fortune Journals

JOURNAL OF BIOINFORMATICS AND SYSTEMS BIOLOGY

ISSN: 2688-5107

Research Article

Discovery of Predictive Genes of Mice Intraocular Pressure based on RNA-Sequencing data using Machine Learning

Xiaoqin Huang[1&], Akhilesh Kumar Bajpai[2&], Yan Gao[2], Michelle Bao[3], Monica M. Jablonski[1], Siamak Yousefi[1,2]*, Lu Lu[1,2*]

Abstract

Purpose: Intraocular pressure (IOP) is a major risk factor for open angle glaucoma. IOP reduction is the only alterable factor for glaucoma treatment other than surgery. Lowering IOP is critical for glaucoma management. This study aims to identify predictive genes of mice IOP.

Methods: Several machine learning models were applied for IOP classification based on RNA-sequencing data of BXD mouse strains. The predictive genes were selected based on feature importance of the best model coupled with sequential feature selection. The collective IOP predictive genes were validated based on IOP phenotypes of mouse strains with different ages.

Results: The best classification model based on IOP phenotype achieved an area under the receiver operating characteristic curve (AUC) of 0.94 (95% CI 0.93-0.96) with an accuracy of 77% (95% CI 74-78%). Fifty genes were identified as predictive genes of mice IOP. The AUC of the model based on the independent dataset (phenotype record ID BXD12303, age 3-5 months) was 0.90 (95% CI 0.89-0.91) with an accuracy of 81% (95% CI 81-81%), and for the IOP (phenotype record ID BXD_12300, age 1-2 months) classification, the AUC was 0.94 (95% CI 0.94-0.94) with an accuracy of 69% (95% CI 69-69%). A total of five genes (out of 50) were previously identified as associated with glaucoma, leading to an enrichment ratio of 2.73.

Conclusions: Machine-learning models identified a group of predictive genes for mice IOP and showed an improvement in the glaucoma gene enrichment ratio compared with the traditional linear association models.

Keywords: Machine learning; IOP; Predictive gene; Glaucoma.

Introduction

Intraocular pressure (IOP) is a measure of the fluid pressure within the eye. It is an important marker for many ophthalmological diseases, including glaucoma, the second leading cause of irreversible blindness worldwide [1-3]. IOP is a primary risk factor for primary open angle glaucoma (POAG), and population-based studies have suggested a 16% increase in risk for developing POAG for every mmHg increase in IOP [4-9]. Predicting IOP behavior is significant because IOP represents a crucial phenotype in the context of glaucoma. Furthermore, the predictive capability can serve as a pivotal reference for early identification of the disease, along with its potential application in developing drug targets for glaucoma treatment. Machine learning models have been widely used in disease-related phenotype

Affiliation:

[1]Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, USA

[2]Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, USA

[3]Ursuline Academy of Dallas, Dallas, USA

***Corresponding author:**

Lu Lu, Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, USA; Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, USA

Siamak Yousefi, Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, USA; Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, USA

& Both authors contributed equally.

Citation: Xiaoqin Huang, Akhilesh Kumar Bajpai, Yan Gao, Michelle Bao, Monica M. Jablonski, Siamak Yousefi, Lu Lu. Discovery of Predictive Genes of Mice Intraocular Pressure based on RNA-Sequencing data using Machine Learning. Journal of Bioinformatics and Systems Biology. 6 (2023): 339-346.

Received: September 20, 2023
Accepted: September 27, 2023
Published: November 29, 2023

prediction [10-13]. In a recent study, fundus photography was combined with systemic variables to develop a deep learning regression model for predicting IOP, yielding a mean absolute error of 2.29 mmHg [1]. However, these variables require several clinical examinations, which can incur significant costs and time investments. Genomic data has been widely used for phenotype prediction [14-16]. For example, Mohammed et al. [17] developed an ensemble deep learning model to classify 5 cancer types using RNA-seq data, and the model performance achieved more than 99% in each class. Coleto-Alcudia et al. applied an Artificial Bee Colony based on Dominance (ABCD) algorithm, which internally uses a support vector machine (SVM) classifier, and identified cancer biomarkers from RNA-seq data and further tested them in different datasets [18]. These results demonstrate that the method we propose is effective in gene selection for the identification of cancer biomarkers from RNA-seq data.

