Supplementary Material for:

Computational classification of different wild-type zebrafish strains based on their variation in light-induced locomotor response

Yuan Gao, Gaonan Zhang, Beth Jelfs, Robert Carmer, Prahatha Venkatraman, Mohammad Ghadami, Skye A. Brown, Chi Pui Pang, Yuk Fai Leung, Rosa H. M. Chan, Mingzhi Zhang

S1. Technical details

S1.1. Sample Entropy

Sample entropy is used as a feature to quantify the randomness or complexity of the movement (Richman and Moorman, 2000). Explicitly, given time series data $x_1,...,x_T$, we define a local (sub) signal of length m, i.e. $X_m(i)=\{x_i, x_{i+1},..., x_{i+m-1}\}$. Then, for every $i, j=1,...,T$ and $i \neq j$, we calculate the Chebyshev distance between $X_m(i)$ and $X_m(j)$, $D_{Chebyshev}(X_m(i), X_m(j))$ and the Chebyshev distance between $X_{m+1}(i)$ and $X_{m+1}(j)$, $D_{Chebyshev}(X_{m+1}(i), X_{m+1}(j))$. The Chebyshev distance for vectors $X_m(i)$ and $X_m(j)$ can be defined by Eq. (1) (Cantrell, 2000):

$$D_{Chebyshev}(X_m(i), X_m(j)) = \max_k(|x_{i+k} - x_{j+k}|), \text{ for } k=0,...,m-1. \quad (1)$$

After that, for a predefined value r, we count the number of local (sub) vector pairs with different lengths (i.e. for length m and $m+1$) that have Chebyshev distance that smaller than r, i.e. we count number of local (sub) vector pairs with $D_{Chebyshev}(X_{m+1}(i), X_{m+1}(j)) < r$ as A and these with $D_{Chebyshev}(X_m(i), X_m(j)) < r$ as B. Finally, the Sample Entropy can be calculated as Eq. (2):

$$SampEn = -\log\frac{A}{B}. \quad (2)$$

In our implementation, we followed the general suggestion to set $m=2$ and $r=0.2\times std$ (Richman and Moorman, 2000), where std is the standard deviation of the whole signal.

S1.2. Support Vector Machine

SVM is a powerful supervised learning method which aims to avoid overfitting through the use of the support vectors. In our model, we used *C*-SVM with Radial Basis Function (RBF) kernel (Bishop, 2006, Chang and Lin, 2011) for non-linear classification.

The kernel *C*-SVM problem can be formulated as Eq. (3):

$$\begin{aligned} &\min_{\omega,b,\varepsilon_i} \frac{1}{2}\omega^{\mathrm{T}}\omega + c\sum_{i=1}^{N}\varepsilon_i \\ &\text{s.t. } y_i(\omega^{\mathrm{T}}\phi(x_i)+b) \geq 1-\varepsilon_i, \quad \varepsilon_i \geq 0, \quad i=1,...,N, \end{aligned} \tag{3}$$

where x_i is a training vector, and y_i is a class indicator such that $y_i \in \{-1, 1\}$. ε_i is a slack variable, and $C > 0$ controls the trade-off between the slack variable penalty and the margin. ω and b are the weight and bias vectors of the linear discriminant function, and ϕ is a predefined feature-space transformation.

By applying the Lagrange multiplier, the dual problem of SVM becomes as Eq. (4):

$$\begin{aligned} &\min_{\alpha_i,\alpha_j} \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j) - \sum_{i=1}^{N}\alpha_i \\ &\text{s.t. } \sum_{i=1}^{N}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i=1,...,N, \end{aligned} \tag{4}$$

where K denotes the predefined kernel, i.e. $K(x_i,x_j) = \phi(x_i)^{\mathrm{T}}\phi(x_j)$, and α_i, α_j are the Lagrange multipliers. Here we use RBF kernel $K(x_i,x_j) = e^{-\gamma\|x_i-x_j\|^2}$, in which γ is a manual parameter which governs the width of the kernel (Bishop, 2006).

After the problem Eq. (4) is solved, the decision function can be given in Eq. (5):

$$\mathrm{sgn}(\omega^{\mathrm{T}}\phi(x_i)+b) = \mathrm{sgn}(\sum_{i=1}^{N}\alpha_i y_i K(x_i,x_j)+b) \tag{5}$$

In our experiment, the slack parameter *C* is set to 2 and the kernel width γ is set to 0.125.

