A High-Throughput Zebrafish Screening Method for Visual Mutants by Light-Induced Locomotor Response

Yuan Gao, Rosa H.M. Chan, Tommy W.S. Chow, Liyun Zhang, Sylvia Bonilla, Chi-Pui Pang, Mingzhi Zhang, and Yuk Fai Leung

Abstract—Normal and visually-impaired zebrafish larvae have differentiable light-induced locomotor response (LLR), which is composed of visual and non-visual components. It is recently demonstrated that differences in the acute phase of the LLR, also known as the visual motor response (VMR), can be utilized to evaluate new eye drugs. However, most of the previous studies focused on the average LLR activity of a particular genotype, which left information that could address differences in individual zebrafish development unattended. In this study, machine learning techniques were employed to distinguish not only zebrafish larvae of different genotypes, but also different batches, based on their response to light stimuli. This approach allows us to perform efficient high-throughput zebrafish screening with relatively simple preparations. Following the general machine learning framework, some discriminative features were first extracted from the behavioral data. Both unsupervised and supervised learning algorithms were implemented for the classification of zebrafish of different genotypes and batches. The accuracy of the classification in genotype was over 80 percent and could achieve up to 95 percent in some cases. The results obtained shed light on the potential of using machine learning techniques for analyzing behavioral data of zebrafish, which may enhance the reliability of high-throughput drug screening.

Index Terms—High-throughput drug screening, zebrafish, machine learning, classification, light-induced locomotor response

1 Introduction

Zebrafish, *Danio rerio*, is a freshwater tropical fish. The fish are small and are easily reared in small aquariums. Sexually mature fish have high fecundity and a pair of healthy adult fish can lay up to 200 eggs at weekly intervals. Moreover, the larval fish develop rapidly and are highly transparent, which allow for easy observation of developmental process (for a comprehensive review, see Zhang et al. [1]). In addition, its light-sensitive tissue in retina is anatomically comparable to human [2]; hence, zebrafish is an excellent model for eye disease and eye drug discovery [1], [2], [3], [4]. A number of fish genetic mutants that affect vision has been identified and/or generated over the years. The visual defects in these mutants affect their response to light stimulus, and result in an alteration of their locomotor response. Taken together, these advantages have made the zebrafish amenable to high-throughput behavioral studies and opened up a tremendous opportunity to rapidly identify chemicals that may alter light-induced behavior [5]. Recently, two light-induced behavioral assays have been developed and used to characterize 14,000 and 4,000 drugs on neural [6] and sleeping behavior [7], respectively.

During the course of their investigation, Emran and colleagues further customized this light-induced behavioral assay as shown in Fig. 1 and demonstrated that visual mutants had a specific alteration during the acute phase of light-transition [8], [9]. This acute response, also known as the visual-motor response (VMR), is an eye-driven startle response triggered by a rapid onset or offset of light stimulation [9], [10]. VMR is largely attenuated in genetic mutants that have defects in vision, and it is completely abolished in a mutant that fails to develop an eye and in larvae after removal of the eye ball [9], [10]. Due to this unique relationship between VMR and vision, it has been proposed that the VMR assay can be used to evaluate drug therapies on different fish genetic mutants to expedite the discovery of eye drugs [1].

To this end, Zhang et al. have recently utilized the VMR assay to demonstrate the visual benefit of *Schisandrin B* (*Sch B*), an active ingredient of a traditional Chinese medicine *Fructus Schisandrae* [11], on a visual mutant $pde6c^{w59}$ [12], [13] (refer to as *pde6c* hereafter). The results from Zhang et al. indicate that *Sch B* could enhance the VMR response of *pde6c* under the light-ON stimulus, supporting the potential eye benefits of *Sch B* [10], [14]. In humans, mutations in *pde6c* cause similar phenotypes in the eyes as

- *Y. Gao, R.H.M. Chan, and T.W.S. Chow are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China. E-mail: ethan.y.gao@my.cityu.edu.hk, {rosachan, eetchow}@cityu.edu.hk.*
- *L. Zhang and S. Bonilla are with the Department of Biological Sciences, Purdue University, 915 W. State Street, West Lafayette, IN 47907. E-mail: {zhang383, sbonilla}@purdue.edu.*
- *C.-P. Pang is with the Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong, China. E-mail: cppang@cuhk.edu.hk.*
- *M. Zhang is with the Joint Shantou International Eye Center, Shantou University and the Chinese University of Hong Kong, Shantou, China. E-mail: zmz@jsiec.org.*
- *Y.F. Leung is with the Department of Biological Sciences and the Center for Drug Discovery, Purdue University, 915 W. State Street, West Lafayette, IN 47907 and the Department of Biochemistry and Molecular Biology, Indiana University School of Medicine Lafayette, 625 Harrison Street, West Lafayette, IN 47907. E-mail: yfleung@purdue.edu.*

Manuscript received 18 June 2013; revised 20 Jan. 2014; accepted 21 Jan. 2014. Date of publication 18 Feb. 2014; date of current version 4 Aug. 2014. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2014.2306829

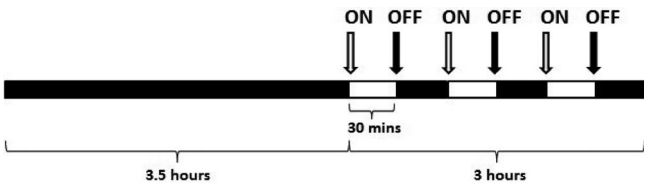

Fig. 1. Conceptual diagram for the timeline of light stimuli in the LLR assay. The larvae were dark-adapted in the Zebrabox for 3.5 hours immediately before the experiment, then the assay lasted 3 hours with light stimuli: ON-OFF-ON-OFF-ON-OFF. Each stimulus was sustained for 30 min.

in the zebrafish mutant. Thus, the study of drug effects on the VMR of various zebrafish visual mutants can potentially facilitate the discovery of novel treatments for patients.