However, to the best of our knowledge, few studies have been reported to predict glaucoma-related phenotypes using genomic data and machine learning analysis. The goal of our study is to identify biomarkers or candidate genes involved in glaucoma pathogenesis using data generated in the BXD recombinant inbred (RI) genetic reference population (GRP) [19,20]. The BXD family has been derived from the crosses between DBA/2J (D2) and C57BL/6J (B6) mice and is currently the best pheno- and geno-characterized inbred animal population [20,21]. GeneNetwork (https://genenetwork.org/), an open-source platform hosts thousands of phenotype datasets and hundreds of omics datasets corresponding to BXD mice, making this GRP a powerful tool for systematic genetics exploration. Furthermore, the parental D2 strain develops a naturally occurring chronic secondary angle-closure glaucoma [22,23] and is considered as a congenital experimental model of glaucoma. The D2 mice develop a form of glaucoma that results from the abnormal liberation of iris pigment (iris pigment dispersion, IPD) in the anterior chamber, which obstructs drainage routes for aqueous humor and results in a marked elevation of IOP. A mutation in the glycoprotein (transmembrane) Nmb gene (GpnmbR150X) results in IPD in D2 mice. A separate phenotype known as iris stromal atrophy, occurs due to a mutation in the tyrosine protein type 1 gene (Tyrp1b) [23-26]. Together, these result in “pigment dispersion syndrome” (PDS) followed by pigmentary glaucoma (PG) [25,27,28]. The BXD RI strains share genotypes from glaucoma-D2 and control-B6 mice, thus making them a perfect animal model to explore the mechanisms underlying glaucoma. In this study, we applied machine learning to identify a collective of predictive genes for IOP classification based on RNA-seq data of BXD mice strains. Subsequently, we validated these genes by employing them for IOP phenotype across mice of varying age groups. Furthermore, gene enrichment analysis was employed to compare the predictive set of genes with the reference glaucoma genes.

Methods

Mice

Mice were handled in accordance with the Guide for the Care and Use of Laboratory Animals. Studies, and all mice experiments were approved by the Animal Care and Use review board of University of Tennessee Health Science Center and were carried out following the ARVO Statement for the Use of Animals in Ophthalmic and Vision Research. Mice were housed as previously described [29], and were maintained in light: dark cycle of 12 h:12 h with lights switched on at 6 AM. A total of 3,856 BXD mice and both parental strains were used by us for generating phenotype and gene expression data. Additional details including mice sex ratio and different age groups can be found in our previous study [30]. For generating gene expression data, the animals were sacrificed under saturated isoflurane. Eyeballs from the animals were dissected and stored at −80°C until RNA extraction.

IOP measurement

The mice were anesthetized by intraperitoneal injection with a mixture of 25 mg/kg ketamine and 5 mg/kg xylazine as previously reported [30]. The IOP of both eyes of all mice was measured immediately after the induction of general anesthesia using an induction–impact tonometer (Tonolab, Colonial Medical Supply, Franconia, NH) as previously described [30]. IOP was measured as soon as the mouse was completely unconscious (typically in 2-3 minutes). Six consecutive IOP readings were averaged. The IOP readings obtained with Tonolab tonometer have been shown to be accurate and reproducible in different mouse strains, including DBA/2J [31]. Furthermore, the impact of the Tonolab probe on the cornea is minimal and is not known to cause either corneal damage or progressive changes in IOP, even after repeated readings. All IOP were measured between 9:00 am and 6:00 pm during the light cycle. We did not observe any obvious effects of sexes on IOP [32], hence the values were averaged among male and female mice. The mice with obvious abnormality in eye, such as bulbar atrophy, corneal ulcer, and staphyloma were excluded from the study.

For the current analysis, we used three IOP phenotypes, which have been deposited in our GeneNetwork portal (https://genenetwork.org/) and can be accessed through the following phenotype record IDs: BXD_15976, BXD_12300 and BXD_12303. Phenotype BXD_15976 (n = 77 strains): Intraocular pressure (IOP) of all ages (1 to 30 months old), both sexes, average of left and right eyes; Phenotype BXD_12300 (n = 67 strains): Intraocular pressure (IOP), 1 to 2 months old, both sexes, average of left and right eyes [mmHg]; Phenotype BXD_12303 (n = 69 strains): Intraocular pressure (IOP), 3 to 5 months old, both sexes, average of left and right eyes [mmHg].