S2. The detailed classification results

The detailed classification results using 3 Nearest Neighbour (3NN), Naïve Bayes (NB), Support Vector Machine (SVM) and Expectation-Maximization method on Gaussian Mixture Model (EM-GMM) for 3 to 9 dpf zebrafish strains are given in Table S1-S2. The results demonstrate that SVM outperformed other methods for the majority of cases.

Stage	Classification Problem	Accuracy			
		3NN	Naive Bayes	SVM	EM-GMM
3 dpf	TLAB vs. TL	55.31%	58.21%	**62.05%**	57.72%
	TLAB vs. AB	64.49%	65.75%	**69.99%**	65.25%
	TL vs. AB	67.82%	65.19%	**71.03%**	64.27%
	TLAB vs. TL vs. AB	44.24%	44.60%	**51.44%**	43.84%
4 dpf	TLAB vs. TL	70.33%	71.30%	**73.18%**	71.15%
	TLAB vs. AB	59.87%	61.88%	**63.33%**	61.59%
	TL vs. AB	75.95%	76.90%	**80.78%**	75.95%
	TLAB vs. TL vs. AB	52.93%	55.10%	**57.79%**	54.84%
5 dpf	TLAB vs. TL	74.61%	76.83%	**79.75%**	76.44%
	TLAB vs. AB	78.53%	83.89%	83.31%	**83.89%**
	TL vs. AB	63.97%	65.00%	**70.79%**	64.25%
	TLAB vs. TL vs. AB	56.96%	61.68%	**63.04%**	61.21%
6 dpf	TLAB vs. TL	63.96%	64.38%	**65.63%**	63.66%
	TLAB vs. AB	56.91%	62.82%	**66.09%**	62.38%
	TL vs. AB	62.34%	67.63%	**69.29%**	66.42%
	TLAB vs. TL vs. AB	43.51%	51.33%	**51.98%**	49.55%
7 dpf	TLAB vs. TL	60.00%	66.57%	**67.92%**	65.51%
	TLAB vs. AB	60.28%	62.58%	**63.39%**	62.08%
	TL vs. AB	74.89%	77.04%	**79.38%**	76.98%
	TLAB vs. TL vs. AB	47.30%	52.67%	**55.22%**	51.71%
8 dpf	TLAB vs. TL	64.34%	64.29%	**68.96%**	64.30%
	TLAB vs. AB	78.14%	78.87%	**79.62%**	78.97%
	TL vs. AB	87.05%	90.52%	**90.61%**	90.47%
	TLAB vs. TL vs. AB	61.15%	63.07%	**65.36%**	62.69%
9 dpf	TLAB vs. TL	67.20%	70.11%	**75.85%**	69.37%
	TLAB vs. AB	64.02%	**72.98%**	72.27%	72.51%
	TL vs. AB	87.22%	87.98%	**89.45%**	88.01%
	TLAB vs. TL vs. AB	53.29%	62.20%	**64.82%**	61.72%

Table S1. The classification accuracies of different methods for 3–9 dpf zebrafish using first 30-sec LLR (VMR). 3NN is the 3 Nearest Neighbour classifier, SVM is Support Vector Machine and EM-GMM is the Expectation-Maximization method on Gaussian Mixture Model. The **bold** value is the highest value of each line. In most cases, SVM outperforms the other classification methods.

