GMR
The Original

HIGH-THROUGHPUT SEQUENCING-BASED IDENTIFICATION AND FUNCTIONAL ANNOTATION OF GENETIC VARIANTS ASSOCIATED WITH PHENOTYPIC TRAITS

Bhavani Ganapathy[1], Mrs. Dhivya N[2], Roshini B[3], Dr. Jayvadan Vaishnav[4], Damanjeet Aulakh[5], Ms. Rucha N. Acharya[6], Mr. S. Mohan[7], Dr. T. Premraj[8]

[1]Associate Professor, Department of Pharmacology, Meenakshi Ammal Dental College and Hospital, Meenakshi Academy of Higher Education and Research
[2]Assistant Professor, Department of Medical Surgical Nursing, Kasturba Gandhi Nursing College, Sri Balaji Vidyapeeth (Deemed to be University), Puducherry, India, Email: ndhivya0710@gmail.com, ORCID: https://orcid.org/0000-0003-0426-0769
[3]Assistant Professor, Meenakshi College of Allied Health Sciences, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research
[4]Assistant Professor, Department of Life Sciences, Parul Institute of Applied Sciences, Parul University, Vadodara, Gujarat, India
Email: jayvadan.vaishnav27033@paruluniversity.ac.in, ORCID: https://orcid.org/0000-0003-2083-403X
[5]Centre of Research Impact and Outcome, Chitkara University, Rajpura – 140417, Punjab, India,
Email: damanjeet.aulakh.orp@chitkara.edu.in, ORCID: https://orcid.org/0009-0009-4840-8228
[6]Assistant Professor, Faculty of Allied and Healthcare, Gokul Global University, Sidhpur, Gujarat, India,
Email: rnacharya.gpc@gokuluniversity.ac.in, ORCID: https://orcid.org/0009-0008-1720-3660
[7]Department of Biochemistry, Aarupadai Veedu Medical College and Hospital, Vinayaka Missions Research Foundation (Deemed to be University), Puducherry, India, Email: mohan.sivanandham@avmc.edu.in, ORCID: https://orcid.org/0000-0001-6905-1023
[8]Assistant Professor, Radiodiagnosis, Sree Balaji Medical College and Hospital, Bharath Institute of Higher Education and Research
ORCID: https://orcid.org/0009-0003-2268-2735

ABSTRACT

High-throughput sequencing has contributed greatly to the discovery of genetic variants and their contribution to phenotypic variation. This paper will seek to identify and annotate genetic variants related to phenotypic traits in Saccharomyces cerevisiae in a systematic and functional way based on an extensive bioinformatics platform. Whole-genome sequencing data of various yeast strains were processed with the help of an integrated pipeline consisting of quality control, sequence alignment, variant calling, and functional annotation. Large amounts of single nucleotide polymorphisms (SNPs) and insertion/deletions (INDELs) were identified in the entire genome with a greater percentage found in non-coding and regulatory regions. The functional classification showed that a considerable number of variants led to non-synonymous mutations that had an impact on the genes related to metabolic pathways and stress response systems. The statistical analysis of association has revealed important variants that were associated with phenotypic phenotypes such as growth rate and stress tolerance to the environment, which suggests a polygenic and complex genetic framework. The results also indicate that the regulatory and rare variants contribute to the variability in phenotypes. On the whole, this paper has shown that high-throughput sequencing together with functional annotation is effective in decoding the genotypephenotype relationships and a scalable model of genomic analysis is possible with model organisms and beyond.