Citation: Xiaoqin Huang, Akhilesh Kumar Bajpai, Yan Gao, Michelle Bao, Monica M. Jablonski, Siamak Yousefi, Lu Lu. Discovery of Predictive Genes of Mice Intraocular Pressure based on RNA-Sequencing data using Machine Learning. Journal of Bioinformatics and Systems Biology. 6 (2023): 339-346.

Eye RNA-seq data generation and preprocessing

For this study, we generated RNA-seq data using the Illumina HiSeq 2000 platform. The dataset encompassed ocular transcriptome data collected from a total of 157 animals across 93 BXD strains. Briefly, total RNA was extracted using Trizol® reagent (Invitrogen, Grand Island, NY, USA) according to the manufacturer's instructions. Both left and right eyeballs from one mouse were added to a single 2 mL tube containing 700 µL QIAzol Lysis Reagent and one 5 mm stainless steel bead (Qiagen, Hilden, Germany). The eye tissue was homogenized for 2 min in a Tissue Lyser II (Qiagen, Hilden, Germany) with a speed frequency of 30 r followed by incubating for 5 min. Then, 140 µL chloroform was added into the homogenate, shaken vigorously for 15 s, and centrifuged for 15 min at 12,000 x g at 4℃, and 280 µL upper aqueous was then transferred into a new collection tube containing 500 µL 100% ethanol. The mixture was loaded into a RNeasy mini Quick spin column (Qiagen, Valencia, CA, USA) followed by Buffer RWT once and Buffer RPE purification twice. All RNA samples were treated with DNase to avoid DNA contamination, and verified by Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). RNA with OD260/280 > 1.8 and RIN > 8.0 were used for library preparation. One microgram of RNA was used for cDNA library construction at Novogene using an NEBNext® Ultra RNA Library Prep Kit for Illumina® (New England Biolabs, Ipswich, MA, USA) according to the manufacturer's protocol. Briefly, mRNA was enriched using oligo(dT) beads followed by two rounds of purification and fragmented randomly by adding fragmentation buffer. The first strand cDNA was synthesized using random hexamers primer, after which a custom second-strand synthesis buffer (Illumina, San Diego, CA, USA), dNTPs, RNase H and DNA polymerase I were added to generate the second strand (ds cDNA). After a series of terminal repair, poly-adenylation, and sequencing adaptor ligation, the double-stranded cDNA library was completed followed by size selection and PCR enrichment. The resulting 250-350 bp insert libraries were quantified using a Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and quantitative PCR. Size distribution was analyzed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Qualified libraries were sequenced on an Illumina HiSeq2000 Platform (Illumina, San Diego, CA, USA) using a paired-end 150 run (2×150 bases). An average of 40 million raw reads were generated from each library.

Mus musculus (mouse) reference genome (GRCm38) and gene model annotation files were downloaded from the Ensembl genome browser (https://useast.ensembl.org/). The paired-end reads were aligned to the reference genome using STAR v2.5.0a [33] aligner. FeatureCount v0.6.1 [34] was used to count the number of reads mapped to each gene. Transcripts Per Million (TPM) was calculated for each gene based on the length of the gene and reads mapped to that gene. The TPM was further rescaled to log2(TPM+1). The dataset used in the current analysis are available on our GeneNetwork website (www.genenetwork.org) under the name "UTHSC BXD Young Adult Eye RNA-Seq (Nov20) TPM Log2".

Model development

The phenotype dataset was merged with the RNA-seq dataset based on the common strains. Samples with no phenotype measurement were excluded from the dataset. Genes with over 90% missing values were filtered. Subsequently, the remaining genes were ranked in descending order based on their MAD (mean absolute deviation) scores. The phenotype was converted to a binary variable. To be specific, phenotype values exceeding the median of the group were assigned as 1, while values equal to or below the median were assigned a label of 0. We split the combined dataset into training and testing sets with allocating proportions of 80% and 20%, respectively. Data in the training set was used for model fitting, and data in the testing set was utilized for model evaluation, as shown in Table 1.

Model training

In the training stage, we fitted three tree-based classifiers for each phenotype prediction: Random Forest (RF), Gradient Boosting (GBM), and XGBoost (XGB). To tune the hyper-parameters of each classifier, we applied a grid search with a 10-fold cross-validation with the scoring method "roc_auc". The hyper-parameter space we searched is shown as below: number of genes with top MAD: 50, 100, 200, 300, 400; number of trees: 60, 100, 140, 180, 220, 260, 300, 340, 380; learning rate (only for Gradient boosting and XGBoost): 0.01, 0.05, 0.1; maximum depth: 2, 3, 4, 5, 6, 7. After the search, the model with the best performance was selected as the final model for evaluation.