Stage	Classification Problem	Accuracy			
		3NN	Naive Bayes	SVM	EM-GMM
3 dpf	TLAB vs. TL	77.77%	78.62%	**81.79%**	78.26%
	TLAB vs. AB	56.77%	54.23%	**61.84%**	52.70%
	TL vs. AB	67.63%	63.98%	**69.62%**	62.15%
	TLAB vs. TL vs. AB	48.65%	52.46%	**54.98%**	52.26%
4 dpf	TLAB vs. TL	70.14%	68.76%	**74.69%**	68.55%
	TLAB vs. AB	59.97%	63.88%	**65.15%**	63.03%
	TL vs. AB	70.45%	71.02%	**72.58%**	70.43%
	TLAB vs. TL vs. AB	52.17%	52.83%	**56.21%**	51.84%
5 dpf	TLAB vs. TL	76.98%	77.92%	**82.77%**	77.78%
	TLAB vs. AB	81.67%	81.23%	**83.65%**	81.01%
	TL vs. AB	69.59%	66.26%	**72.41%**	66.05%
	TLAB vs. TL vs. AB	64.27%	62.46%	**67.45%**	62.21%
6 dpf	TLAB vs. TL	71.15%	65.84%	**71.95%**	64.23%
	TLAB vs. AB	74.19%	69.11%	**77.89%**	68.31%
	TL vs. AB	72.34%	61.59%	**72.38%**	61.12%
	TLAB vs. TL vs. AB	**60.92%**	49.40%	60.79%	48.65%
7 dpf	TLAB vs. TL	73.70%	71.69%	**77.41%**	70.13%
	TLAB vs. AB	71.07%	64.90%	**75.16%**	64.49%
	TL vs. AB	77.62%	68.43%	**78.73%**	68.34%
	TLAB vs. TL vs. AB	60.89%	54.27%	**64.69%**	53.78%
8 dpf	TLAB vs. TL	71.21%	64.23%	**74.39%**	62.84%
	TLAB vs. AB	71.32%	74.05%	**74.59%**	73.94%
	TL vs. AB	**87.70%**	83.70%	85.95%	83.44%
	TLAB vs. TL vs. AB	60.37%	58.59%	**65.14%**	57.76%
9 dpf	TLAB vs. TL	69.82%	58.69%	**74.23%**	57.30%
	TLAB vs. AB	**68.74%**	66.81%	67.04%	64.46%
	TL vs. AB	77.73%	69.41%	**80.43%**	68.66%
	TLAB vs. TL vs. AB	58.47%	49.83%	**60.14%**	48.32%

Table S2. The classification accuracies of different methods for 3–9 dpf zebrafish using all 30-min LLR. 3NN is the 3 Nearest Neighbour classifier, SVM is Support Vector Machine and EM-GMM is the Expectation-Maximization method on Gaussian Mixture Model. The **bold** value is the highest value of each line. In most cases, SVM outperforms the other classification methods.

S3. Plots of mean activity

In the following figures, S1-S7, we illustrate the plots of mean activity for the different zebrafish strains (i.e. TL, AB and TLAB) from 3 to 9 days post-fertilization (dpf) for both light-ON and light-OFF stimuli.

