computation

MDPI

Article

A Parallel Computing Approach to Gene Expression and Phenotype Correlation for Identifying *Retinitis Pigmentosa* Modifiers in *Drosophila*

Chawin Metah [1], Amal Khalifa [1,*] and Rebecca Palu [2]

[1] Department of Computer Science, Purdue University Fort Wayne, Fort Wayne, IN 46805, USA; metac01@pfw.edu
[2] Department of Biology, Purdue University Fort Wayne, Fort Wayne, IN 46805, USA; palur@pfw.edu
* Correspondence: khalifaa@pfw.edu

Abstract: As a genetic eye disorder, *retinitis pigmentosa* (RP) has been a focus of researchers to find a diagnosis through either genome-wide association (GWA) or RNAseq analysis. In fact, GWA and RNAseq are considered two complementary approaches to gaining a more comprehensive understanding of the genetics of different diseases. However, RNAseq analysis can provide information about the specific mechanisms underlying the disease and the potential targets for therapy. This research proposes a new approach to differential gene expression (DGE) analysis, which is the heart of the core-analysis phase in any RNAseq study. Based on the *Drosophila* Genetic Reference Panel (DGRP), the gene expression dataset is computationally analyzed in light of eye-size phenotypes. We utilized the foreach and the doParallel R packages to run the code on a multicore machine to reduce the running time of the original algorithm, which exhibited an exponential time complexity. Experimental results showed an outstanding performance, reducing the running time by 95% while using 32 processes. In addition, more candidate modifier genes for RP were identified by increasing the scope of the analysis and considering more datasets that represent different phenotype models.

Keywords: retinal apoptosis; ER stress; modifier genes; gene expression; phenotypic variation; degenerative models; parallel computing; multithreading

Citation: Metah, C.; Khalifa, A.; Palu, R. A Parallel Computing Approach to Gene Expression and Phenotype Correlation for Identifying *Retinitis Pigmentosa* Modifiers in *Drosophila*. *Computation* **2023**, *11*, 118. https://doi.org/10.3390/computation11060118

Academic Editor: Rainer Breitling

Received: 16 May 2023
Revised: 5 June 2023
Accepted: 6 June 2023
Published: 14 June 2023

1. Introduction

Retinitis pigmentosa (RP) is a genetic eye disorder that causes progressive degeneration of the cells in the retina, leading to significant visual impairment or blindness in some cases. RP is caused by mutations in genes that are essential for the normal functioning of the retina, and there are several different genetic forms of the disorder. There is currently no cure for RP, but there are treatments that can help manage the symptoms and slow the progression of the disease [1].

Recent studies on the human genetic database show a significant comparability between *Drosophila Melanogaster*, commonly known as the fruit fly, and human genes. Thus, the *Drosophila* Genetic Reference Panel (DGRP) was created in 2007, which includes a standardized set of over 200 inbred lines and more than 20,000 annotated genes. It provides a powerful resource for geneticists and evolutionary biologists to investigate, in a controlled setting, the genetic basis of a wide range of traits including behaviors, morphology, and disease susceptibility [2,3]. For example, by exposing the DGRP strains to environmental conditions that mimic the effects of RP, such as exposure to light-induced oxidative stress, researchers can observe differences in the severity and progression of the disease across the different strains.

Through genome-wide association (GWA) studies, researchers can then identify genetic variants that are significantly associated with disease as well as potential targets for

therapeutic intervention. In their GWA, Chow et al. [4] identified 100 candidate genes that were linked to retinal degeneration and may play a role in the progression of RP. However, this type of analysis cannot fully identify potential modifiers that are indirectly related to changes in eye size and may be differentially regulated several steps downstream of the associated gene. In this case, RNAseq analysis can be used in ways that are complementary to GWA to investigate the genetics of RP [5]. While GWA focuses on identifying genetic variants associated with the disease, RNAseq analysis can be used to investigate the effects of these variants on gene expression and pathway dysregulation [6]. This provides insights into the specific mechanisms underlying the disease and potential targets for therapy. Additionally, RNAseq analysis can be used to validate the functional relevance of genetic variants identified by GWA [7].

As explained by Conesa et al. [5], differential gene expression (DGE) analysis is the heart of the core-analysis phase in any RNAseq study. Various techniques have been developed to evaluate expression data among multiple samples and identify suspect genes. Some of the methods rely on specific discrete probability distributions, such as the Poisson and negative binomial distributions [8]. edgeR is a comprehensive package in R that takes raw read data and carries out both normalization and differential expression analysis simultaneously [9]. More packages, including baySeq [10] and EBseq [11], on the other hand, follow the Bayesian negative binomial paradigm to identify differential gene expression. Furthermore, methods such as NOISeq [12] and SAMseq [13] are non-parametrically designed and introduce fewer false premises and estimations based on existing data. In a comparative study by Soneson and Delorenzi [14], several DGE approaches were investigated using simulated and real RNA-seq data. The results showed that all methods performed better for large sample sizes using more than three samples per condition for an RNA-seq experiment. EBSeq showed the strongest dependency on the sample size while DESeq offered the best results for smaller sample size datasets.