Model evaluation

We used the area under the receiver operating characteristic curve (AUC) and accuracy as the evaluation metrics for each model. We set the cutoff threshold to 0.5 as the data is highly balanced.

Table 1: Sample distribution for machine learning classifiers

Phenotype ID. (No. of samples)	No. of case/ control	No. of genes	No. of genes in the merged dataset	No. of case/control in the training set	No. of case/control in the testing set
15976 (89)	44/45	24575	17530	30/41	14/4
12300 (85)	39/46	24575	17492	31/38	8/8
12303 (85)	42/43	24575	17492	36/33	6/10

Table 2: Results of the best model optimized using grid search

Phenotype ID. (No. of samples)	Phenotype description	AUC (95% CI)	Accuracy (%) (95% CI)	No. of genes	Enrichment ratio
15976(89)	IOP (1-30 month)	**0.94 (0.93,0.96)**	**77 (74, 78)**	400	1.43
12300(85)	IOP (1-2 month)	0.8 (0.78, 0.81)	69 (67, 71)	50	1.64
12303(85)	IOP (3-5 month)	0.54 (0.52, 0.55)	68 (67, 69)	400	1.3

Predictive gene set selection

For each fitted model (RF, GBM, XGB), we selected a set of genes with feature importance higher than 0. Furthermore, we applied a sequential feature selection (SFS) method to find the predictive gene set, a subset of the genes with the best predictive performance.

Cross phenotype validation

We evaluated the predictive genes from phenotype (No. 15976, 1-30 months) with other IOP datasets (No. 12303, 3-5 months, 12300, 1-2 months). We applied the same procedure for this experiment except with a fixed number of features (n = 50). The prediction performance of the model was compared with the initial best model in the above section.

Gene enrichment analysis

The gene-sets selected based on the different models were further validated for their biological significance using the enrichment of known glaucoma genes. The glaucoma-related genes were obtained from multiple publicly available resources, including DISEASES database [35] (https://diseases.jensenlab.org/), UniProtKB [36] (https://www.uniprot.org/uniprotkb), GeneCards [37] (https://www.genecards.org/), and Alliance database [38](https://www.alliancegenome.org/), as described in our previous study [39]. Finally, a comprehensive list of glaucoma-related genes was derived by combining the gene-sets obtained from the above resources and removing duplicates. This final set obtained is henceforth referred to as the “glaucoma-reference set”. The gene lists identified based on different models were compared with the glaucoma-reference set and the number of overlapping genes was determined. To compare the results across multiple lists from different models, we derived an enrichment ratio using the following formula:

$$\text{Enrichment ratio} = \frac{(g/n)}{(G/N)} \quad (1)$$

where g = the number of overlapping genes between glaucoma-reference set and model set;

G = the total number of genes in glaucoma-reference set; n = the number of genes selected from the model; and N = the total number of genes in the mouse genome used in the analysis.

Results

After filtering the genes which were expressed in at least 10% of the samples, the RNA-seq dataset included approximately 17,000 genes with different number of samples for each phenotype listed in Table 1.

The results of the best model for the three selected phenotypes are presented in Table 2. The best model for IOP BXD_15976 classification is RF, with the top 400 genes sorted by MAD, 60 estimators and maximum depth of 2. The AUC was 0.94 (95% CI 0.93-0.96) and accuracy was 77% (95% CI 74-78%). The best model for phenotype 12300 was based on top 50 genes, with AUC of 0.80 (95% CI, 0.78-0.81), and accuracy of 69% (95% CI 67-71%). The lowest model performance was for phenotype 12303 with top 400 genes, achieving AUC of 0.54 (95% CI 0.52-0.55), and accuracy of 68% (95% CI 67-69%). Apparently, the best result was the phenotype BXD_15976 classification. Thus, we further selected the predictive genes from the result of this model.

The top 400 genes were extracted and sorted by the feature importance (FI). There were 106 genes with feature importance higher than 0; hence, they were selected for optimizing the model by including the genes additively. The model with the best performance on the testing set was by using the top 50 genes (FI-50 genes) with AUC of 0.94 (95% CI 0.93-0.96) and accuracy of 77% (95% CI 74-78%), as shown in Table 3.

