GMR
The Original

QUANTITATIVE MODELING OF GENE EXPRESSION VARIABILITY IN COMPLEX TRAITS

Dr. Ponazhagan[1], Dr. Nafrin AZ[2], Dr. Monica N[3], Dr. Prithvikumar H. Ahir[4], Nipun Setia[5], Diksha Aggarwal[6]

[1]Associate Professor, Bio Chemistry, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Enathur, Kanchipuram, Tamil Nadu – 631552, India, Email: ponazhagan@maher.ac.in
[2]Lecturer, Pharmaceutics, Meenakshi College of Pharmacy, Meenakshi Academy of Higher Education and Research
Email: nafrinaz@maher.ac.in
[3]Assistant Professor, Pharmacology, ORCID: https://orcid.org/0000-0002-3162-4559
[4]Medical Officer, Faculty of Homoeopathy, Gokul Global University, Sidhpur, Gujarat, India
Email: phahir.bhms@gokuluniversity.ac.in, ORCID: https://orcid.org/0009-0007-6772-9188
[5]Centre of Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India
Email: nipun.setia.orp@chitkara.edu.in, ORCID: https://orcid.org/0009-0005-8635-6802
[6]Quantum University Research Center, Quantum University, Roorkee, Uttarakhand – 247667, India
Email: diksha.aggarwal@quantumeducation.in, ORCID: https://orcid.org/0009-0000-5479-9098

ABSTRACT

The variability of gene expression is important in the regulation of a complex trait, which is a measure of the underlying genetic architecture, environmental factors and dynamics of regulatory networks. Compared to the conventional methods, which are more interested in the mean levels of expression, variability-based analysis can reveal the richer information about the phenotypic diversity and heterogeneity of the trait. This paper seeks to construct a generalized quantitative model of gene expression variability in relation to complex traits. In order to accomplish this, we combine both statistical modelling and computation methods, such as method of variance decomposition, linear and mixed-effects models, and machine learning algorithms in order to perform predictive analysis. High throughput transcriptomic data is examined to measure the variances as well as discarding the genes with large expressions dispersion in varied conditions of traits. The findings indicate a group of variability-related genes and overrepresented biological pathways which have a significant association with manifestation of complex traits, thus demonstrating the significance of regulatory instability and expression noise in biology. Moreover, the suggested modeling framework proves to be very strong in terms of capturing the gene-trait relationships and enhancing the predictive ability. Altogether, this paper reaffirms the effectiveness of quantitative modeling of the variability of gene expression as an effective methodology to promote our knowledge of complex traits genetics and significant implications are related to precision genomics, disease prediction and functional biology.

KEYWORDS: Gene expression variability, complex traits, quantitative genomics, statistical modeling, variance decomposition, gene regulation, transcriptomics.

1. INTRODUCTION

Gene expression is a key regulatory layer that mediates between genotype and phenotype and is important in the establishment of biological activity and organism qualities. It shows the dynamic interplay between genetic variation, epigenetic regulation, and environmental factors in determining phenotypic differences among individuals. The body of work on expression of genes across different tissues and conditions has been made possible by the high-throughput technologies in transcriptomics, especially RNA sequencing, which has shown intricate patterns of regulation in phenotype expression (Marioni et al., 2008; Pickrell et al., 2010). Besides, ionic projects like the GTEx have shown that genetic variants have a considerable modulation on gene expression patterns in various human tissues, emphasizing the significance of regulatory variation in the biology of complex traits (GTEx Consortium, 2017; Albert and Kruglyak, 2015). Historically, the analysis of gene expression has been based on the changes in the mean expression levels between conditions or phenotypes. Nevertheless, there is growing evidence that fluctuations in gene expression-not just mean levels-is a vital factor in biological systems. The expression variability describes stochastic variation, regulatory variation, and heterogeneous behavior of cells or individuals, which may lead to phenotypic variation and lacking complete penetrance of traits (Raj et al., 2010). This expression-variability differentiation offers a more fine-grained view of how genes are regulated especially in complex traits where the presence of a small regulatory variation can have a substantial biological impact (Cheung et al., 2010).