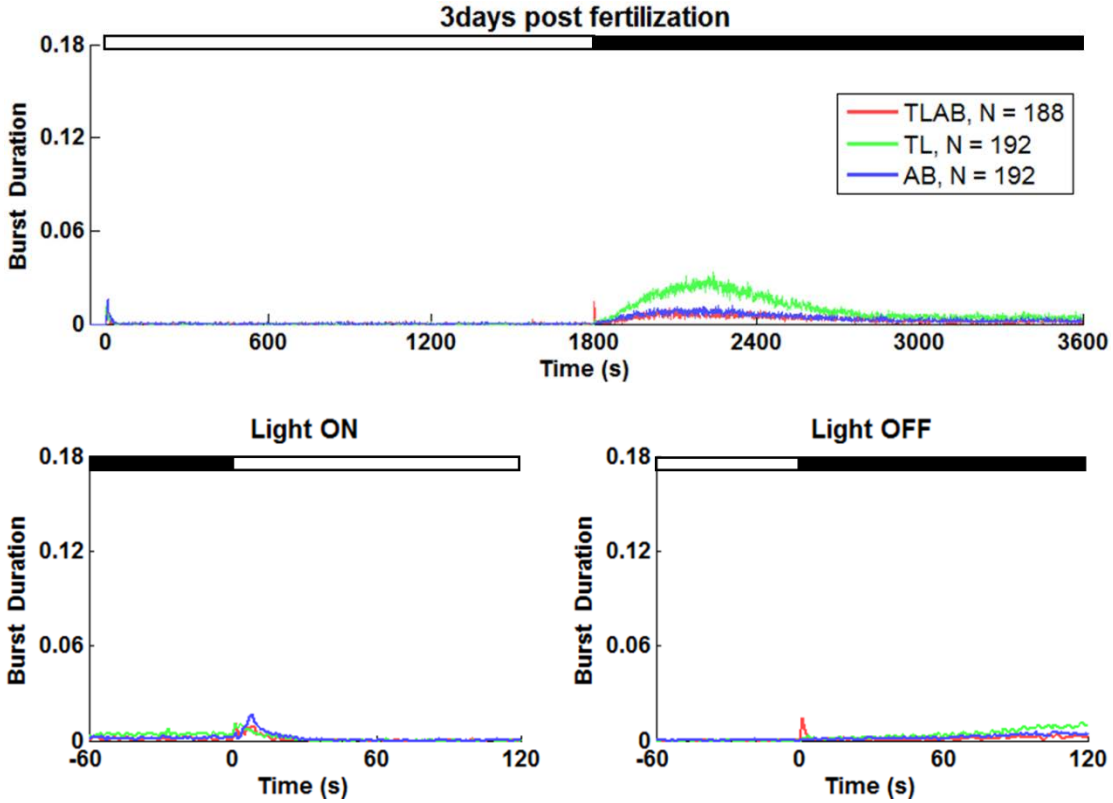

Figure S1. Plots of the mean activity for each of the TLAB, TL and AB zebrafish strains at 3 dpf. Top Subfigure: the overall plots of the averaged three light-ON and light-OFF trials; Bottom Left Subfigure: the mean plots showing 1 minute before and 2 minutes after light-ON stimulus; Bottom Right Subfigure: the mean plots showing 1 minute before and 2 minutes after light-OFF stimulus.

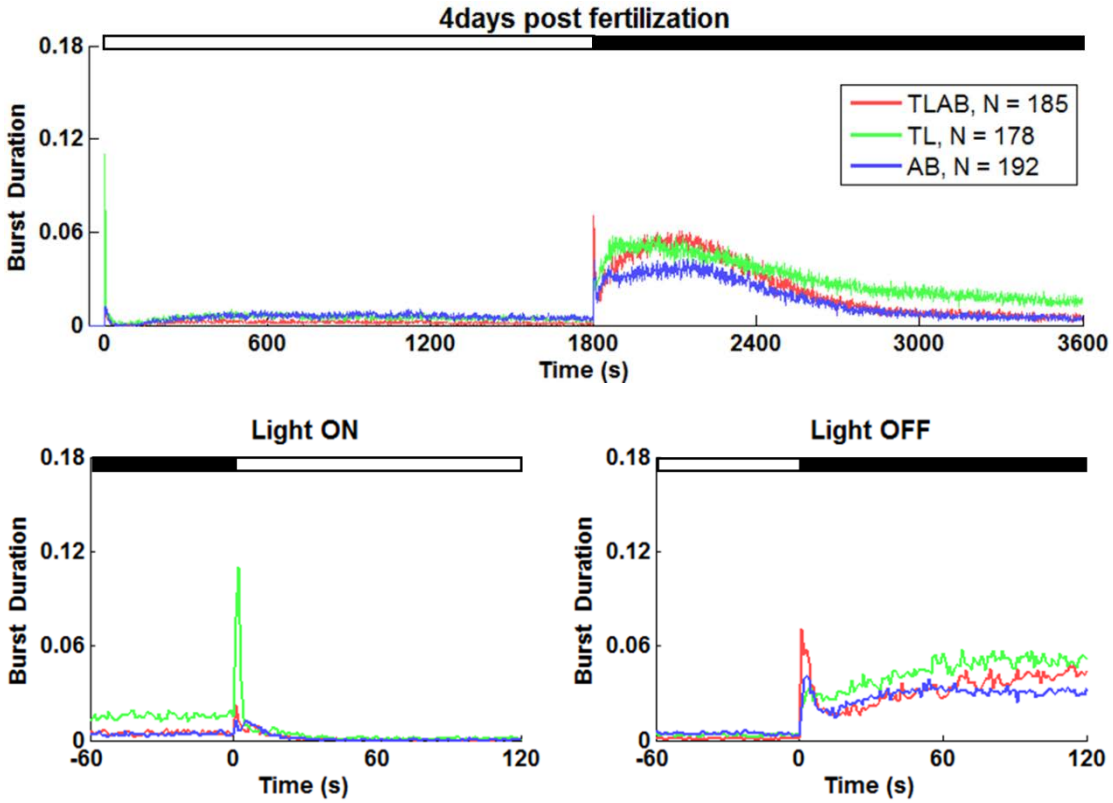

Figure S2. Plots of the mean activity for each of the TLAB, TL and AB zebrafish strains at 4 dpf. Top Subfigure: the overall plots of the averaged three light-ON and light-OFF trials; Bottom Left Subfigure: the mean plots showing 1 minute before and 2 minutes after light-ON stimulus; Bottom Right Subfigure: the mean plots showing 1 minute before and 2 minutes after light-OFF stimulus.