There are multiple implementations of the light-induced behavioural assays in the literature [9], [15]. The differences in the settings led to controversy in the visual contribution to the VMR. In this work, our usage of VMR and experimental collection of data are restricted in the same configurations established by Emran et al., which will be discussed in Section 4.3. In order to avoid confusion, we describe this (Emran's) collection scheme [9] as an assay for measuring light-induced locomotor response (LLR). The systematic analysis of this LLR output is essential for determining the visual components and realizing the potential use of VMR for drug screening for visual impairment. We characterized the LLR results obtained from our recent VMR study using normal wild-type (WT) and visually-impaired mutant (*pde6c*) zebrafish larvae in order to elucidate the influence of genetic mutations on the LLR [10], [14]. However, biological data are intrinsically noisy and the individual variation may mask the biological difference. To maximize the detection of the true biological effect, these biological data were often averaged to eliminate the noise or random movements [7], [8], [9], [16].

With the rapid development of computation biology, various machine learning methods are implemented to assist multi-dimensional behavioral data analysis for specific purposes. For example, advanced studies in machine vision investigated the animal behavioral data by categorizing them into discrete behavioral pattern. Kabra et al. developed an interactive tool to annotate the animal behavior automatically by Gentle AdaBoost [17]. Mirat et al. used support vector machine (SVM) to categorized the episodes of movement for zebrafish into several maneuvers, e.g., slow forward swim, routine turn and escape [18]. Rather than the widely used spatio-temporal features, Burgos-Artizzu et al. proposed trajectory features to represent the video data for behavioral pattern recognition using AdaBoost [19]. On the other hand, hidden Markov model (HMM) [20] was also implemented to investigate the underlying states by observed animal behavior. Franke et al. used HMM to analyze the movements of caribous and to predict their behavioral states such as bedding, feeding and relocating [21]. HMM was also used to study the behavioral change before and after formaldehyde treatment in zebrafish to detect the change of underlying state in pre-defined behavioral pattern [22], [23].

Instead of categorizing behavior into predefined patterns of machine vision based behavior analysis [17], [18], [19], biological sample screening aims to classify different biological samples according to their difference in behavior [1], [5], [6], [7], [8]. Additionally, being different from machine vision based research which use videos as input [17], [18], [19], the input of biological sample screening analysis is a specific time-series behavioral curve generated by specific metric (e.g., velocity) [1], [5], [6], [7], [8]. In a pioneering drug screening study by zebrafish, Rihel et al. used Locomotor activity curve recorded by Zebrabox [24] to extract some behavioral activity features, and then applied clustering methods to study the similarity and dissimilarity in the effectiveness of different drugs according to the change in behavior [7]. Nevertheless, the behavioral data in [7] were still averaged before feature extraction, which left individual differences unaddressed. In view of the limitations of the current statistical studies [25] and the benefits of utilizing machine learning techniques in biological data analysis [26], a novel machine learning based framework is introduced in this study to model the behavioral data of individual zebrafish for high-throughput screening.

In our proposed machine learning framework, LLR of WT and *pde6c* could be studied individually. The influence of different batches, i.e., embryos collected from different heterozygous parents, could also be identified by our methods. Similar to [7], the behavioral data used in our study is also time-series curve recoded by Zebrabox [24], and data was labeled in advance to build the model. Discriminative features were first extracted from the preprocessed data. After that, supervised learning algorithms, including K nearest neighbor (KNN) [20], Naive Bayes [27], SVM [28] were implemented for classification. In addition, we also conducted an unsupervised learning algorithm, expectation-maximization algorithm with Gaussian mixture model (EM-GMM) [20], with the labeled data utilized for initialization.

The contribution of this paper is fourfold:

- We developed a machine-learning framework for quantifying LLR of different zebrafish subtypes.
- We demonstrated the importance of LLR, especially the acute phase, in identifying visually-impaired larvae.
- The novel framework allows us to screen *pde6c* mutants efficiently with relatively simple experimental preparations.
- Through our framework, the zebrafish larvae that do not respond differently from the majority of the larvae from the same class can be automatically identified as outliers for further experimental characterization.

2 Methodology

2.1 Zebrafish Maintenance

All fish lines used in this study were originated from the strain AB [29] and maintained according to standard procedures [30]. Two genotypes, WT and *pde6c*, were used. For the WT, the embryos were collected from two different batches of adults which were designated as #1 and

TABLE 1
Relationship of Different Fish

	WT-#1 (W1)	WT-#2 (W2)	*pde6c*-#2 (P2)
WT-#1 (W1)	-	Same Genotype Unrelated Family	Different Genotype Unrelated Family
WT-#2 (W2)	Same Genotype Unrelated Family	-	Different Genotype Cousins
pde6c-#2 (P2)	Different Genotype Unrelated Family	Different Genotype Cousins	-

The nomenclature of them are genotype-batch notation.

#2. And all *pde6c* embryos were collected from the heterozygous parents from batch #2. The WT embryos collected from #2 were cousins of the *pde6c* embryos collected from #2, while the WT embryos collected from #1 were from an unrelated family. The two genotypes were robustly discriminated by the lack of optokinetic response (OKR), a stereotypic eye reflex in response to any movements in the environment [31], [32], in the *pde6c* larvae [12]. Using genotype-batch notation, three classes of zebrafish larvae were used: WT-#1, WT-#2 and *pde6c*-#2, which are abbreviated as W1, W2, P2 in Section 4 for drawing the result figures. The nomenclature of fish and their relationship are summarized in Table 1. In each class, 24 larvae were used. The embryos used for the experiments were collected and raised at 28°C in E3 medium [33]. All protocols were approved by the Purdue Animal Care and Use Committee.

2.2 Behavioral Data Collection

The behavioral data were collected by Zebrabox (Viewpoint Life Sciences) [24]. Before the assay, all the zebrafish larvae were accommodated in a 96-well plate for overnight (In this study, 72 wells are filled with zebrafish larvae while 24 wells are empty). Immediately before the experiment, the larvae in the plate were dark-adapted in the Zebrabox for 3.5 hours. During the assay, a bright light stimulus was given in the following sequence: ON-OFF-ON-OFF-ON-OFF, in which each stimulus lasted for 30 min, as illustrated in Fig. 1. The movement of the larvae was detected by an infrared camera that captured videos in 30 frames per second. As the metric of the movement, *burst duration* was defined and detected by the following extraction scheme. First, a detection sensitivity was set to compare two successive frames to define *active pixels*, i.e., the pixels which have a change in grey level more than the detection sensitivity. In this study, the detection sensitivity was set at 6. These *active pixels* represent the movement of a larva in successful frames. And then, a burst threshold (BT) was set to define *actual movement* (i.e., burst) when the number of *active pixels* between two frames was larger than a predefined value. Finally, the *burst duration* was defined as the summarized *actual movement* duration per second, i.e., the fraction of second a larva actually move (bin size = 1 s). In this study, we tested two BTs: 0 and 4. For BT = 0, all movements are regarded as the *actual movement*, while for BT = 4, small movements lower than 4 *active pixels* were not regarded as the *actual movement*. The latter value, i.e., BT = 4 pixel, was used in previous studies [7], [9]. This arrangement allowed us to determine the contribution of the information from the small movements in classification. And we used this scheme to collect larval movement data at five days post-fertilization (dpf) and 8 dpf.