Table 3: Performance of the optimized model on different number of genes

No. of top genes	AUC (95% CI)	Accuracy (%) (95% CI)
10	0.91 (0.90,0.93)	75 (73, 78)
20	0.93 (0.92,0.95)	75 (73, 78)
30	0.93 (0.92,0.95)	75 (73, 78)
40	0.88 (0.86,0.89)	75 (73, 78)
50	**0.94 (0.93,0.96)**	**77 (74,78)**
60	0.92 (0.91, 0.93)	75 (73, 78)
70	0.93 (0.91,0.95)	75 (73,78)
80	0.91 (0.90,0.93)	77 (75,83)
90	0.93 (0.90,0.95)	70 (68,72)
100	0.91 (0.90,0.93)	75 (73,78)
106	0.91 (0.90,0.93)	75 (73,78)

Note: The model with the best performance is highlighted in bold font.

These FI-50 genes were further validated by applying them to another IOP phenotype classification, with different age groups: BXD_12300 (IOP, 1-2 months) and BXD_12303 (IOP 3-5 months). The performance of the best model was: IOP BXD_12300: AUC of 0.94 (95% CI 0.94-0.94), and accuracy of 69% (95% CI 69-69%); IOP BXD_12303: AUC of 0.90 (95% CI 0.89-0.91), and accuracy of 81% (95% CI 81-81%), as shown in Table 4. These values were much higher than their best models optimized based on the sets of genes sorted by MAD, which is shown in Table 2, where IOP BXD_12303 only obtained AUC of 0.54 and accuracy of 68%. When FI-50 genes were applied, the performance improved significantly. Table 5 lists the selected FI-50 genes that were used for testing the performance of the IOP classification model.

Because IOP is an important phenotype of glaucoma, we used a set of glaucoma reference genes (749 genes) generated in our previous study [39] to compare with the FI-50 selected genes and calculated the enrichment ratio. Our results indicated that 5 of the FI-50 genes (H2-Ab1, Lyz2, Pvalb, Crygd and Crygb) were glaucoma-associated genes. Furthermore, the enrichment ratio of glaucoma genes was 2.73, which was significantly improved compared to the typical linear model (where the enrichment ratio was 1). However, the enrichment ratio of the MAD-sorted 400 genes for the phenotype BXD_15976 was 1.43, while that of 50 genes of phenotype BXD_12300 was 1.64, and for 400 genes of phenotype BXD_12303 as 1.30.

Discussion

In this study, we used machine learning to select a set of predictive genes for IOP classification using step-by-step

Table 4: Performance of the IOP classification model optimized using 50 selected genes

Phenotype ID	Phenotype description	AUC (95% CI)	Accuracy (%) (95% CI)
15976	IOP (1-30 month)	0.94 (0.93,0.96)	77 (74,78)
12300	IOP (1-2 month)	0.94 (0.94,0.94)	69 (69, 69)
12303	IOP (3-5 month)	0.90 (0.89,0.91)	81 (81,81)