KEYWORDS: High-throughput sequencing, genetic variants, SNPs, INDELs, functional annotation, genotype–phenotype association, Saccharomyces cerevisiae, bioinformatics pipeline, genome analysis, variant discovery

1. INTRODUCTION

The molecular basis of phenotypic variation continues to be a basic question in the field of genetics and genomics. The phenotypic traits are determined by a large range of genetic variants, such as single nucleotide polymorphism (SNP), insertions/deletions, and regulatory variants, that have a cumulative effect on the visible traits in populations. The comprehensive genomic studies have given important understanding about the evolutionary process, environmental adaptation and relationship between the genotype and phenotype in Saccharomyces cerevisiae (Bai et al., 2022; Peter et al., 2018). Nevertheless, genetic architecture of complex traits tends to be regulated by complex interactions between multiple loci such as additive, dominant, and epistatic effects with interpretation being extremely difficult (Matsui et al., 2022; Nguyen Ba et al., 2022). Recent research has emphasized that rare and low-frequency variants can greatly contribute to phenotypic diversity and have commonly contributed a large share of unaccounted variability in complex phenotypes (Bloom et al., 2019; Fournier et al., 2019). Furthermore, the genetic background effects and regulatory differences also complicate genotype–phenotype mapping; a single variant can have diverse phenotypic consequences, based on their genomic conditions (Galardini et al., 2019; Jakobson et al., 2019). The GWAS and quantitative trait locus (QTL) mapping have progressed in terms of understanding the variants that are associated with traits, but they tend to be incapable of capturing the underlying biological mechanisms due to their weakness in functional interpretation (Sardi et al.,

2018; Maclean et al., 2017). In spite of these developments, the current studies do not contain a lot of integrated frameworks that join the high-throughput variant discovery with thorough annotation of functions and biological interpretation. Although the gene-level effects (Sirr et al., 2018) and functional validation of variants (Sharon et al., 2018) have been investigated individually, a more comprehensive approach to correlating large-scale genomic variation with phenotypic traits has not been explored. The lack of this leads to a critical need to have powerful analytical pipelines that can be used to combine the sequencing data with functional genomics information.
This research paper presents a high-throughput sequencing-based model of identifying and functionally annotating genetic variants relating to phenotypic characters in Saccharomyces cerevisiae. This study will combine variant identification and functional classification and phenotypic association analysis to gain us a complete picture of the genetic factors of complex phenotypes and enhance the development of functional genomics studies.

2. RELATED WORK

There has been a lot of success in trying to explain the genetic basis of the phenotype variation in Saccharomyces cerevisiae using a wide range of genomic and functional methods. Quantitative trait locus (QTL) mapping has successfully been applied to the large-scale finding of genomic regions that are related to complex traits. As an example, Eder et al. (2018) have shown the use of QTL analysis to determine the relationship between genetic variations and fermentation-related phenotypes, whereas Wang et al. (2019) have revealed the identification of key genomic loci that are linked to high-temperature fermentation performance of industrial-use yeast strains. Likewise, Haas et al. (2019) examined the ethanol tolerance and found that there is a significant genetic variation that leads to stress response characteristics. All these studies point to the usefulness of QTL-based methodologies in the discovery of genotype-phenotype relations. In addition to the mapping using locus, regulatory variation has become a very important factor that affects the phenotypic consequences. The study carried out by Renganaath et al. (2020) reported cis-regulatory variants that led to varying gene expression whereas Duveau et al. (2021) examined how the trans-regulatory mutations affect the variability of gene expression. Besides that, Shih et al. (2021) have shown that cis-regulatory variants have a significant impact on gene expression changes in response to different environmental conditions. These results underline the value of regulation processes in defining the diversity of phenotype and the shortcoming of methods, which concentrate on coding variants only. With recent improvements in high-throughput experimental and computational models, gene interactions can now be explored in greater detail and functional validation can be performed. Jackson et al. (2020) have used single-cell RNA sequencing to recapitulate gene regulatory networks, which revealed information about genotype-dependent expression patterns in a variety of environments. Moreover, the technologies of genome editing have allowed making the accurate validation of functional variants easier. Roy et al. (2018) designed a multiplexed genome editing system by a trackable genomic barcode, which facilitates functional interrogation of genetic variants in large scale.
Although there are methodological developments, there are a number of challenges. Numerous published studies concentrate on single aspects like QTL mapping, regulatory variation or functional validation on their own, which restricts the possibility of obtaining a comprehensive measure of genotype-phenotype relationships. Also, the incorporation of high-throughput sequencing data with a detailed functional annotation and association with phenotype analysis is not well-explored. Such fragmentation limits the creation of comprehensive frameworks with the ability of properly interpreting multi-faceted trait structures. Consequently, there exists an urgent need to develop combined methods that involve variant discovery, functional annotation and phenotypic analysis as a single analysis pipeline.