TABLE 2
Features Used

Maximal Amplitude	Sample Entropy
Number of Rest Bout	Number of Active Bout
Averaged Rest Bout Length	Averaged Active Bout Length
Length of First Active Bout	Length of First Rest Bout
Mean of the Total Response	Mean of the Active Response

2.3 Machine Learning Framework

The Machine Learning Framework for data analysis consists of three stages: pre-processing, feature representation and classification. Most of the pre-processing in our work has been done by Zebrabox. For instance, various threshold parameters discussed above were used to filter out the noise. Other necessary pre-processing will be discussed in Section 4. Here, we focus on feature representation and classification. Feature representation aims to use small amount of features to accurately describe the data set, so that large inputs for algorithm in classification stage, which may lead to curse of dimensionality, can be avoided. And by using the features from former stage, classification is used to distinguish different samples into groups.

2.3.1 Feature Representation

We compared two candidate feature representation methods, i.e., Symbolic Aggregate approXimation (SAX) features used in computer science for time-series analysis [34], and the features used in the pioneering work by Rihel et al. SAX features represent the histogram of segmentally coded, symbolized time-series curve as features, where the symbolized curve is obtained by discretizing the original time-series data. And The feature set used by Rihel et al. consists of empirically selected metrics of the activity curve by the biologists [7]. Note that although Rihel's feature set was designed for a much longer period (24 hours for each rest +wake phases) than ours (1 hours for each ON+OFF stimuli), the bin size of Rihel's collection scheme is 1 min while ours is 1 second. Thus, the durations of Rihel's assay and ours are comparable when considering the total number of bins, which makes Rihel's feature set as a candidate feature set for our assay. Classification tested on our zebrafish behavioral data using the feature sets extracted by Rihel's method [7] and SAX [34] indicates that features used by Rihel et al. offered best accuracy. Considering the wake/rest phase of Rihel's assay is also triggered by light ON/OFF, the result demonstrates that Rihel's features are sensitive to light changes, and they are suitable to represent our zebrafish activity curve, even though this feature set was originally proposed to study rest/wake phase. In this paper, based on the work by Rihel et al. [7], the extended feature set in Table 2 is extracted from both light-ON and light-OFF stimuli. Thus, for each zebrafish in each ON-OFF trial, there are 20 features in total.

Rest bout corresponds to the period with continual 1-second bins when zebrafish larvae do not move, i.e., the continuous bins of *burst duration* = 0. And the opposite case is

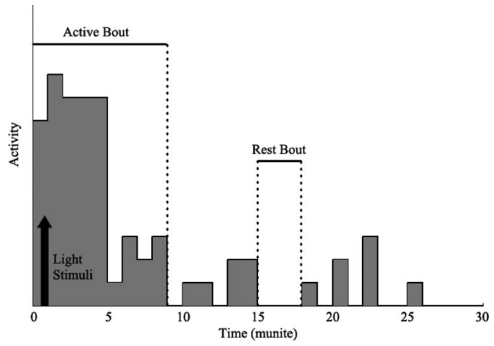

Fig. 2. Illustration of key features used.

defined as *active bout*, as illustrated in Fig. 2. *Sample entropy* is a widely applied measurement of complexity and regularity of biological time series [35].

2.3.2 Classification

Both unsupervised and supervised learning methods were used in this study. They are: 1) K Nearest Neighbor (KNN) classifier; 2) Naive Bayes classifier; 3) Support Vector Machine (SVM); 4) Expectation-Maximization algorithm with Gaussian Mixture Model (EM-GMM). The former 3 methods are supervised learning methods. While the latter one, EM-GMM, is an unsupervised learning method which can be incorporated with label information for initialization.

The neighbor parameter $K = 3$ of KNN classifier [20] was chosen as it was neither too global nor too local. The standard version of Naive Bayes classifier as described in [27] has been used in our experiment. For SVM method, a"soft" margin SVM, C-SVM [20], [27], with Radial Basis Function (RBF) kernel is used to gain a non-linear classification. The coefficient of slack variables C and the kernel width parameter γ is set to 2 and 0.125. A free C++ library *LIBSVM* is available in [28]. For EM-GMM [20], to avoid overfitting, the mean μ_m and variance σ_m^2 of each Gaussian were fixed as initialization by the labeled data, only the prior of each Gaussian π_m was updated.

3 Results

The proposed framework was implemented to identify individual activity profiles of different genotypes and/or batches. In order to investigate the effect of LLR under light stimulus, the baseline activity after a prolonged dark adaptation was collected from the last 30 min of the initial 3.5-hour dark phase in Fig. 1 as control.

3.1 Evaluation Metric

Ten-fold 500 times cross validation was used for generating training sets and validation sets for supervised methods KNN, Naive Bayes and SVM. For unsupervised method EM-GMM, the algorithm was initialized by the mean of the training set from the cross validation to cluster the corresponding validation set.The evaluation metric is the classification accuracy, as (1):

$$\text{accuracy} = \frac{\#\ \text{correctly predicted samples}}{\#\ \text{total testing samples}} \times 100\%. \quad (1)$$

3.2 Classification Accuracy

To examine the classification accuracy using only baseline activities without light stimulus, the last 0.5-hour of the 3.5-hour dark adaptation data were used for classification. And for LLR analysis, The data obtained from 3 ON-OFF trials were averaged as the input of the classification algorithms, as illustrated in Fig. 3. To identify the significant component of LLR, the data were segmented in three different durations: the first minute, the first two minutes as well as the whole 30-min LLR after the change in stimulus for testing. Since similar results were obtained among the four classification methods, for precision, only the classification results for 8 dpf using all 30-min LLR data were summarized here to illustrate the performance of all four methods presented in Fig. 4. Considering SVM could slightly outperform others for most cases (especially for classifying genotypes of same batch, i.e., WT-#2 versus *pde6c*-#2), and it is easy to implement (free libraries are available such as *LIBSVM* [28]), we only illustrated SVM results hereafter. For the complete results of the four methods, please refer to supplementary data, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2306829.