Table 5: List of FI-50 selected genes for IOP classification

Gene name	Gene description
Crct1	Cysteine-rich C-terminal 1
Morc2b	Microrchidia 2B
Atp6v0c-ps2	ATPase, H+ transporting, lysosomal V0 subunit C, pseudogene 2
H2-K1	Histocompatibility 2, K1, K region
H2-Aa	Histocompatibility 2, class II antigen A, alpha
H2-Ab1	**Histocompatibility 2, class II antigen A, beta 1**
Cdsn	Corneodesmosin
Prss33	Protease, serine 33
BC051142	cDNA sequence BC051142
Krt4	Keratin 4
Ucp3	Uncoupling protein 3 (mitochondrial, proton carrier)
H2-K2	Histocompatibility 2, K region locus 2
Tex35	Testis expressed 35
Lyz2	**Lysozyme 2**
Ncr3-ps	Natural cytotoxicity triggering receptor 3, pseudogene
Cryge	Crystallin, gamma E
Rps18-ps3	Ribosomal protein S18, pseudogene 3
Col11a2	Collagen, type XI, alpha 2
Krt17	Keratin 17
Spink5	Serine peptidase inhibitor, Kazal type 5
Pvalb	**Parvalbumin**
Crygd	**Crystallin, gamma D**
Chrne	Cholinergic receptor, nicotinic, epsilon polypeptide
mt-Nd4l	Mitochondrially encoded NADH dehydrogenase 4L
Mir184	MicroRNA 184
Crygb	**Crystallin, gamma B**
Tmem181c-ps	Transmembrane protein 181C, pseudogene
Rsph3b	Radial spoke 3B homolog (Chlamydomonas)
Rec8	REC8 meiotic recombination protein
H2-DMb1	Histocompatibility 2, class II, locus Mb1
Ces1a	Carboxylesterase 1A
Tff2	Trefoil factor 2 (spasmolytic protein 1)
Gm9390	Wilms' tumour 1-associating protein pseudogene
Myl2	Myosin, light polypeptide 2, regulatory, cardiac, slow
Scn4b	Sodium channel, type IV, beta
Mlycd	Malonyl-CoA decarboxylase
Il36a (Il1f9)	Interleukin 36A
Nexn	Nexilin
Klhl31	Kelch-like 31
Sprr1a	Small proline-rich protein 1A
Art1	ADP-ribosyltransferase 1
Rps18	Ribosomal protein S18
Myom3	Myomesin family, member 3
mt-Te	Mitochondrially encoded tRNA glutamic acid
Tnni2	Troponin I, skeletal, fast 2
Eqtn	Equatorin, sperm acrosome associated
Tpsb2	Tryptase beta 2
Zfp949	Zinc finger protein 949
Rps27a-ps1	Ribosomal protein S27A, pseudogene 1
Mpz	Myelin protein zero

Note: Bold font genes are known to be associated with glaucoma, based on text mining.

Citation: Xiaoqin Huang, Akhilesh Kumar Bajpai, Yan Gao, Michelle Bao, Monica M. Jablonski, Siamak Yousefi, Lu Lu. Discovery of Predictive Genes of Mice Intraocular Pressure based on RNA-Sequencing data using Machine Learning. Journal of Bioinformatics and Systems Biology. 6 (2023): 339-346.

methods. Firstly, we ranked the genes based on MAD. Due to a small number of samples, only top 400 genes were included to avoid overfitting. The genes were further filtered based on their importance in providing the highest accuracy based on the best model. The set of genes were validated in a similar phenotype classification task with different age groups (IOP BXD_12300 and IOP BXD_12303). In all circumstances, the model performed better using the selected genes than their best model optimized by using a series of gene-sets (50, 100, 200, 300, 400) sorted by MAD, indicating that the FI-50 genes were predictive for IOP classification.

We were also interested to know if those FI-50 predictive genes were functionally meaningful, because being predictive does not always mean association. As it is well known that elevated IOP is a significant risk factor for developing glaucoma and higher IOP is also the primary risk factor of glaucoma, a reference set of 749 glaucoma related genes was used to compare with our model genes. By comparing with the MAD-sorted model genes, we obtained an enrichment ratio of 1.43, 1.64 and 1.30 for IOP phenotype BXD_15976, BXD_12300 and BXD_12303, respectively. Then, we compared the most predictive set of FI-50 genes with the reference glaucoma gene set and discovered that 5 of them (H2-Ab1, Lyz2, Pvalb, Crygd and Crygb)were glaucoma-related genes, with an enrichment ratio of 2.73. This outcome showed significant improvement compared to the traditional linear model (enrichment ratio =1). H2-Ab1 has several functions including peptide and protein antigen binding activity. It is also known to be involved in processes, such as B cell affinity maturation, cellular response to interferon-gamma, and positive regulation of T-helper 1 type immune response [40]. Lyz2 is involved in defense response to gram-negative and gram-positive bacterium. A study by Panagis et al. [41], measured the downregulation of Lyz2 in 9-month-old animals with high IOP exposure when compared with age-matched animals with low IOP exposure. Pvalb is known to be involved in excitatory and inhibitory chemical synaptic transmission. A recent study suggested that the expression of Pvalb is downregulated as the mice aged and developed glaucoma with retinal ganglion cell loss [42]. Crygd and Crygb, both belonging to the crystallin family of proteins, are best known as structural constituents of eye lens. Mutations in the genes of this family are known to be associated with cataract [43,44]. Furthermore, recent studies have stressed the neuroprotective roles of the crystallins in glaucoma [45,46]. Liu et al. [45] showed downregulation of crystallins including both Crygd mRNA and protein in an experimental animal model of glaucoma. Mirzaei et al. [46] demonstrated that crystallins including CRYGB and CRYGD were up to 18-fold downregulated at the protein level, in glaucoma condition compared to control.