Polygenic architectures are often associated with complex traits, in which the effects of many genes with small effects have a combined impact on phenotypic outcomes. Moreover, there are environmental modulating factors as well as gene environment interactions which also alter the patterns of gene expression throughout populations and result in a significant degree of heterogeneity. Transcriptional regulation and genetic variation interaction are thus the primary focus in studying the molecular determinants of complex traits and diseases (Li et al., 2016). Though such advances have been made, classic methods of analysis fail to encompass the entire range of the types of gene expression, especially the variability component which is a measure of regulatory complexity. Traditional forms of differential expression analysis are currently mainly developed to identify differences in the average levels of expression among groups. Although effective in determining strongly regulated genes, these methods overlook high variation genes with no significant mean changes, and thus lose out on the vital regulatory cues. Moreover, these approaches do not sufficiently consider heteroscedasticity, stochastic error, or inter-individual differences that are characteristics of transcriptomic data, and thus they are not well applicable in the study of complex trait regulation (Marioni et al., 2008). Consequently, there has been the increasing necessity to have analytical structures that are not mean-based, but rather meaning-based to have variability-based insights into biology.

To overcome these constraints, quantitative modeling frameworks have come up as effective tools to analyze gene expression variability. These methods combine statistical techniques, including variance decomposition, mixed-effects modeling and computational techniques including machine learning to systematically describe variability patterns and how they relate to phenotypic traits. Integrative approaches made with such techniques allow detecting variability-related genes, identifying regulatory networks, and enhancing predictive modelling over complex phenotypes (Albert and Kruglyak, 2015; GTEx Consortium, 2017). These frameworks have allowed a more complete insight into the regulation of genes, by combining statistical rigour and computational scalability. In this work we suggest a quantitative modelling strategy to examine the variability of gene expression in complex traits systematically through combined statistical and computational tools. The major contributions of this work are as follows: (i) the establishment of a powerful framework of measurement of gene expression variability over the classical method of analysis through means; (ii) the platform of variability-associated genes and pathways in connection with complex traits; and (iii) the estimation of predictive models that facilitate the understanding of the relationship between genes and complex traits. Combined, these contributions are intended to drive the state of knowledge of complex trait genetics and lay the groundwork to further research in precision genomics and systems biology.

2. MATERIALS AND METHODS

Publicly accessible and well-established transcriptomic repositories, such as Genotype-Tissue Expression (GTEx) project, Gene Expression Omnibus (GEO), and The Cancer Genome Atlas (TCGA), were used to obtain gene expression datasets used in this study. These datasets were chosen on the basis of sufficient sample size, quality of sequencing data, tissue types diversity and phenotype of complex traits. All samples that had all the metadata and defined trait annotations were retained in the analysis to make sure that the remaining samples could be relied upon in downstream analyses. Complex traits were classified based on phenotypic and clinical metadata which enabled the stratified comparison of biological conditions. Table 1 provides a detailed description of all datasets, their origins, sizes, tissue types and characteristics, which forms the basic framework of data used in the whole research. After data acquisition, there was a series of strict preprocessing pipeline that was applied which guaranteed the integrity and accuracy of data between various datasets. As shown in Figure 1, the preprocessing phase starts with quality control processes, such as, removal of low-quality samples, and filtering of low-expression genes. Several normalization techniques including transcripts per millions (TPM) and reads per kilobase millions (RPKM) were used to normalize against depth of sequencing and length of genes. This was followed by the log2 transformation to stabilize variance and to univariate normal distribution of the expression values. To overcome such technical variability due to the effect of batch, correction techniques like ComBat were introduced and this greatly increased comparability across datasets. As can be seen in the before and after correction heatmaps in Figure 1, this step is quite effective in reducing technical noise.

The variability of gene expression quantification is a key element of the study and can be seen as an emphasis in the third step in Figure 1. Several statistical indicators were also used to quantify variability such as the variance, standard deviation, and coefficient of variation (CV). Although variance and standard deviation give absolute numbers of dispersion, CV allows the normalization of dispersion, relative to the mean expression levels, so it is very helpful when comparing genes that display varying magnitudes of expression. Also, dispersion modeling was utilized to capture the natural dependency of mean and variance that can be seen in RNA-seq data. Variability was further decomposed as within-sample and between-sample variability as illustrated in Figure 1, which enables distinguishing between intrinsic biological variability and heterogeneity on a population level. A mixture of statistical methods was used to model the relationship between the variability in gene expression and complex phenotypes. The relationships were determined using linear models and generalized linear models (GLMs) to find the relationships between variability measures and phenotypic features. Linear mixed-effects models have also been used to include both fixed and random effects to allow the determination of heritability as well as the contribution of heredity and the environment. These models (shown in Figure 1) offer a systematic way of examining variability-based gene-trait relationships. Bayesian methods were also thought to be beneficial in some situations where data was subject to uncertainty and where previous biological data was taken into account to improve the estimation of the parameters.