The SVM results of 5 and 8 dpf zebrafish larvae using both baseline activity and LLR are shown in Fig. 5. The results demonstrated that the accuracy could achieve as high as 95 percent (the classification accuracy WT-#1 versus *pde6c*-#2 using first 1-min LLR, as the highest value of BT = 0 at Row 1, Column 2 of Fig. 5). And for all the cases in zebrafish genotype classification using LLR, the accuracy is over 80 percent.

4 Discussion

This paper presented the implementation of various machine learning methods for analyzing the behavioral

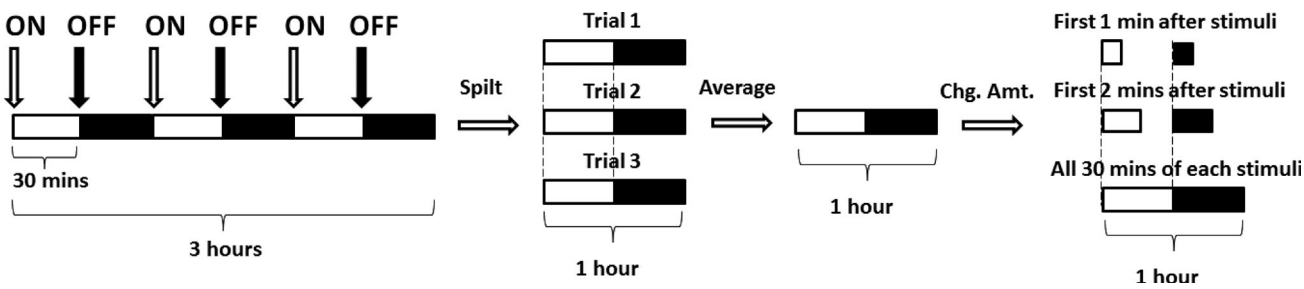

Fig. 3. Conceptual illustration of data extracted from each individual zebrafish for different classification schemes.

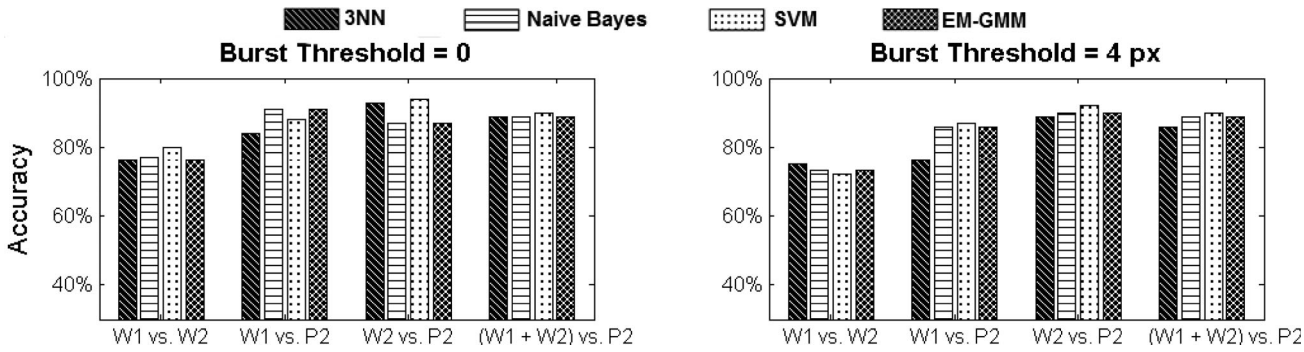

Fig. 4. Classification results of 8 dpf zebrafish larvae using 30-min LLR. Activities of both BT = 0 and BT = 4 pixel were analyzed. We used four different classification methods, as bars from left to right: 3 Nearest Neighbor (3NN), Naive Bayes (NB), Support Vector Machine (SVM) and Expectation-Maximization algorithm with Gaussian Mixture Model (EM). In addition, 4 different classification problems were tested, in which W1, W2, P2 denoted WT-#1, WT-#2, *pde6c*-#2, respectively.

data of zebrafish larvae by individual. The behavioral data used in our work is named as LLR to avoid confusion by different VMR measuring schemes of Erman [9] and Fernandes [15]. We tested our methods on zebrafish larvae of two age groups, which consisted of three classes in each group. The experimental results indicate that our machine learning methods are efficient and effective. The accuracy can reach over 80 percent in general with the highest up to 95 percent for classifying zebrafish larvae of different genotypes. Note that although the following discussion are based on SVM results as Fig. 5, same conclusion can be deduced from the other three methods because they share similar classification results.

4.1 Importance of Developing an Efficient Method for Mutant Screening

Many visual mutants are recessive in nature. For example, the *pde6c* mutant used in this study is a recessive mutant. These recessive mutants, which comprise 25 percent of a cross of heterozygous parents, have to be first identified before any drug study can be conducted. While both alleles of the gene are mutated in these individuals, these visual mutants often do not have any obvious morphological defects and this precludes an easy identification of them. Currently, they have to be identified individually based on pre-inserted reporter genes and/or optokinetic response, which can be extremely inefficient to implement in large-scale studies. By the behavioral models of mutants built with the machine learning method, new individual mutants can be identified automatically in parallel by a single VMR assay. Thus, the computational algorithms developed by us can significantly facilitate the genetic screening and in turn drug screening.

