3-030-01054-6_73

Amezquita R., Lun A., Becht E., Carey V., Carpp L., Geistlinger L., Marini F., Rue-Albrecht K., Risso D., Soneson C., Waldron L., Pagès H., Smith M., Huber W., Morgan M., Gottardo R., and Hicks S., 2019, Orchestrating single-cell analysis with Bioconductor, Nature Methods, 17: 137-145.
https://doi.org/10.1038/s41592-019-0654-x

Argelaguet R., Cuomo A., Stegle O., and Marioni J., 2021, Computational principles and challenges in single-cell data integration, Nature Biotechnology, 39: 1202-1215.
https://doi.org/10.1038/s41587-021-00895-7

Ashraf M., Anowar F., Setu J., Chowdhury A., Ahmed E., Islam A., and Al-Mamun A., 2023, A survey on dimensionality reduction techniques for time-series data, IEEE Access, 11: 42909-42923.
https://doi.org/10.1109/ACCESS.2023.3269693

Atta L., and Fan J., 2021, Computational challenges and opportunities in spatially resolved transcriptomic data analysis, Nature Communications, 12(1): 5283.
https://doi.org/10.1038/s41467-021-25557-9

Baliarsingh S.K., Vipsita S., Muhammad K., Dash B., and Bakshi S., 2019, Analysis of high-dimensional genomic data employing a novel bio-inspired algorithm, Appl. Soft Comput., 77: 520-532.
https://doi.org/10.1016/J.ASOC.2019.01.007

Davis-Turak J., Courtney S.M., Hazard E.S., Glen W.B., Silveira W.A., Wesselman T., Harbin L., Wolf B., Chung D., and Hardiman G., 2017, Genomics pipelines and data integration: challenges and opportunities in the research setting, Expert Review of Molecular Diagnostics, 17: 225-237.
https://doi.org/10.1080/14737159.2017.1282822

Devine J., Kurki H.K., Epp J.R., Gonzalez P.N., Claes P., and Hallgrímsson B., 2023, Classifying high-dimensional phenotypes with ensemble learning, bioRxiv, 2023.
https://doi.org/10.1101/2023.05.29.542750

Ding D.Y., 2024, The role and challenges of genome-wide association studies in revealing crop genetic diversity, Bioscience Method, 14(1): 8-19.
https://doi.org/10.5376/bm.2024.15.0002

Erfani S.M., Rajasegarar S., Karunasekera S., and Leckie C., 2016, High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning, Pattern Recognit, 58: 121-134.
https://doi.org/10.1016/j.patcog.2016.03.028

Fan J.Q., Han F., and Liu H., 2013, Challenges of big data analysis, National Science Review, 1(2): 293-314.
https://doi.org/10.1093/nsr/nwt032

Gangavarapu T., and Patil N., 2019, A novel filter-wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets, Applied Soft Computing, 81: 105538.
https://doi.org/10.1016/J.ASOC.2019.105538

He S., Ye X., Sakurai T., and Zou Q., 2023, MRMD3.0: A python tool and webserver for dimensionality reduction and data visualization via an ensemble strategy, Journal of Molecular Biology, 435(14): 168116.
https://doi.org/10.2139/ssrn.4258941

Imoto Y., Nakamura T., Escolar E.G., Yoshiwaki M., Kojima Y., Yabuta Y., Katou Y., Yamamoto T., Hiraoka Y., and Saitou M., 2022, Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis, Life Science Alliance, 5(12).
https://doi.org/10.26508/lsa.202201591

Jiang X.Y., Kong X.Y., and Ge Z.Q., 2023, Augmented industrial data-driven modeling under the curse of dimensionality, IEEE/CAA Journal of Automatica Sinica, 10(6): 1445-1461.
https://doi.org/10.1109/JAS.2023.123396

Jo T., Nho K., and Saykin A.J., 2019, Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data, Frontiers in Aging Neuroscience, 11: 220.
https://doi.org/10.3389/fnagi.2019.00220

Juan H.F., and Huang H.C., 2023, Quantitative analysis of high-throughput biological data, Wiley Interdisciplinary Reviews: Computational Molecular Science, 13(4): e1658.
https://doi.org/10.1002/wcms.1658

Kalina J., 2014, Classification methods for high-dimensional genetic data, Biocybernetics and Biomedical Engineering, 34: 10-18.
https://doi.org/10.1016/J.BBE.2013.09.007

Karim M.R., Beyan O., Zappa A., Costa I.G., Rebholz-Schuhmann D., Cochez M., and Decker S., 2020, Deep learning-based clustering approaches for bioinformatics, Briefings in Bioinformatics, 22(1): 393-415.
https://doi.org/10.1093/bib/bbz170

Kaur P., Singh A., and Chana I., 2021, Computational techniques and tools for omics data analysis: state-of-the-art challenges and future directions, Archives of Computational Methods in Engineering, 28: 4595-4631.
https://doi.org/10.1007/s11831-021-09547-0