3. MATERIALS AND METHODS

3.1 Dataset Description

Genome-sequencing data of Saccharomyces cerevisiae strains was taken in publicly available genomic repositories such as datasets generated through a wide range of environmental and industrial sources. The chosen data set included several strains with variation in such phenotypic characteristics as rate of growth, stress resistance, and fermentation efficiency. The raw reads of the sequence were obtained in the form of FASTQ files with sufficient coverage to detect variants. The alignment was done with the S. cerevisiae S288C reference assembly, which was obtained at the Saccharomyces Genome Database (SGD). The general workflow of the suggested variant identification and functional annotation pipeline is presented in Fig. 1.

Fig. 1. Workflow of High-Throughput Sequencing-Based Variant Identification and Functional Annotation Pipeline

3.2 Quality Control and Preprocessing

Raw sequencing reads were initially quality checked with FastQC (v0.11.9) which examines parameters like base quality scores, GC content and the level of duplication of sequences and adapter contamination. The quality of data was enhanced by filtering the low-quality bases and adapters with Trimmomatic (v0.39) with the standard filtering options, including a sliding window method (4:20), a minimum read length of 50 bp, and the elimination of leading and trailing low-quality bases. A repeat of post-trimming quality was also done to verify that sequencing artifacts were eliminated and, to provide, high quality reads to proceed with downstream analysis.

3.3 Sequence Alignment and Post-processing

High-quality reads were mapped to the reference genome with the help of the Burrows-Wheeler Aligner (BWA-MEM algorithm, v0.7.17) which is optimized to map short-read sequencing data. SAM files were then converted to Binary Alignment/Map (BAM) format with SAMtools (v1.15). To reduce biases of library preparation, sorted BAM files were obtained and duplicate reads were tagged. The quality measures of the alignment process such as mapping rate and coverage distribution were checked to be sure that the alignment process was reliable.

3.4 Variant Calling and Filtering

The Genome Analysis Toolkit (GATK, v4.2) was used to run variant discovery based on best practices in variant calling. SNPs and INDELs in all the samples were identified by the HaplotypeCaller module. Raw variant calls have undergone stringent filtering according to several quality parameters, such as minimum depth of coverage ($DP \geq 10$), mapping quality ($MQ \geq 30$), and variant quality score ($QUAL \geq 30$). Further filtering was performed to filter low-confidence variants and other possible sequencing artifacts and a high-confidence variant dataset was obtained that could be further analyzed.

3.5 Functional Annotation of Variants

ANNOVAR was used to functionally annotate the filtered variants (latest version) and allowed the classification of variants based on their genomic location and predicted functional impact. Variants were divided into coding and non-coding regions, and it was further divided into synonymous, non-synonymous, frameshift, and regulatory variants. Annotation was done based on gene using the available gene ontology and pathway databases to perform pathway enrichment analysis to identify affected genes and to determine the biological significance of identified variants.

3.6 Phenotypic Association Analysis

In order to examine the interaction between the genetic variants and phenotypic traits, correlational and regression-based statistical association analysis was performed using the standard methods. Variants were compared based on their correlation with phenotypic traits which included: growth rate, ethanol tolerance and response to environmental stress. Variants that had strong associations were identified using significance levels, and multiple testing corrections were taken into consideration in order to minimize false positives. The trait-related variants were further analyzed to identify their functional implication and overall genetic makeup to phenotypic variation.