4.2 The Benefits of the Proposed Classification Framework to Drug Screening

In the current implementation of the LLR for drug screening, the behavioural profiles were averaged according to one main parameter (e.g., the genotype of normal and mutant samples) to investigate their mean differences. Instead, our proposed method, which analyzes individual behavioral data, would facilitate the determination of the

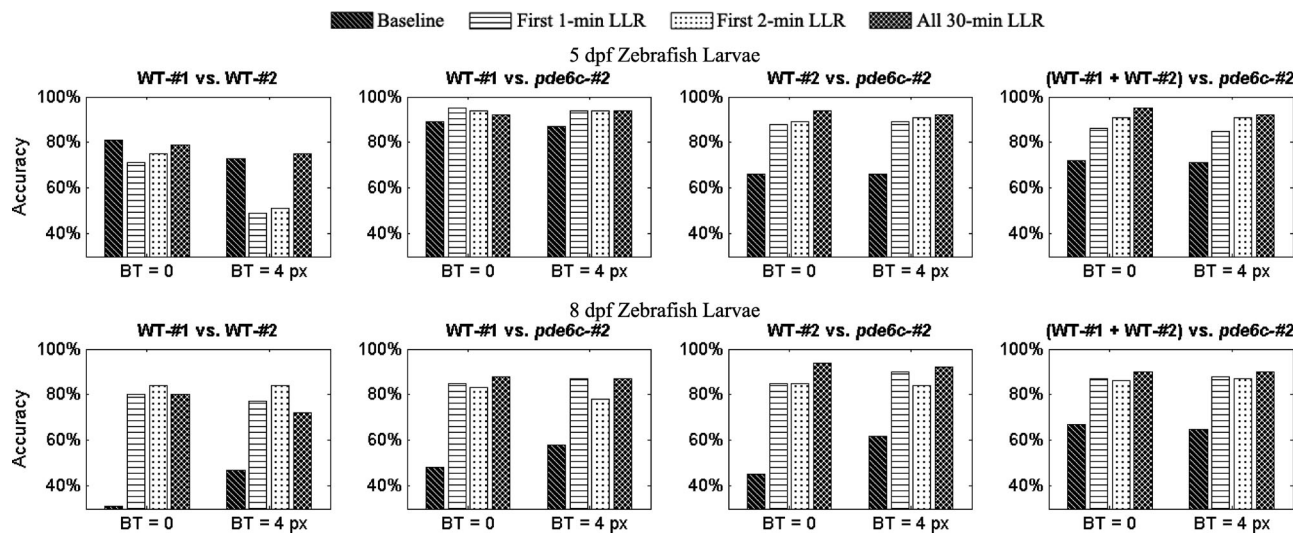

Fig. 5. SVM classification results using both Baseline and LLR data of 5 dpf (Top Row) and 8 dpf (Bottom Row) zebrafish larvae. For classification using baseline data, the activities recorded in the last 30 min of the initial 3.5-hour dark period before the first light-ON stimulus was used as input. While for LLR classification, three different amounts of data were used, i.e., first 1-min, first 2-min and all 30-min LLR data. Activities of both BT = 0 and BT = 4 pixel were analyzed. Four different classification problems were tested, as the four subfigure columns. The other parameters are the same with Fig. 4.

most useful drugs by identifying those making mutants resume similar LLR profile as the healthy controls. Our classifiers would be also an important first step to grouping similar drugs by the LLR of treated mutants using a relatively simple experimental setup. This in turn would facilitate the determination of unknown drug action through the known drugs with similar LLR profiles. In addition, our method can efficiently identify outliers that behaved differently from other typical zebrafish in the same treatment group for further neurobiological analysis. These outliers may carry other natural genetic variations/single nucleotide polymorphisms that act as a modifier of the drug response. The identification of these individuals may facilitate the further investigation of the drug action.

4.3 The Relationship between VMR and LLR

The VMR was first introduced by Emran and colleagues to demonstrate the acute phase of light-transition that would cause a specific alteration of visual mutants [8], [9]. Recently, Fernandes and colleagues attempted to characterize the neurological basis of photokinesis and claimed that there are two additional non-visual components of VMR including inputs from pineal gland and a region in the hypothalamus [15]. In particular, they showed that an eyeless mutant *chokh* and eye enucleation would not completely abolished VMR in normal larvae while Emran et al., showed that the VMR was abolished in the same *chokh* mutant [9]. However, it should be noted that there are major differences between the data collection and summarization in these studies. In particular, Fernandes et al., calculated mean displacement per minute while Emran et al., calculated mean fraction of movement per second. In other words, Fernandes' collection approach summarized the slow response to light stimuli while Emran's approach was used for capturing fast response, which is more compatible for measuring visual startle. This is likely the reason why the *chokh* mutant did not have VMR in Emran's study. In addition, non-visual contribution to locomotor response is by definition not "visual" motor response; thus, we restrict our usage of VMR and experimental collection of data in the format established by Emran et al. (as shown in Fig. 1). Since in this collection scheme [9], the stimulus is changed every 30 mins, it is likely that non-visual contribution of locomotor response would play a role in the later phase of the stimulus. In other words, one would expect the time immediately after light stimuli change would be primarily visual-driven, while the full 30-min data after light stimuli change would be both visual and non-visual driven. Thus, we describe the behaviour measured by the collection scheme in Fig. 1 as light-induced locomotor response (LLR), which acknowledges the fact that the collection scheme would detect locomotor behaviour comprises visual and non-visual inputs.

4.4 Significance of Using LLR to Identify WT and *pde6c* Zebrafish

Comparing the classification results using the baseline data alone, the classification accuracy using 30 min LLR data achieved the highest increase to more than 40 percent, as shown between the first bar and the last bar of BT = 0, in Row 2, Column 3 of Fig. 5. And this increase is more significant in 8 dpf zebrafish. Moreover, by using LLR data, the classification of different genotypes can be carried out with over 90 percent accuracy.

4.5 The Influence on Identifying WT and *pde6c* Zebrafish by Different Amount of Behavioral Data After Light Stimulus Change

Fig. 5 also shows that the classification of different genotypes using data collected from different durations after light stimulus change were not significantly different from each other. It demonstrates that the early component immediately after stimulus change (i.e., the first 1-min LLR) is the most important for classification. Since the *pde6c* used in this study is a visual mutant, this further implies that the early component is primarily vision driven. This observation corroborates the empirical experience that the VMR is an acute response of the LLR.

4.6 The Optimal Burst Threshold for Classification

Classification results were different for data extracted from different BTs. A low BT allowed small movements to be extracted, but such movements were filtered out when a higher BT was set. High BT was often used to capture the major effect of drug treatment or to identify different genotypes [7], [9].

However, our results shown in Fig. 5 indicate that activity profiles collected with BT $= 0$ generally provide better discriminative power. This indicate that potentially useful information can be neglected by filtering out noise with BT $= 4$. Thus, BT $= 0$ is likely a more optimal setting for classification studies like ours.