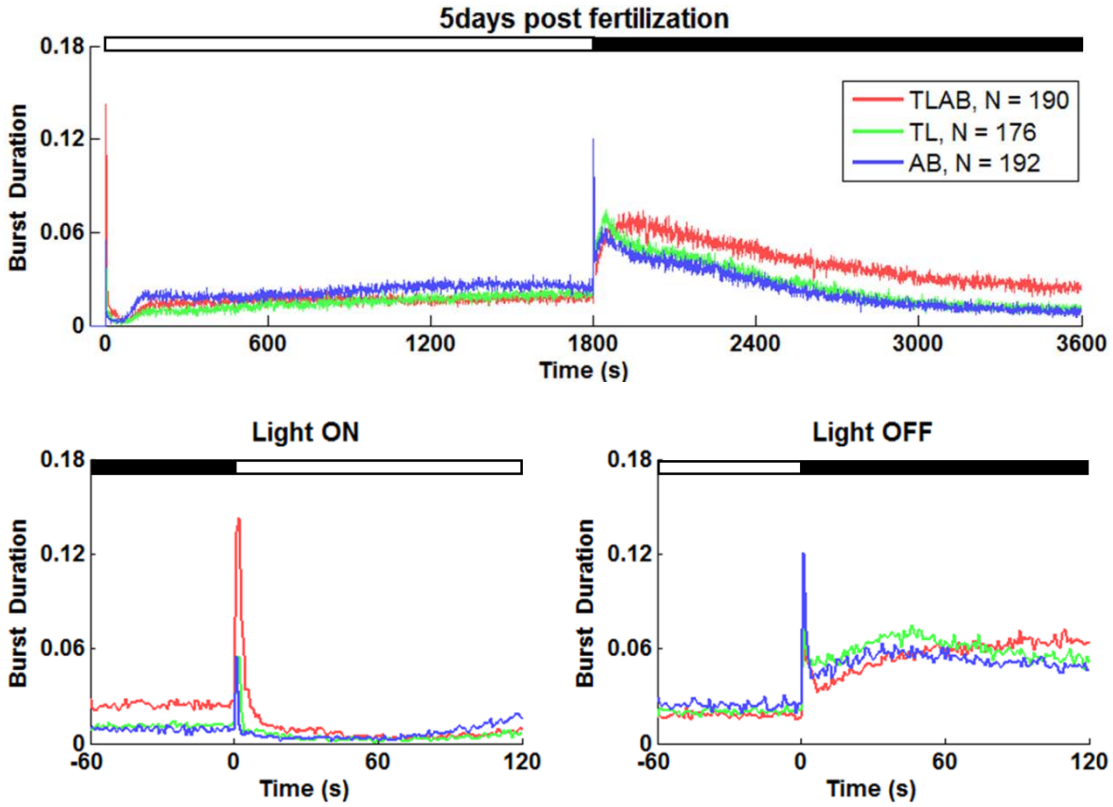

Figure S3. Plots of the mean activity for each of the TLAB, TL and AB zebrafish strains at 5 dpf. Top Subfigure: the overall plots of the averaged three light-ON and light-OFF trials; Bottom Left Subfigure: the mean plots showing 1 minute before and 2 minutes after light-ON stimulus; Bottom Right Subfigure: the mean plots showing 1 minute before and 2 minutes after light-OFF stimulus.