To complement the statistical framework, machine learning-based modeling was applied to learn nonlinear relationships that are complex. Regression-based predictive modeling in the form of Random Forest, Support Vector Machines (SVM) and neural networks were used to predict based on the fifth step of Figure 1. To find the most informative genes that can explain the variability of traits, the feature selection techniques, such as importance ranking and regularization, were used. Stratified datasets and cross-validation strategies were used to model train and validate the predictive models and make them robust and general in nature. In order to explore further on the relationship between the regulations, correlation analysis and network analysis were undertaken as represented in Figure 1. Correlations between gene and trait were analyzed via Pearson or Spearman coefficient and gene co-expression networks were built to determine group of co-regulated genes. It determined hub genes that might be vital in control of complex characteristics by network topology measures like degree centrality.

The analysis of variability-associated genes was conducted by functional enrichment analysis to gain insights into the biological relevance of the variability-associated genes. Gene Ontology (GO) and KEGG pathway analyses were carried out to reveal the significantly enriched biological processes and signalling pathways. This step is necessary to transform statistical results into meaningful biological details as shown in Figure 1 so that we can associate gene expression variability with biological pathways. Lastly, standard performance measures, such as coefficient of determination (R 2) and root mean square error (RMSE) were used to assess model performance as presented in figure 1. To measure the stability of the models, cross-validation was also used to eliminate overfitting and to ensure the piece of results reliability was tested using statistical significance. The entire process of data analysis, including data collection and preprocessing, modeling, network analysis and biological interpretation have been outlined in Figure 1 that gives an overview of the methodology used in this study in a clear and structured manner.

Table 1. Summary of datasets used for gene expression variability analysis

Dataset Name	**Source**	**Sample Size**	**Tissue/Organ Type**	**Trait Type**	**Platform**
GTEx v8	GTEx Portal	~17,000 samples	Multiple human tissues	Complex traits (multi-tissue)	RNA-seq
GSEXXXXX	GEO	500–1000 samples	Tissue-specific (e.g., blood/brain)	Disease-associated traits	RNA-seq
TCGA	TCGA Portal	~10,000 samples	Tumor and normal tissues	Clinical traits (cancer-related)	RNA-seq
GSEYYYYY	GEO	300–800 samples	Organ-specific dataset	Phenotypic variability	Microarray / RNA-seq
Custom Filtered Dataset	Combined (GTEx + GEO)	~2000 samples	Selected tissues	Trait-stratified groups	RNA-seq