4. RESULTS

4.1 Variant Discovery

Upon rigorous filtering, a total of about 145,000 SNPs and 18,500 INDELs were discovered in the genomes of Saccharomyces cerevisiae that were analyzed. Variations distribution showed that about 72% of SNPs were observed in non-coding regions and 28% in coding regions (Fig. 2). Chromosomal analysis showed an imbalanced number of variants with chromosomes IV and XII having increased accumulation variants implying the likelihood of genomic hotspots of variation.

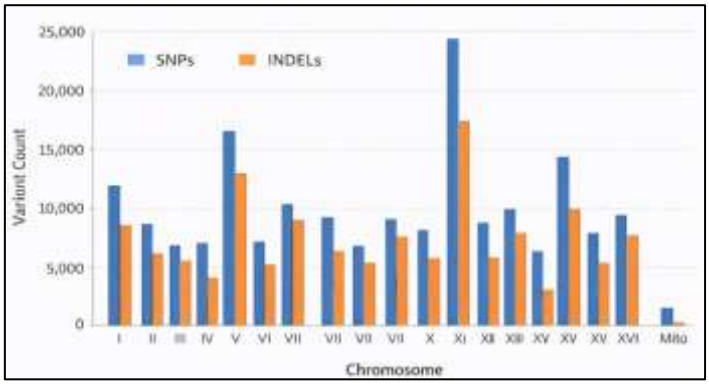

Fig. 2. Genome-wide distribution of SNPs and INDELs across chromosomes

4.2 Functional Annotation

Functional annotation revealed that approximately 60 percent of coding-region variants were synonymous, and approximately 35 percent were non-synonymous, and these may have an effect on protein functionality. A lesser percentage (approximately 5) was made up of frameshift and high-impact variations (Table 1). Interestingly, the enrichment of non-synonymous variants was observed in genes related to metabolic pathways and stress response processes, which meant that they were important in their function.

Table 1: Functional classification of identified variants

Variant Type	Percentage (%)
Synonymous	60
Non-synonymous	35
High-impact (frameshift, stop gain/loss)	5

4.3 Phenotype Association

The association analysis revealed about 120 strong variants ($p < 0.05$ after correction) that were associated with such phenotypic characteristics as ethanol tolerance and growth rate. The genes that are related to energy metabolism and the cellular stress response had several variants found which justify their biological significance. The genome-wide distribution of major variants indicates the presence of a polygenic architecture, meaning that a set of locus affect trait variability (Fig. 3).

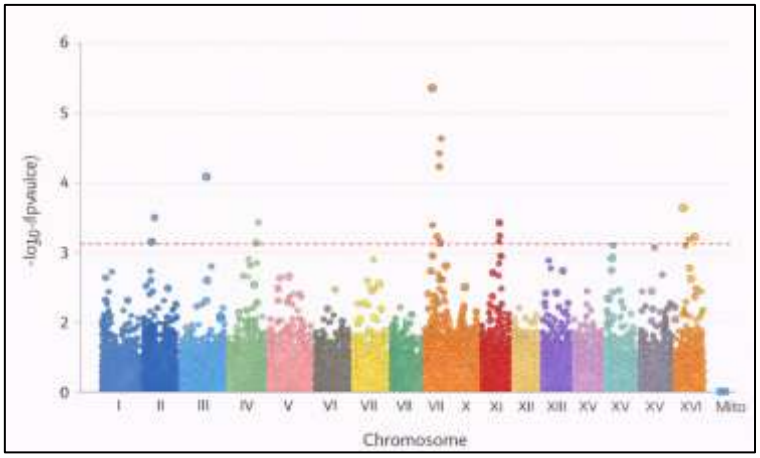

Fig. 3. Manhattan plot showing variant–trait associations

4.4 Pathway Analysis

Pathway enrichment analysis demonstrated that genes involved in glycolysis, oxidative stress response and signal transduction pathways were significantly overrepresented ($p < 0.01$). The identified variants contribute to functional adaptation as these pathways are vital to cellular adaptation and survival in response to changes in environmental conditions (Fig. 4).