4.7 The Necessity of Matching Controls

Our observations also indicate the importance of using appropriate matching controls to maximize the discriminative power of the visual-behaviour analysis. For example, WT-#1 and WT-#2 are both normal larvae and yet there are instances that they could be classified fairly up to 80 percent accurate with our algorithms (Fig. 5, Row 1, Column 1). In addition, WT-#2 is genetically more related to *pde6c*-#2 as they are cousins, while WT-#1 is not. It can be inferred that (1) WT-#1 would be easier to be differentiated from *pde6c*-#2, i.e., the accuracy of classification of WT-#1 and *pde6c*-#2 was generally high, especially for 5 dpf zebrafish (Row 1, Column 2 of Fig. 5, 95 percent); (2) WT-#2 would be intrinsically harder to be differentiated from *pde6c*-#2 using baseline activity profile. Column 3 of Fig. 5 indicate that the accuracy using baseline activity are 66 percent for 5 dpf zebrafish and 45 percent for 8 dpf; and (3) WT-#2 should be robustly separated from *pde6c*-#2 if the part of the activity profile that is highly influenced by vision is used. This agrees with our result in Column 3 of Fig. 5. Once we used the 1-min LLR data for classifying WT-#2 and *pde6c*-#2, the classification accuracy substantially increased (66 to 89 percent for 5 dpf, 45 to 85 percent for 8 dpf). Additionally, using longer duration data only provided a modest increase in classification accuracy for WT-#2 and *pde6c*-#2. Since *pde6c* is a visual mutant, this observation indicates that the 1-min LLR (i.e., the VMR as defined by Emran et al. [9]), which supposedly captures the visual startle efficiently,

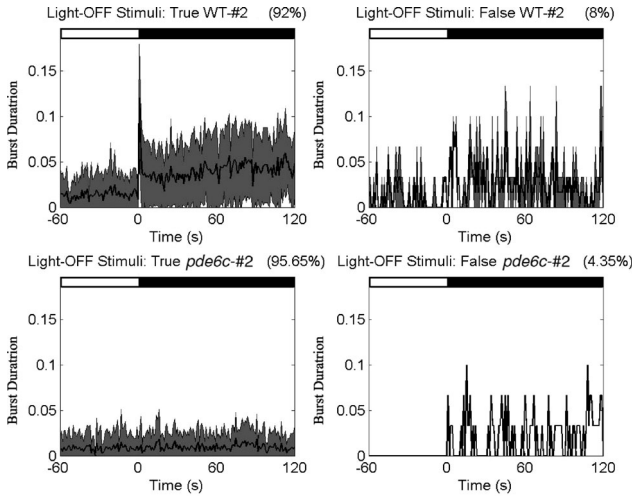

Fig. 6. Light-OFF activity profiles of inliers and outliers of 8 dpf zebrafish WT-#2 versus *pde6c*-#2, where BT was 0 and behavioral data length was 30 min. The activity was measured by Burst Duration, which has been defined in Section 2. The light-OFF stimulus was given at time 0.

provides a good discriminative power for analyzing visual mutants. Our results suggested that analysis conducted between visual mutants and closely related WT controls by VMR is critical for screening drugs that may affect vision.

4.8 Identification of Outliers

The outliers detected by our methods, which can be used for further neurobiological study, are illustrated in the following. As stated above, the results from data extracted by lower threshold were generally more accurate, and different amount of data do not influence much for the classification. Thus, we used 30-min data extracted from BT = 0 to illustrate inliers and outliers detected from our framework. To alleviate the impact caused by the difference in algorithms, individuals that all four algorithms cannot correctly classify were indicated as outliers, while the remaining was considered as inliers. For instance, Fig. 6 demonstrated the light-OFF activity profiles of zebrafish classed as True WT-#2, True *pde6c*-#2, False WT-#2 and False *pde6c*-#2, for 8 dpf zebrafish larvae, respectively.

5 Conclusion

In this paper, we proposed an integrated machine learning framework for behavior based individual zebrafish screening of different genotypes and different batches. The behavioral data of individual zebrafish was recorded as time-series curve according to its *burst duration*. After extracting the features from the behavioral curve, both supervised and unsupervised learning methods, including KNN, Naive Bayes Classifier, SVM and EM-GMM, were used to classify different zebrafish types. The experimental results show that our framework can accurately classify different zebrafish, with the highest accuracy up to 95 percent for classifying zebrafish from different genotypes. We also discussed the effectiveness of LLR for identifying WT and *pde6c* zebrafish, the benefits of using lower burst threshold in pre-processing, as well as the significance of the acute phase of LLR.

In particular, our results suggest that LLR is imperative for the classification of normal zebrafish and zebrafish with visual defects. We also showed that first minute data after the light change was sufficient to capture the light-induced movements initiated by the eyes and would give sufficient accuracy for classifying visual mutants. Additionally, our results show that activity data collected with burst threshold at 0 would give the best discriminative power between different genotypes. In addition, it is essential to use matching controls that are closely related to visual mutant maximize the utility of the difference of the visual behavior for classification and drug study. In summary, our framework has laid down an important foundation to use visual behaviour, particularly VMR, for high-throughput drug screening.

Acknowledgments

This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project no. CityU 123312) and a grant from City University of Hong Kong (Project no. 7003013). Y.F. Leung was partially supported by a collaborative fund from the JSIEC.

References

[1] L. Zhang, L. Chong, J. Cho, P.-C. Liao, S. Feichen, and Y.F. Leung, "Drug Screening to Treat Early-Onset Eye Diseases: Can Zebrafish Expedite the Discovery?" *The Asia-Pacific J. Ophthalmology*, vol. 1, no. 6, pp. 374-383, 2012.

[2] J.M. Fadool and J.E. Dowling, "Zebrafish: A Model System for the Study of Eye Genetics," *Progress in Retinal and Eye Research*, vol. 27, no. 1, pp. 89-110, 2008.

[3] J. Gross and B. Perkins, "Zebrafish Mutants as Models for Congenital Ocular Disorders in Humans," *Molecular Reproduction and Development*, vol. 75, no. 3, pp. 547-555, 2008.

[4] A.C. Morris, "The Genetics of Ocular Disorders: Insights from the Zebrafish," *Birth Defects Research Part C, Embryo Today: Rev.*, vol. 93, no. 3, pp. 215-228, 2011.

[5] C.A. Lessman, "The Developing Zebrafish (*Danio rerio*): A Vertebrate Model for High-Throughput Screening of Chemical Libraries," *Birth Defects Research Part C, Embryo Today: Rev.*, vol. 93, no. 3, pp. 268-280, 2011.