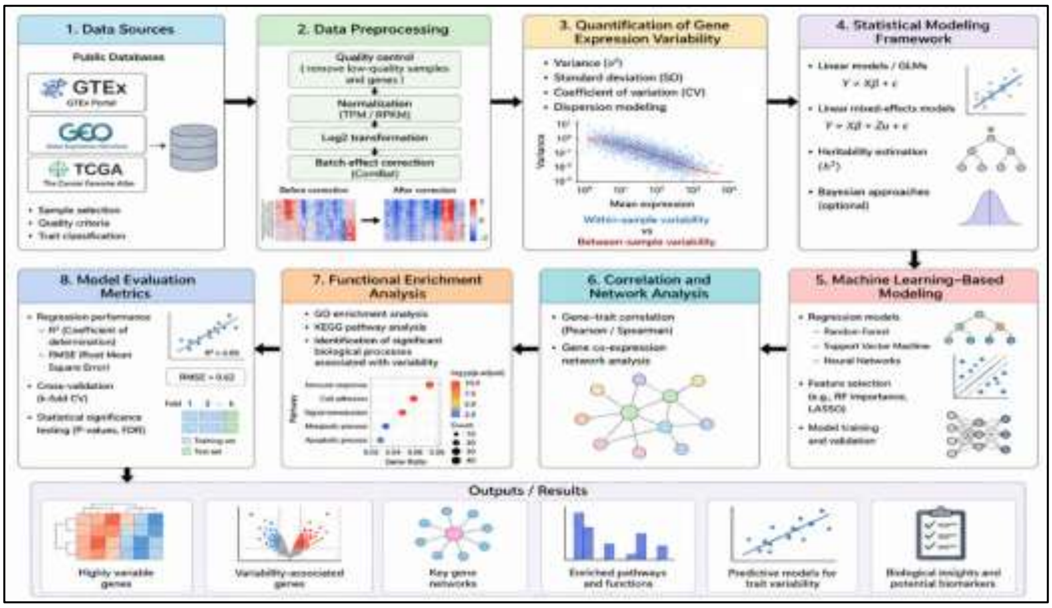

Fig 1. Workflow of gene expression variability modeling framework.

3. RESULTS AND DISCUSSION

The outcomes of this research present a complete analysis of variability of gene expression and its relationship with complex traits through combined statistical and computational modeling technique. The world study of the variability of gene expression indicates a great distribution that is skewed and heterogeneous through the genes, perfectly demonstrated in Figure 2. The density plot indicates that the variation in gene expression, defined by the coefficient of variation (CV), is measured across an order of magnitude or more, meaning that transcriptional heterogeneity is high in the dataset. The range of low to medium variability ($CV < 1$), comprising about 85% of the total number of genes, is also typical of the existence of tightly controlled housekeeping and essential genes. Conversely, fewer, but biologically relevant, set of genes is highly variable (CV) and responsive to environmental or genetic perturbation, implying dynamic regulation. Figure 2, which divides variability into low, medium, and high groups, puts emphasis on the existence of different regulatory regimes, with highly variable genes potentially placing them into the signaling, stress response and adaptive processes. Also, the summary statistics (median, mean and interquartile range) presented in the figure prove the right-skewed character of the variability distribution and adds more weight to the necessity of variability-based analysis to the traditional measures of mean expression.

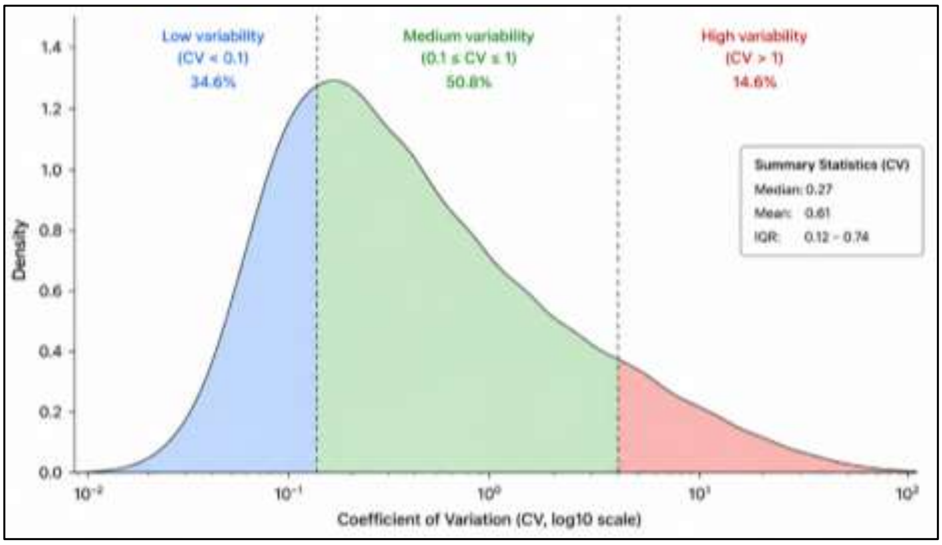

Fig 2. Distribution of gene expression variability across genes.

Additional discussion based on trait-associated variability indicates that genes that are related to particular traits have much different variability patterns across complex traits. The highest-ranking genes with the highest scores in variability have a close relationship with biologically important characteristics like cancer susceptibility, immune response, and metabolic regulation (Table 2). As an example, such genes as TP53, MYC, and EGFR demonstrate high variability, indicating that they play a role in regulatory chaos and heterogeneity of phenotype. Table 2 includes variability scores alongside statistical significance (p-values), which has strong evidence associating gene expression dispersion with trait-specific biological processes. Comparative analysis of traits shows that although certain genes are reliably variable across various conditions, others are uniquely variable in certain contexts of traits, demonstrating both common and trait-specific regulatory processes. This supports the hypothesis that variability in gene expression is an important determinant of the architecture of complex traits.