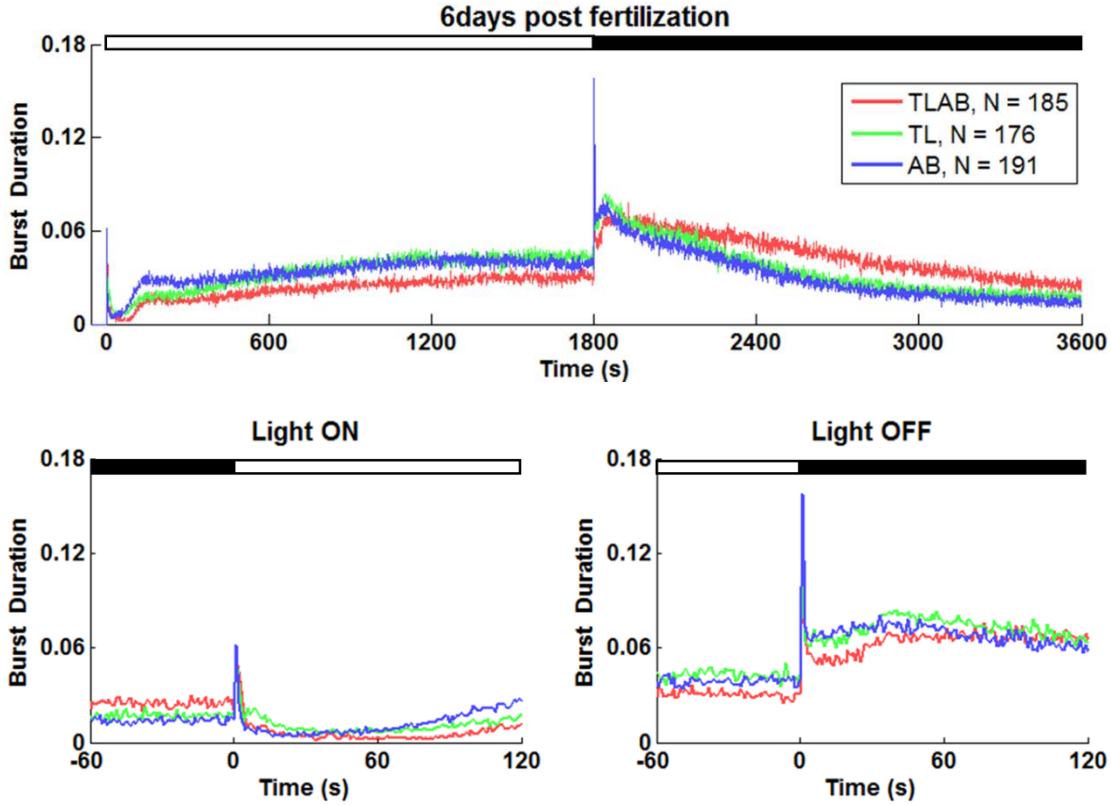

Figure S4. Plots of the mean activity for each of the TLAB, TL and AB zebrafish strains at 6 dpf. Top Subfigure: the overall plots of the averaged three light-ON and light-OFF trials; Bottom Left Subfigure: the mean plots showing 1 minute before and 2 minutes after light-ON stimulus; Bottom Right Subfigure: the mean plots showing 1 minute before and 2 minutes after light-OFF stimulus.