This study has a few limitations. Firstly, the dataset used is small and the generalizability of the genes needs to be validated in a larger dataset once available. Secondly, the function of those selected genes needs to be validated through experimental research.

Conclusion

This study developed various machine learning models to select a subset of genes that were predictive of phenotypes based on IOP. Findings may lead to approaches for screening patients with ocular hypertension in the early stage. Furthermore, this study sheds light on providing guidance for clinicians to identify patients who may need closer monitoring. Further work is required to validate the identified genes in a larger dataset as well as in experimental research.

References

1. Ishii K, Asaoka R, Omoto T, et al. Predicting intraocular pressure using systemic variables or fundus photography with deep learning in a health examination cohort. Sci Rep 11 (2021): 3687.

2. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. Br J Ophthalmol 90 (2006): 262-267.

3. Zhang N, Wang J, Li Y, et al. Prevalence of primary open angle glaucoma in the last 20 years: a meta-analysis and systematic review. Sci Rep 11 (2021): 13762.

4. Khawaja AP, Cooke Bailey JN, Wareham NJ, et al. Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. Nat Genet 50 (2018): 778-782.

5. De Voogd S, Ikram MK, Wolfs RC, et al. Incidence of open-angle glaucoma in a general elderly population: the Rotterdam Study. Ophthalmol 112 (2005): 1487-1493.

6. Nemesure B, Honkanen R, Hennis A, et al. Incident open-angle glaucoma and intraocular pressure. Ophthalmol 114 (2007): 1810-1815.

7. Leske MC, Heijl A, Hyman L, et al. Predictors of long-term progression in the early manifest glaucoma trial. Ophthalmol 114 (2007): 1965-1972.

8. Sommer A, Tielsch JM, Katz J, et al. Relationship between intraocular pressure and primary open angle glaucoma among white and black Americans. The Baltimore Eye Survey. Arch Ophthalmol 109 (1991): 1090-1095.

9. Chan MP, Grossi CM, Khawaja AP, et al. Associations with Intraocular Pressure in a Large Cohort: Results from the UK Biobank. Ophthalmol 123 (2016): 771-782.

10. Jia J, Wang R, An Z, et al. RDAD: A Machine Learning System to Support Phenotype-Based Rare Disease Diagnosis 9 (2018).

11. Guo T, Li X. Machine learning for predicting phenotype

from genotype and environment. Curr Opin Biotechnol 79 (2023): 102853.