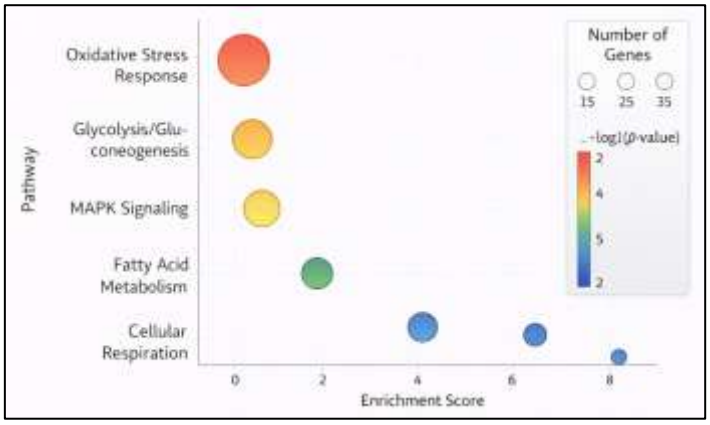

Fig. 4. Pathway enrichment analysis of genes harboring significant variants

5. DISCUSSION

The findings reveal that high-throughput sequencing allows identifying genetic variants and their functional consequences of phenotypic variation in a comprehensive manner. The non-coding and regulatory variants predominance indicates that the regulation of gene expression is even more important than thought before, which aligns with the results of Renganaath et al. (2020) and Shih et al. (2021). This emphasizes the need to take regulatory variants of the genotype-phenotype studies into account along with coding mutations. The enrichment of non-synonymous variants of metabolic and stress-related genes was observed to be associated with the literature of QTL and association and indicated that the mentioned pathways were identified by the former as the key predictors of phenotypic traits (Eder et al., 2018; Wang et al., 2019). Moreover, multiple large variants that are related to phenotypic traits are identified, which favors the polygenic model, with many loci of small to moderate effects on traits. This fact aligns with the literature that tends to stress the complexity and epistaticity of genetic architectures (Duveau et al., 2021; Haas et al., 2019). Moreover, integrating variant discovery and functional annotation gives a more complete insight into the relationship between genotype and phenotype than single methods. Although previous research has restricted its research on regulatory variation, or QTL mapping, or functional validation separately (Jackson et al., 2020; Roy et al., 2018), the current framework has shown the benefit of integrating all parts into a single pipeline. This integrative methodology increases the interpretability of genomic data and helps to identify variants of biological interest.

In sum, the results affirm a complex interplay of coding, regulatory, and interacting genetic factors in regulating phenotypic variation in S. cerevisiae. The study offers a generalizable system of writings to future genomic researches and the need to contemplate the interfaces of multi-level data to elucidate complicated expression of traits fully.

CONCLUSION

This paper introduces a high-throughput sequencing-based work-flow framework to comprehensively identify and functionally annotate genetic variations that are correlated with phenotypic characteristics in Saccharomyces cerevisiae. Through the incorporation of quality-monitored sequencing data, as well as a set of uniform bioinformatics pipelines, the paper was able to discover a big repertoire of SNPs and INDELs and define their functional consequences in coding and regulatory elements of the genome. The outcomes show that a significant fraction of variants is found in the non-coding regions, which underscores the importance of regulatory factors in determinants of phenotypic diversity. Moreover, the association analysis showed that the phenotypic traits including growth performance and stress tolerance are all controlled by polygenic architecture, which incorporates a number of loci as well as multifaceted interactions. The enrichment of the functional and biologically relevant variants in metabolic and stress-response pathways emphasize their biological significance and confirms the evidence found to date on the complexity of genotype-phenotype relationships. The suggested frame is more comprehensive and integrative in its approach as compared to the traditional models that consider individual elements of variant analysis and propose them as a single workflow, which is more complex and global in its implementation. Nevertheless, limitations include the fact that the study is based on single-layer genomic data, which are not likely to be reflective of the complexity of regulatory and epigenetic processes. The next research initiative should be to combine multi-omics data, such as transcriptomics and proteomics, to enhance the interpretation of functions. Moreover, the additional integration of machine learning and predictive modeling strategies may improve the prioritization of variants and allow to identify more causative variants more precisely. Altogether, this study offers a scalable and repeatable model that can be applied to other organisms and complex traits in studies of functional genomics and precision biology to contribute to improvements.