Kim Y.S., Hao J., Mallavarapu T., Park J., and Kang M., 2019, Hi-lasso: high-dimensional lasso, IEEE Access, 7: 44562-44573.
https://doi.org/10.1109/ACCESS.2019.2909071

Leonavicius K., Nainys J., Kučiauskas D., and Mazutis L., 2019, Multi-omics at single-cell resolution: comparison of experimental and data fusion approaches, Current Opinion In Biotechnology, 55: 159-166.
https://doi.org/10.1016/j.copbio.2018.09.012

Ma S., and Dai Y., 2011, Principal component analysis based methods in bioinformatics studies, Briefings in bioinformatics, 12(6): 714-722.
https://doi.org/10.1093/bib/bbq090

Mirza B., Wang W., Wang J., Choi H., Chung N.C., and Ping P.P., 2019, Machine learning and integrative analysis of biomedical big data, Genes, 10(2): 87.
https://doi.org/10.3390/genes10020087

Moon K.R., Dijk D., Wang Z., Gigante S., Burkhardt D.B., Chen W.S., Yim K., Elzen A., Hirn M.J., Coifman R.R., Ivanova N.B., Wolf G., and Krishnaswamy S., 2019, Visualizing structure and transitions in high-dimensional biological data, Nature Biotechnology, 37(12): 1482-1492.
https://doi.org/10.1038/s41587-019-0336-3

Palit S., Heuser C., Almeida G.P., Theis F.J., and Zielinski C., 2019, Meeting the challenges of high-dimensional single-cell data analysis in immunology, Frontiers in Immunology, 10: 1515.
https://doi.org/10.3389/fimmu.2019.01515

Peralta D., and Saeys Y., 2020, Robust unsupervised dimensionality reduction based on feature clustering for single-cell imaging data, Appl.Soft Comput, 93: 106421.
https://doi.org/10.1016/j.asoc.2020.106421

Ramanathan A., Chennubhotla C.S., Agarwal P.K., and Stanley C.B., 2015, Large-scale machine learning approaches for molecular biophysics, Biophysical Journal, 108(2): 370a.
https://doi.org/10.1016/J.BPJ.2014.11.2027

Rocha A., Groen T., Skidmore A., Darvishzadeh R., and Willemen L., 2017, The Naïve Overfitting Index Selection (NOIS): a new method to optimize model complexity for hyperspectral data, Isprs Journal of Photogrammetry and Remote Sensing, 133: 61-74.
https://doi.org/10.1016/J.ISPRSJPRS.2017.09.012

ShyamMohanJ., S., 2016, Data reduction techniques for high dimensional biological data, International Journal of Research in Engineering and Technology, 05: 319-324.
https://doi.org/10.15623/IJRET.2016.0502058

Song X.F., Zhang Y., Guo Y.N., Sun X.Y., and Wang Y.L., 2020, Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data, IEEE Transactions on Evolutionary Computation, 24(5): 882-895.
https://doi.org/10.1109/TEVC.2020.2968743

Viegas F., Rocha L., Gonçalves M., Mourão F., Sá G., Salles T., Andrade G., and Sandin I., 2018, A genetic programming approach for feature selection in highly dimensional skewed data, Neurocomputing, 273: 554-569.
https://doi.org/10.1016/j.neucom.2017.08.050

Vinga S., 2020, Structured sparsity regularization for analyzing high-dimensional omics data, Briefings in bioinformatics, 22(1): 77-87.
https://doi.org/10.1093/bib/bbaa122

Wörheide M., Krumsiek J., Kastenmüller G., and Arnold M., 2021, Multi-omics integration in biomedical research-A metabolomics-centric review, Analytica Chimica Acta, 1141: 144-162.
https://doi.org/10.1016/j.aca.2020.10.038

Xu Y., Yu Z., Cao W., Chen C.L.P., and You J., 2021, Adaptive classifier ensemble method based on spatial perception for high-dimensional data classification, IEEE Transactions on Knowledge and Data Engineering, 33(7): 2847-2862.
https://doi.org/10.1109/TKDE.2019.2961076

Yan C.K., Ma J.J., Luo H.M., and Patel A., 2019, Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets, Chemometrics and Intelligent Laboratory Systems, 184: 102-111.
https://doi.org/10.1016/J.CHEMOLAB.2018.11.010

Yan J., and Wang X.F., 2022, Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology, The Plant Journal : for Cell and Molecular Biology, 111(6): 1527-1538.
https://doi.org/10.1111/tpj.15905

Zhao Y.Z., Chang C., and Long Q., 2019, Knowledge-guided statistical learning methods for analysis of high-dimensional -omics data in precision oncology, JCO Precision Oncology, 3: 1-9.
https://doi.org/10.1200/po.19.00018

## 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.