12. Maintz L, Welchowski T, Herrmann N, et al. Machine Learning–Based Deep Phenotyping of Atopic Dermatitis: Severity-Associated Factors in Adolescent and Adult Patients. JAMA Dermatol 157 (2021): 1414-1424.
13. Strauss MJ, Niederkrotenthaler T, Thurner S, et al. Data-driven identification of complex disease Phenotype 18 (2021): 20201040.
14. Cheng C-Y, Li Y, Varala K, et al. Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. Nat Commun 1 (2021): 5627.
15. Anilkumar C, Muhammed Azharudheen TP, Sah RP, et al. Gene based markers improve precision of genome-wide association studies and accuracy of genomic predictions in rice breeding. Heredity 130 (2023): 335-345.
16. Muneeb M, Feng S, Henschel A. Transfer learning for genotype–phenotype prediction using deep learning models. BMC Bioinformatics 23 (2022): 511.
17. Mohammed M, Mwambi H, Mboya IB, et al. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Sci Rep 11 (2021): 15626.
18. Coleto-Alcudia V, Vega-Rodríguez MA. A multi-objective optimization approach for the identification of cancer biomarkers from RNA-seq data. Expert Syst Appli 193 (2022): 159-180.
19. Andreux PA, Williams EG, Koutnikova H, et al. Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. Cell 150 (2012): 1287-1299.
20. Ashbrook DG, Arends D, Prins P, et al. A platform for experimental precision medicine: The extended BXD mouse family. Cell Syst 12 (2021): 235-247.
21. Williams EG, Auwerx J. The Convergence of Systems and Reductionist Approaches in Complex Trait Analysis. Cell 162 (2015): 23-32.
22. John SW, Smith RS, Savinova OV, et al. Essential iris atrophy, pigment dispersion, and glaucoma in DBA/2J mice. Invest Ophthalmol Vis Sci 39 (1998): 951-962.
23. Libby RT, Anderson MG, Pang IH, et al. Inherited glaucoma in DBA/2J mice: pertinent disease features for studying the neurodegeneration. Vis Neurosci 22 (2005): 637-648.
24. Anderson MG, Nair KS, Amonoo LA, et al. GpnmbR150X allele must be present in bone marrow derived cells to mediate DBA/2J glaucoma. BMC Genet 9 (2008): 30.
25. Anderson MG, Smith RS, Hawes NL, et al. Mutations in genes encoding melanosomal proteins cause pigmentary glaucoma in DBA/2J mice. Nat Genet 30 (2002): 81-85.
26. Howell GR, Libby RT, Marchant JK, et al. Absence of glaucoma in DBA/2J mice homozygous for wild-type versions of Gpnmb and Tyrp1. BMC Genet 8 (2007): 45.
27. Chang B, Smith RS, Hawes NL, et al. Interacting loci cause severe iris atrophy and glaucoma in DBA/2J mice. Nat Genet 21 (1999): 405-409.
28. Ritch R. A unification hypothesis of pigment dispersion syndrome. Trans Am Ophthalmol Soc 94 (1996): 381-405.
29. Swaminathan S, Lu H, Williams RW, et al. Genetic modulation of the iris transillumination defect: a systems genetics analysis using the expanded family of BXD glaucoma strains. Pigment Cell Melanoma Res 26 (2013): 487-498.
30. Lu H, Lu L, Williams RW, et al. Iris transillumination defect and its gene modulators do not correlate with intraocular pressure in the BXD family of mice. Mol Vis 22 (2016): 224-233.
31. Nagaraju M, Saleh M, Porciatti V. IOP-dependent retinal ganglion cell dysfunction in glaucomatous DBA/2J mice. Invest Ophthalmol Vis Sci 48 (2007): 4573-4579.
32. Savinova OV, Sugiyama F, Martin JE, et al. Intraocular pressure in genetically distinct mice: an update and strain survey. BMC Genet 2 (2001): 12.
33. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29 (2013): 15-21.
34. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30 (2014): 923-930.
35. Pletscher-Frankild S, Palleja A, Tsafou K, et al. DISEASES: text mining and data integration of disease-gene associations. Methods 74 (2015): 83-89.
36. UniProt C. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51 (2003): 523-531.
37. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics 54 (2016): 30-33.
38. Alliance of Genome Resources C. Harmonizing model organism data in the Alliance of Genome Resources. Genet 220 (2022).
39. Huang X, Bajpai AK, Sun J, et al. A new gene-scoring method for uncovering novel glaucoma-related genes using non-negative matrix factorization based on RNA-seq data. Front Genet 14 (2003): 1204909.

40. Zhou G, Ding ZC, Fu J, et al. Presentation of acquired peptide-MHC class II ligands by CD4+ regulatory T cells or helper cells differentially regulates antigen-specific CD4+ T cell response. J Immunol 186 (2011): 2148-2155.

41. Panagis L, Zhao X, Ge Y, et al. Retinal gene expression changes related to IOP exposure and axonal loss in DBA/2J mice. Invest Ophthalmol Vis Sci 52 (2011): 7807-7816.

42. Liu Yuan HRC, Munguba Gustavo C, Lee Richard K. Parvalbumin expression changes with retinal ganglion cell degeneration. Front Neurosci 12 (2023).

43. Santana A, Waiswol M, Arcieri ES, et al. Mutation analysis of CRYAA, CRYGC, and CRYGD associated with autosomal dominant congenital cataract in Brazilian families. Mol Vis 15 (2009): 793-800.

44. Gao Y, Ren X, Fu X, et al. Case Report: A Novel Mutation in the CRYGD Gene Causing Congenital Cataract Associated with Nystagmus in a Chinese Family. Front Genet 13 (2002): 824550.

45. Liu H, Bell K, Herrmann A, et al. Crystallins Play a Crucial Role in Glaucoma and Promote Neuronal Cell Survival in an In Vitro Model Through Modulating Muller Cell Secretion. Invest Ophthalmol Vis Sci 63 (2022): 3.

46. Mirzaei M, Gupta VB, Chick JM, et al. Age-related neurodegenerative disease associated pathways identified in retinal and vitreous proteome from human glaucoma eyes. Sci Rep 7 (2017): 12685.