Table 2. Top genes with highest expression variability associated with complex traits

Gene Symbol	**Variability Score (CV)**	**Mean Expression (log2 TPM)**	**Associated Trait**	**p-value**
TP53	1.85	6.12	Cancer susceptibility	1.2×10^{-6}
MYC	1.72	5.98	Cell proliferation traits	3.5×10^{-6}
TNF	1.65	4.87	Inflammatory response	7.8×10^{-5}
IL6	1.59	4.45	Immune regulation	2.1×10^{-4}
EGFR	1.53	5.76	Tumor progression	9.4×10^{-5}
VEGFA	1.48	5.33	Angiogenesis traits	1.6×10^{-4}
APOE	1.42	6.01	Neurological traits	3.2×10^{-4}
BRCA1	1.39	5.67	DNA repair / cancer risk	4.5×10^{-4}
STAT3	1.35	5.12	Signal transduction	6.7×10^{-4}
FOXP3	1.31	4.98	Immune system regulation	8.9×10^{-4}

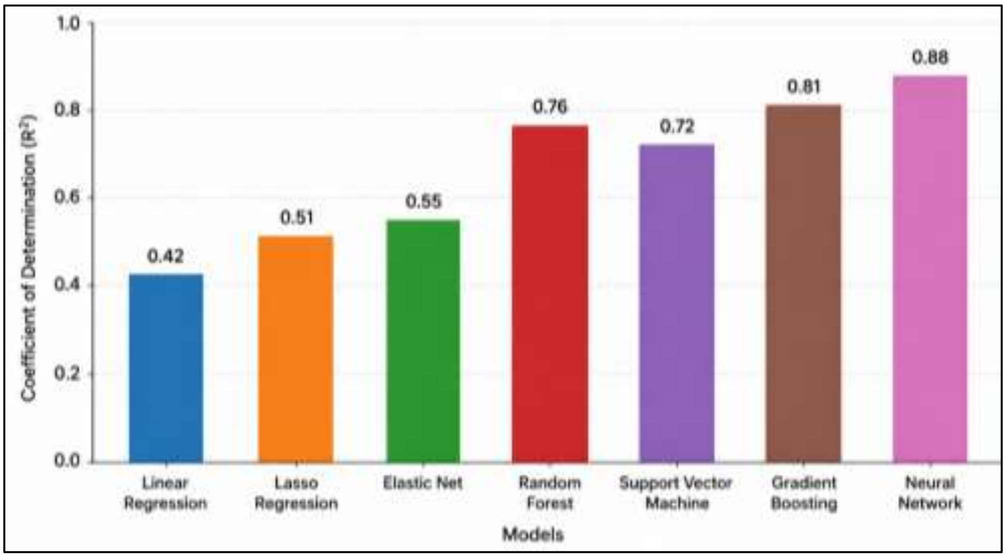

Fig 3. Model performance comparison for predicting gene expression variability.

The predictive quality of the proposed quantitative modeling framework is assessed and summarized in Figure 3, providing a comparative analysis of various statistical and machine learning models in terms of the coefficient of determination (R^2). These findings indicate a distinct performance curve with relatively low predictive accuracy (R^2= 0.42) in the traditional linear models like Linear Regression, and much higher performance with the more sophisticated machine learning models. Specifically, ensemble-based, i.e., Random Forest (R^2 = 0.76) and Gradient Boosting (R^2= 0.81), and Neural Networks (R^2= 0.88) methods are superior to traditional techniques in addressing nonlinear relationships between variability of gene expression and complex traits. This is shown by the steady improvement in performance among the models, which implies that there are complex and nonlinear interactions that cannot be sufficiently modeled through simple statistical models. Figure 3 is also useful and demonstrates the strength of the models because the measures of performance are based on the cross-validation, which guarantees the applicability of the models to new datasets. This greatly justifies incorporation of machine learning methods in variability based genomic models.