[6] D. Kokel, J. Bryan, C. Laggner, R. White, C.Y.J. Cheung, R. Mateus, D. Healey, S. Kim, A.A. Werdich, S.J. Haggarty, C.A. Macrae, B. Shoichet, and R.T. Peterson, "Rapid Behavior-Based Identification of Neuroactive Small Molecules in the Zebrafish," *Nature Chemical Biology*, vol. 6, no. 3, pp. 231-237, 2010.

[7] J. Rihel, D.A. Prober, A. Arvanites, K. Lam, S. Zimmerman, S. Jang, S.J. Haggarty, D. Kokel, L.L. Rubin, R.T. Peterson, and A.F. Schier, "Zebrafish Behavioral Profiling Links Drugs to Biological Targets and Rest/Wake Regulation," *Science*, vol. 327, no. 5963, pp. 348-351, 2010.

[8] F. Emran, J. Rihel, and J. Dowling, "A Behavioral Assay to Measure Responsiveness of Zebrafish to Changes in Light Intensities," *J. Visualized Experiments*, no. 20, pp. 1-6, 2008.

[9] F. Emran, J. Rihel, A.R. Adolph, K.Y. Wong, S. Kraves, and J.E. Dowling, "OFF Ganglion Cells Cannot Drive the Optokinetic Reflex in Zebrafish," *Proc. Nat'l Academy of Sciences USA*, vol. 104, no. 48, pp. 19 126-19 131, 2007.

[10] L. Zhang, L. Chong, J. Cho, W. Zhong, K.M. Ko, and Y.F. Leung, "*Schisandrin B* Enhanced the Visual Motor Response and Protected the Rod Photoreceptors of the Zebrafish *pde6c* Retinal Degeneration Mutant," submitted for publication, 2013.

[11] R. K.-M. Ko and D.H.F. Mak, "*Schisandrin B* and Other Dibenzocyclooctadiene Lignans," *Herbal and Traditional Medicine: Biomolecular and Clinical Aspects*, pp. 289-314, CRC Press, 2004.

[12] G. Stearns, M. Evangelista, J.M. Fadool, and S.E. Brockerhoff, "A Mutation in the Cone-Specific *pde6* Gene Causes Rapid Cone Photoreceptor Degeneration in Zebrafish," *J. Neuroscience*, vol. 27, no. 50, pp. 13 866-13 874, 2007.

[13] A. Lewis, P. Williams, O. Lawrence, R.O.L. Wong, and S.E. Brockerhoff, "Wild-Type Cone Photoreceptors Persist Despite Neighboring Mutant Cone Degeneration," *J. Neuroscience*, vol. 30, no. 1, pp. 382-389, 2010.
[14] Y.F. Leung, L. Zhang, L. Chong, J. Cho, and K.M. Ko, "*Schisandrin B* Improved the Visual Moter Response and Preserves Photoreceptors in the Zebrafish *pde6c* Cone Dystrophy Mutant," *Investigative Ophthalmology and Visual Science*, vol. 54, p. 1943, ARVO E-Abstract, 2013.
[15] A.M. Fernandes, K. Fero, A.B. Arrenberg, S.A. Bergeron, W. Driever, and H.A. Burgess, "Deep Brain Photoreceptors Control Light-Seeking Behavior in Zebrafish Larvae," *Current Biology*, vol. 22, no. 21, pp. 2042-2047, Nov. 2012.
[16] F. Emran, J. Rihel, A.R. Adolph, and J.E. Dowling, "Zebrafish Larvae Lose Vision at Night," *Proc. Nat'l Academy of Sciences USA*, vol. 107, no. 13, pp. 6034-6039, Mar. 2010.
[17] M. Kabra, A.A. Robie, M. Rivera-Alba, S. Branson, and K. Branson, "JAABA: Interactive Machine Learning for Automatic Annotation of Animal Behavior," *Nature Methods*, vol. 10, no. 1, pp. 64-67, 2013.
[18] O. Mirat, J.R. Sternberg, K.E. Severi, and C. Wyart, "ZebraZoom: An Automated Program for High-Throughput Behavior Analysis and Categorization," *Frontiers in Neural Circuits*, vol. 7, pp. 107:1-107:12, 2013.
[19] X.P. Burgos-Artizzu, P. Dollár, D. Lin, D.J. Anderson, and P. Perona, "Social Behavior Recognition in Continuous Video," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1322-1329, 2012.
[20] C.M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[21] A. Franke, T. Caelli, and R.J. Hudson, "Analysis of Movements and Behavior of Caribou (*Rangifer tarandus*) Using Hidden Markov Models," *Ecological Modelling*, vol. 173, no. 2-3, pp. 259-270, 2004.
[22] Y. Liu, S.-H. Lee, and T.-S. Chon, "Analysis of Behavioral Changes of Zebrafish (*Danio rerio*) in Response to Formaldehyde Using Self-Organizing Map and a Hidden Markov Model," *Ecological Modelling*, vol. 222, no. 14, pp. 2191-2201, 2011.
[23] Y. Li, J.-M. Lee, T.-S. Chon, Y. Liu, H. Kim, M.-J. Bae, and Y.-S. Park, "Analysis of Movement Behavior of Zebrafish (*Danio rerio*) under Chemical Stress Using Hidden Markov Model," *Modern Physics Letters B*, vol. 27, no. 2, pp. 1 350 014:1-1 350 014:13, 2013.
[24] "Zebrabox," http://www.vplsi.com/content.php?content.72, 2014.
[25] J.J. Ingebretson and M.A. Masino, "Quantification of Locomotor Activity in Larval Zebrafish: Considerations for the Design of High-Throughput Behavioral Studies," *Frontiers in Neural Circuits*, vol. 7, pp. 109:1—109:9, 2013.
[26] H.M. Ashtawy and N.R. Mahapatra, "A Comparative Assessment of Ranking Accuracies of Conventional and Machine-Learning-Based Scoring Functions for Protein-Ligand Binding Affinity Prediction," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 9, no. 5, pp. 1301-1313, Sept./Oct. 2012.
[27] K. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
[28] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1-27:27, 2011.
[29] "ZFIN: The Zebrafish Model Organism Database," http://zfin.org/ZDB-GENO-960809-7, 2014.
[30] M. Westerfield, *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio rerio)*. Univ. of Oregon Press, 2000.
[31] Y.-Y. Huang and S.C.F. Neuhauss, "The Optokinetic Response in Zebrafish and Its Applications," *Frontiers in Bioscience: A J. and Virtual Library*, vol. 13, pp. 1899-1916, 2008.
[32] S.E. Brockerhoff, "Measuring the Optokinetic Response of Zebrafish Larvae," *Nature Protocols*, vol. 1, no. 5, pp. 2448-2451, 2006.
[33] C. Nusslein-Volhard and R. Dahm, *Zebrafish: A Practical Approach*. Oxford Univ. Press, 2002.
[34] J. Lin, E. Keogh, and L. Wei, "Experiencing SAX: A Novel Symbolic Representation of Time Series," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 107-144, 2007.
[35] J.S. Richman and J.R. Moorman, "Physiological Time-Series Analysis Using Approximate Entropy and Sample Entropy," *Am. J. Physiology—Heart and Circulatory Physiology*, vol. 278, pp. H2039-H2049, 2000.