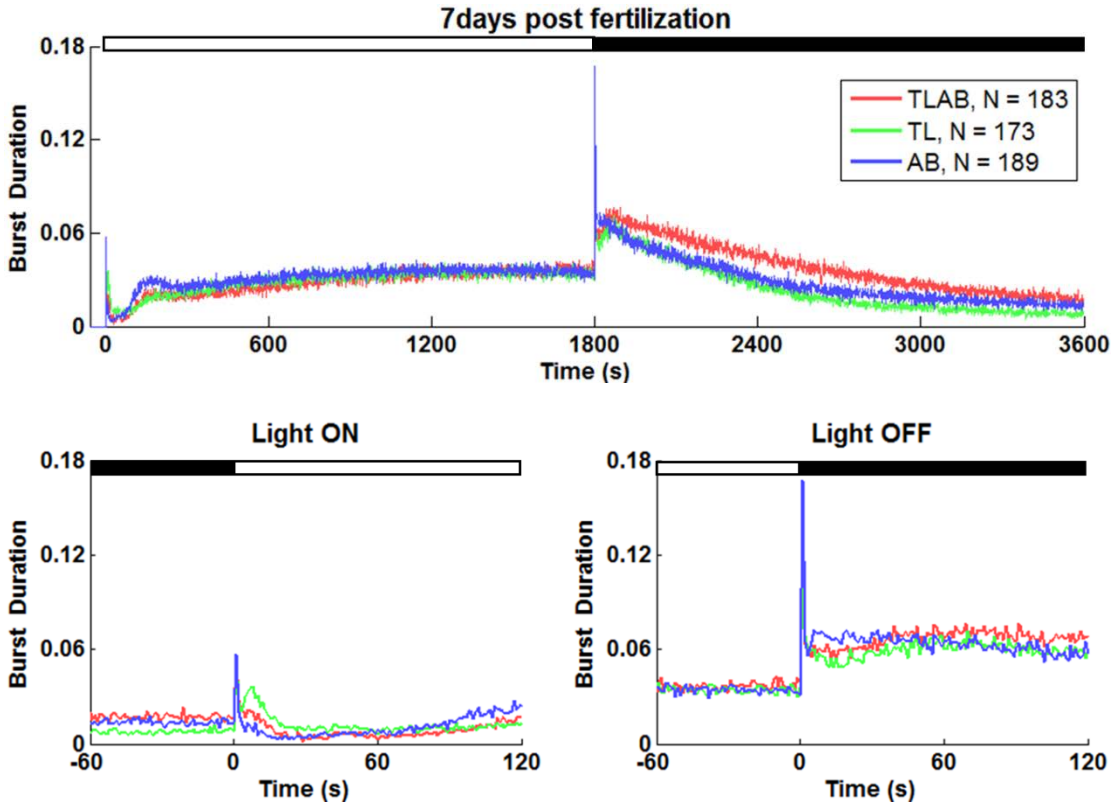

Figure S5. Plots of the mean activity for each of the TLAB, TL and AB zebrafish strains at 7 dpf. Top Subfigure: the overall plots of the averaged three light-ON and light-OFF trials; Bottom Left Subfigure: the mean plots showing 1 minute before and 2 minutes after light-ON stimulus; Bottom Right Subfigure: the mean plots showing 1 minute before and 2 minutes after light-OFF stimulus.