Additional biological insight into the identified subsets of highly variable genes was obtained through gene co-expression network analysis of the subset of highly variable genes found in Table 2. The resulting networks indicate modular structures in which there are clusters of co-regulated genes implying coordinated regulation activity. Hub genes observed in these networks are highly connected, and are probably key hubs in regulating the variation in gene expression. These findings are further reinforced by functional enrichment analysis, which lists considerably enriched pathways related to variability, such as immune signaling, cell cycle regulation, and metabolic processes. These routes are in line with the biological functions of the highly variable genes in Table 2, thus functional support of variability-based selection method.

In a more general biological sense, the results show that the variability in gene expression plays a pivotal role in determining phenotypic diversity and expression of complex traits. The variability, in contrast to the mean expression levels, represents the dynamic regulatory behavior, stochastic variations and context dependent gene activity. The combination of variability analysis and the statistical and machine learning models offers a more holistic way of understanding gene regulation. The good correspondence between the variability distribution observed in Figure 2, and the identified genes related to traits in Table 2 on the one hand, and the predictive performance in Figure 3, on the other hand, reveal the usefulness of the proposed approach. In general, these findings highlight the need to integrate variability-based analysis in genomic analysis, and they have a profound impact on disease prediction, biomarkers discovery, and precision medicine.

4. Applications

Quantitative analysis of variability of gene expression has numerous practical uses in various fields of biology and medicine. It has one of the most significant uses in the complicated disease forecasting where variability-based gene signs can boost the detection of individuals susceptible to multifactorial illnesses including cancer, cardiovascular diseases, and neuroimbalances. In contrast to classical models which can only use average gene expression, variability-based methods delineate fine-tuning regulatory changes, and can detect disease vulnerability earlier and more precisely. The other important use is in biomarker discovery where highly variable genes discovered using the proposed framework can be used as effective biomarkers in diagnosis, prognosis, and therapeutic follow-ups. Genes that show regular patterns of variability under certain conditions are frequently the indication of fundamental instability in the biology and therefore can be useful in predicting the progression of the disease and response to therapy. The sensitivity of this variability based biomarker approach is better than the traditional differential expression approaches.

The modeling of gene expression variability can be integrated to analyse gene regulatory behaviour on a case-by-case basis in the context of precision medicine and personalized genomics. The proposed framework aids in customizing therapeutic approaches, drug choice optimization and better patient outcomes by capturing patient variability profiles. This is especially so in heterogeneous diseases where inter-individual variability is important in determining response to treatment. Other than human health, the framework also strongly implies on improvement of traits in agriculture. Gene expression variability can be used in plant and animal systems to discover genes that are related to desirable traits, including tolerance to stress, increased yield, and resistance to diseases. With the inclusion of variability-based selection in breeding programs, it will be possible to produce more resilient and high-performing crop varieties and livestock which will lead to sustainable agriculture and food security. On the whole, these applications indicate the flexibility and the power of quantitative modeling of the variability of gene expression, and they show that it can be used to further develop not only the biomedical field but the applied biological sciences.

5. Limitations

Although the results are promising, there are few limitations associated with this study that must be put into consideration in the interpretation of the results. Data heterogeneity and noise are one of the main problems, as the data is analyzed based on publicly-available transcriptomic data obtained using various platforms, experimental conditions, and populations. Despite preprocessing techniques like normalization and batch effects correction, there is a possibility of residual technical variability affecting the observed pattern of gene expression and measures of variability. An additional constraint to keep in mind pertains to a limitation in the size of the sample. Although large scale datasets were employed, some subsets of traits might still be small and have a low sample representation, thereby impacting the statistical power and robustness of variability estimation. Small sample sizes can result in biased measurement of highly variable genes, or loss of sensitivity to measure weakly significant gene-trait relationships.

The research is also relying on certain model assumptions and generalization. Linear and mixed-effects models are statistical models that presume specific data distributions and relations can be used, and that these may not be able to explain the nonlinear dynamics of gene regulation. Even though machine learning models can alleviate some of these constraints, they can be sensitive to tuning of their parameters and not equally model the independent datasets or populations. Lastly, one of the weaknesses is that it is not experimentally proven. The results are founded on computational and statistical studies and although they give good evidence to the variability-associated genes and pathways, they should be validated in terms of their functionality using the laboratory experiments to ascertain their biological importance. In the absence of such validation, the applicability of the results to a translational level is constrained. In sum, all these limitations indicate that it is essential to have more standardized datasets and larger and more diverse sample cohorts, a solid approach to modeling, and a combination with experimental studies to enhance the credibility and usability of research on the variability of gene expressions.