REFERENCES

1. Bai, F.-Y., et al. (2022). The ecology and evolution of the baker's yeast *Saccharomyces cerevisiae*. *Genes, 13*(5), 1–15.

2. Bloom, J. S., et al. (2019). Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife, 8*, e49212.
3. Duveau, F., et al. (2021). Mutational sources of trans-regulatory variation affecting gene expression in *Saccharomyces cerevisiae*. *eLife, 10*, e67821.
4. Eder, M., et al. (2018). QTL mapping of volatile compound production in *Saccharomyces cerevisiae* during alcoholic fermentation. *BMC Genomics, 19*, 1–13.
5. Fournier, T., et al. (2019). Extensive impact of low-frequency variants on the phenotypic landscape at population scale. *eLife, 8*, e49258.
6. Galardini, M., et al. (2019). The impact of the genetic background on gene deletion phenotypes in *Saccharomyces cerevisiae*. *Molecular Systems Biology, 15*(8), e8831.
7. Haas, R., et al. (2019). Mapping ethanol tolerance in budding yeast reveals high genetic variation in a wild isolate. *Frontiers in Genetics, 10*, 1–12.
8. Jackson, C. A., et al. (2020). Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife, 9*, e51254.
9. Jakobson, C. M., et al. (2019). Molecular origins of complex heritability in natural genotype-to-phenotype relationships. *Cell Systems, 8*(5), 363–379.
10. Maclean, C. J., et al. (2017). Deciphering the genic basis of yeast fitness variation by simultaneous forward and reverse genetics. *Molecular Biology and Evolution, 34*(8), 2051–2063.
11. Matsui, T., et al. (2022). The interplay of additivity, dominance, and epistasis on fitness in a diploid yeast cross. *Nature Communications, 13*, 1–12.
12. Nguyen Ba, A. N., et al. (2022). Barcoded bulk QTL mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. *eLife, 11*, e73939.
13. Peter, J., et al. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature, 556*(7701), 339–344.
14. Renganaath, K., et al. (2020). Systematic identification of cis-regulatory variants that cause gene expression differences in a yeast cross. *eLife, 9*, e62669.
15. Roy, K. R., et al. (2018). Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nature Biotechnology, 36*(6), 512–520.
16. Sardi, M., et al. (2018). Genome-wide association across *Saccharomyces cerevisiae* strains reveals substantial variation in underlying gene requirements for toxin tolerance. *PLoS Genetics, 14*(2), e1007217.
17. Sharon, E., et al. (2018). Functional genetic variants revealed by massively parallel precise genome editing. *Cell, 175*(2), 544–557.
18. Shih, C.-H., et al. (2021). Cis-regulatory variants affect gene expression dynamics in yeast. *eLife, 10*, e66069.
19. Sirr, A., et al. (2018). Natural variation in SER1 and ENA6 underlie condition-specific growth defects in *Saccharomyces cerevisiae*. *G3: Genes, Genomes, Genetics, 8*(12), 3957–3967.
20. Wang, Z., et al. (2019). QTL analysis reveals genomic variants linked to high-temperature fermentation performance in industrial yeast. *Biotechnology for Biofuels, 12*, 1–14.