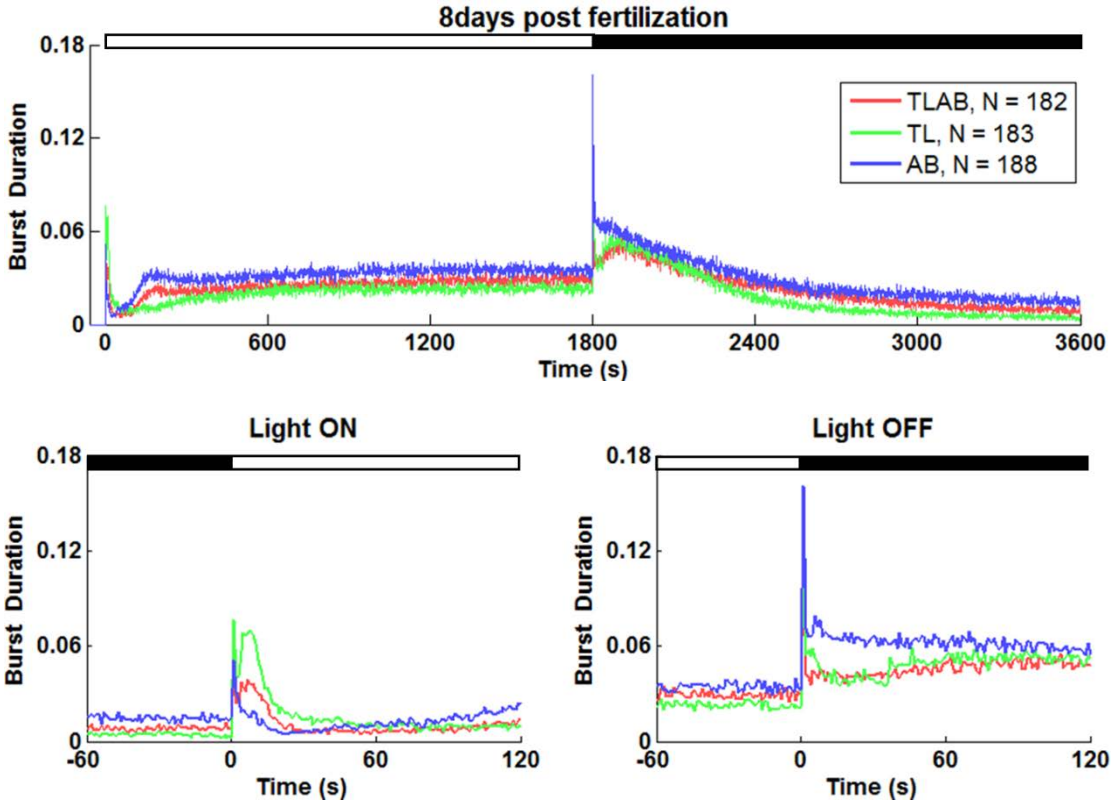

Figure S6. Plots of the mean activity for each of the TLAB, TL and AB zebrafish strains at 8 dpf. Top Subfigure: the overall plots of the averaged three light-ON and light-OFF trials; Bottom Left Subfigure: the mean plots showing 1 minute before and 2 minutes after light-ON stimulus; Bottom Right Subfigure: the mean plots showing 1 minute before and 2 minutes after light-OFF stimulus.

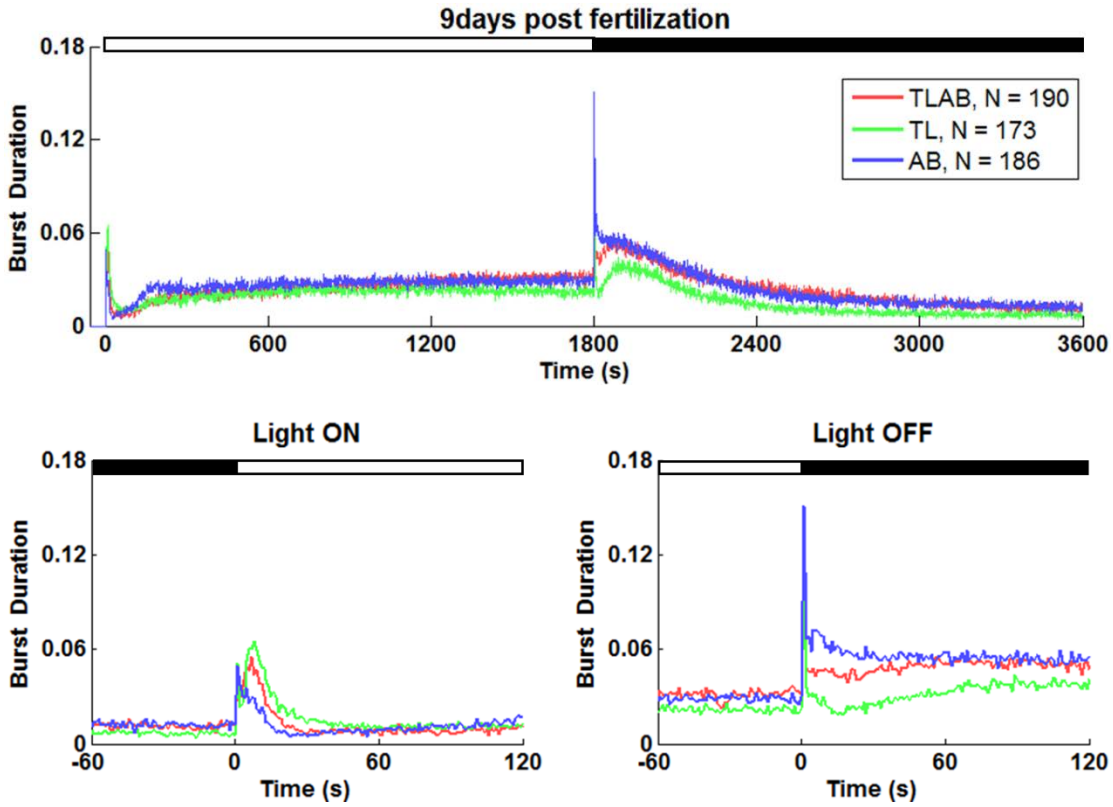

Figure S7. Plots of the mean activity for each of the TLAB, TL and AB zebrafish strains at 9 dpf. Top Subfigure: the overall plots of the averaged three light-ON and light-OFF trials; Bottom Left Subfigure: the mean plots showing 1 minute before and 2 minutes after light-ON stimulus; Bottom Right Subfigure: the mean plots showing 1 minute before and 2 minutes after light-OFF stimulus.