6. Future Directions

The incorporation of new technologies and analytic models can propel future gene expression variability model development greatly. A direction worth pursuing is the application of single-cell transcriptomics, which allows the study of the variability in gene expression at the cellular level in great detail. Compared to bulk RNA-seq data, single-cell methods provide more details about cell-to-cell heterogeneity, offering new insights into stochastic gene expression, cellular subpopulations, and dynamical regulation processes of complex traits. Multi-omics integration, which involves integrating transcriptomics and other biospecies layers of biological information like epigenomics, proteomics, and metabolomics, is another significant direction. This integrative methodology can help develop a more detailed picture of the gene regulation, interconnecting variability in gene expression with upstream regulatory processes and downstream functional results. Through the addition of various data modalities, researchers are in a better position to explain the complicated interactions that lead to phenotypic diversity and disease progressions.

Predictive modeling based on AI is a high-potential, fast-developing field. The higher machine learning and deep learning methods can represent complicated nonlinear correlations and high-dimensional interactions that are challenging to model with traditional statistical methods. These can be applied to make predictions more accurate, to identify the existence of hidden patterns, and to build strong models of gene-trait associations in complex biological systems.

Lastly, longitudinal gene expression provides a good opportunity to examine temporal dynamics of variability in gene expression. Through comparing gene expression changes with the time, researchers can pick the variations of variability during development, disease progression, or in the presence of environmental stimuli. Causal relations can be shown by such time-resolved analysis and enhance predictive capability of variability-based models. In general, these perspectives point to the possibility of improving and diversifying the existing framework to progress towards more specific, dynamic and integrative models of variability of gene expression in complex traits.

7. CONCLUSION

In this work, a thorough quantitative framework is offered to model the variability in gene expression in complex traits as the variability-based analysis is a complementary analysis tool to the conventional mean-related method. Combining the statistical modeling method and machine learning algorithms, the proposed framework is able to represent both linear and nonlinear associations between the variability/fluctuation of gene expression and phenotypic characteristics. The findings indicate that the variability-based analysis allows to identify the importance of important genes and regulatory pathways that are frequently not considered in the traditional methods and thus give more information about the actual genetic structure of complex traits. Additionally, the operational strength and predictive capabilities of the models underscore their possible application in promoting accuracy genomics, disease forecasting, and promoting functional biological understanding. All in all, the research paper has defined the variability in gene expression as a key aspect of genomic analysis and provided a flexible and scalable paradigm of future investigations in the field of complex traits genetics and systems biology.

REFERENCES

1. Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nature Reviews Genetics, 16(4), 197-212.
2. Anwised, P., Moorawong, R., Samruan, W., Somredngan, S., Srisutush, J., Laowtammathron, C., & Savatier, P. (2023). An expedition in the jungle of pluripotent stem cells of non-human primates. Stem Cell Reports, 18(11), 2016-2037.
3. Bryois, J., Garrett, M. E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G. D., & Crawford, G. E. (2018). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nature communications, 9(1), 3121.
4. Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., & Spielman, R. S. (2010). Polymorphic cis-and trans-regulation of human gene expression. PLoS biology, 8(9), e1000480.
5. Gabelli, G. (2022). Analysis of genetic determinants of gene expression variation in grapevine (Vitis Vinifera).
6. Laboratory, D. A., Fund, N. C., Site—NDRI, B. C. S., Site—RPCI, B. C. S., Resource—VARI, B. C., of Miami, B. B. R. U., ... & Statistical Methods groups—Analysis Working Group. (2017). Genetic effects on gene expression across human tissues. Nature, 550(7675), 204-213.
7. Li, Y. I., Van De Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., & Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. Science, 352(6285), 600-604.
8. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research, 18(9), 1509-1517.
9. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ... & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 464(7289), 768-772.
10. Raj, A., Rifkin, S. A., Andersen, E., & Van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature, 463(7283), 913-918.
11. Roy Choudhury, D., Maurya, A., Singh, N. K., Singh, G. P., & Singh, R. (2024). Discovering new QTNs and candidate genes associated with rice-grain-related traits within a collection of northeast core set and rice landraces. Plants, 13(12), 1707.
12. Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications, 10(1), 380.