Yuan Gao received the BS degree in biomedical engineering in 2009 and the MS degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 2012. He is currently working toward the PhD degree in the Department of Electronic Engineering, City University of Hong Kong. His current interests include machine learning, pattern recognition, and applications.

Rosa H.M. Chan received the BEng (first Hons.) degree in automation and computer-aided engineering and a minor in computer science from the Chinese University of Hong Kong in 2003. She received the PhD degree in biomedical engineering from the University of Southern California (USC) in 2011, where she also received the MS degrees in biomedical engineering, electrical engineering, and aerospace engineering. She is currently an assistant professor in the Department of Electronic Engineering, City University of Hong Kong. Her research interests include mathematical modeling of neural system, development of neural prosthesis, and brain-machine interface applications. She received the Croucher Scholarship and Sir Edward Youde Memorial Fellowship for Overseas Studies in 2004. In the summer of 2010, she received Google Scholarship and participated in the Singularity University Graduate Studies Program at NASA AMES.

Tommy W.S. Chow (M'93-SM'03) received the BSc (First Hons.) and the PhD degrees from the University of Sunderland, Sunderland, United Kingdom. He joined the City University of Hong Kong, Hong Kong, as a lecturer in 1988. He is currently a professor in the Electronic Engineering Department. His research interests include the area of machine learning including supervised and unsupervised learning, data mining, pattern recognition and fault diagnostic. He worked for NEI Reyrolle Technology at Hebburn, England, developing digital simulator for transient network analyser. He then worked on a research project involving high current density current collection system for superconducting direct current machines, in collaboration with the Ministry of Defense (Navy) at Bath, England, and the International Research and Development at Newcastle upon Tyne. He has authored or coauthored more than 170 technical papers in international journals, 5 book chapters, and more than 60 technical papers in international conference proceedings.

Liyun Zhang received bachelor of medicine degree from the Beijing Medical University, China (current name: Peking University, Health Science Center) in 1995. She worked as an ophthalmologist in the Department of Ophthalmology at Beijing Tong Ren Hospital from 1995 to 2004. During the clinical practice, she also attended a Master program at Capital University of Medical Sciences, Beijing, China, and received her master of medicine degree in 2003. She received the PhD degree in ophthalmology from The Chinese University of Hong Kong (CUHK) in 2007. She pursued her postdoctoral research at Purdue University, Indiana, USA from 2009 to 2012. During this period, she received a Pediatric Ophthalmology Research Grant from the Knights Templar Eye Foundation in 2010 and a Charles D. Kelman, MD Scholar Award from The International Retinal Research Foundation in 2011. Currently, she is working as a postdoctoral research fellow at the University of Cincinnati, Ohio, USA. Her research interest is identifying new treatments for eye diseases.

Sylvia Bonilla received the BS degree in biology and a minor in organic and biochemistry from the California State University, Dominguez Hills, in 2008. She is currently working toward the PhD degree in the Biological Sciences Department at Purdue University.

Chi-Pui Pang received the BSc degree in biochemistry in 1978 from the University of London and the DPhil degree in 1982 from the University of Oxford on an EP Abraham Research Fund Scholarship, followed by two years postdoctoral research in Oxford. He is S.H. Ho professor of Visual Sciences, a professor of Ophthalmology and Visual Sciences, and the chairman of the Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong. He is also the director of The Shantou University/The Chinese University of Hong Kong Joint Shantou International Eye Center. His research interests include genomic studies and gene mapping of glaucoma, retinal diseases, myopia, congenital cataracts, retinoblastoma, thyroid-associated orbitopathy, diabetic retinopathy, retinitis pigmentosa, uveitis, and corneal dystrophies. He also works on ocular stem cells and herbal molecules on their effects in eye diseases. He has more than 310 publications and 13 book chapters. He is a reviewer for the Wellcome Trust (UK), National Eye Institute (USA), National Medical Research Council (Singapore), Health Research Board (Ireland), Catalan Agency for Health Technology Assessment and Research (Spain), and National Science Foundation, China. He is honorary or visiting professor of more than 20 clinical or research institutions in mainland China.

Mingzhi Zhang has been working on ophthalmology for more than 30 years, with subspecialty in Cataract, Glaucoma, and Optometry. She is the pioneer of phacoemulsification and refractive surgery for cataract. Her study of surgical treatment of glaucoma with cataract has attracted great attention in these fields. Recently, by combining both basic research and clinical research together, she focused on mapping of disease causing genes of congenital cataract as well as susceptibility genes of glaucoma and high myopia. She has conducted six research projects founded national wide, two International cross-cutting projects, three provincial research projects in recent years. One-hundred thirty one papers have been published in national and provincial journals, of which there are 61 papers accepted in international peer reviewed SCI journals.

Yuk Fai Leung received the BSc (first Hons.) and the MPhil degrees in biochemistry from the Hong Kong University of Science and Technology in 1996 and 1998, respectively. He received the PhD degree in ophthalmology from the Chinese University of Hong Kong in 2002. He received a Croucher Foundation Postdoctoral Fellowship the same year and pursued his postdoctoral research at Harvard University until 2007. In 2005, he was awarded with a Pediatric Ophthalmology Research Grant from the Knights Templar Eye Foundation. In 2008, he established his own research group at Purdue University in the Department of Biological Sciences. He also received a Hope for Vision Award in the same year. His current research focuses on using zebrafish eye disease models to elucidate disease-causing gene network and identify new drug therapies