An alternative approach would be observing extreme phenotypes (such as extremely large or small eye sizes) and trying to understand how they are linked to changes in gene expression when they are significantly altered in strains. Although this approach has been used successfully to identify modifier genes in a variety of diseases and cancers [15,16], it has not yet been widely applied to *Drosophila*. Two previous studies investigated candidate RP genes using eye-size phenotypes and the gene expression level of DGRP. Amstutz et al. [17] utilized the unsupervised learning algorithm of K-Mean clustering to group strains based on the profile of their expression levels. The extreme minimum and maximum eye-size profiles were selected from a total of six clusters. The two replicates for each selected line were averaged and passed to the differential gene expression analysis algorithm. The study further validated the RP candidate genes by comparing the control eye sizes with RNA interference (RNAi) strains.

On the other hand, Nguyen et al. [18] proposed selecting sixteen strains that exhibit extreme eye size and generating a list of all their possible replicate combinations instead of using the average expression value between two replicates of a DGRP strain. The "best combination" is then selected, such that only the strain replicate that offers the best expression/eye-size correlation among all the genes is selected and considered for further analysis. However, the main issue is the intensive computations required to find the best replicate, especially when considering a large number of strains.

The objective of this research is to speed up the execution time of Nguyen's approach by exploiting some parallel computing constructs in the R language. Hence, the main contributions are (1) restructuring the code to run in parallel on a multicore processor, (2) filtering out the input datasets before the analysis, and (3) using the *p*-value to assess the statistical significance of the results. The rest of the paper is organized as follows: Section 2 describes the structure of the input datasets and the steps of the proposed computational approach. In Section 3, the results of the study are presented and discussed in the light of several experiments considering different performance criteria. Potential candidate modifiers are then identified and their relation to eye development and degeneration are

highlighted in Section 4. Finally, Section 5 summarizes the conclusions as well as some future research directions.

2. Materials and Methods

2.1. Input Datasets

As the goal of this research was to explore the correlation between the genotype of RP and the phenotype of degenerative eye-size, we needed two types of datasets. The first is the DGRP gene expression dataset which was developed by Huang et al. [19], measuring the expression levels for 185 DGRP strains over 18,140 genes. As shown in Table 1, this dataset is represented by a matrix with the genes as rows and the DGRP stains as columns. The gene names have the FBgn/XLOC prefix followed by an identification number. In fact, the FBgn genes have been recognized and fully annotated in Flybase, while little is known about the XLOC genes. In addition, the values stored in the matrix cells are the RNA expression value for each strain. With two replicates for each strain line, the columns are annotated using the postfix ":1" and ":2" representing the first and the second replicate, respectively.

Table 1. A sample of the DGRP expression-level dataset.

Gene	Expression Level			
	RAL021:1	**RAL021:2**	**RAL026:1**	**RAL026:2**
FBgn0000014	4.244723137	4.216353088	4.028685457	3.965513774
FBgn0000015	3.234859699	3.199773952	3.266073855	3.514853684
FBgn0000017	8.066864662	7.962031505	8.016965853	8.081375654
FBgn0000018	5.317033088	5.268665083	5.583749674	4.949218486
FBgn0000022	3.000683083	3.000127343	4.033542617	3.364429304
FBgn0000024	6.120670813	6.023183171	6.363472661	6.83930746
FBgn0000028	4.101309578	4.050933404	4.581349626	4.276622648
FBgn0000032	7.460913282	7.68689799	7.782455553	7.635495636
FBgn0000036	3.988090417	3.789139103	3.979189512	3.95396714
FBgn0000037	4.475747359	4.323271618	4.457239171	4.378994365
...	...	...	...	...
XLOC_006439	2.414951288	2.612959863	3.717652528	2.090561202

The eye-size phenotype, on the other hand, is captured by three different datasets: $Rh1^{G69D}$, *rpr*, and *p53*. Those resulted from studies that observed different apoptosis (programmed cell death)-induced retinal degeneration models across the DGRP. Chow et al. [4] studied the $Rh1^{G69D}$ model of degeneration, representing 173 DGRP lines. Overexpression of *p53* or reaper (*rpr*) were both studied by Palu et al. [20]. The 204 strains from the DGRP were used for the GMR > *p53* study and 202 were used for the GMR-*rpr* study. The average two-dimensional eye area was measured for at least 10 female flies that were flash-frozen prior to imaging. The eye area is defined as the area of the polygon connecting certain landmark points around the eye. Usually, the eye area is manually measured using tools such as Adobe Photoshop on frontal images of the head [21]. As shown in Table 2, the three datasets are saved in a text file that contains two columns: strain IDs and the average eye size measured in pixels $\times 10^3$. A strain ID comprises an indicator, "RAL", followed by a three-digit identification number. It is also worth noting that some strains in eye-size datasets don't exist in the expression-level matrix. Hence, only the intersection between the two will be analyzed in each case.

Table 2. A sample of the eye-size dataset.

Strain ID	Average Eye Size (Pixels $\times 10^3$)
RAL021	19,976.8
RAL026	21,473.2
RAL038	19,981.5
RAL040	16,992.9
RAL042	21,481.4
...	...
RAL913	19,488.5

2.2. Correlation Analysis

The correlation coefficient is used in the fields of science and finance to statistically assess the relationship between two variables, factors, or datasets. Its values can range from -1 to 1, where -1 indicates a perfect inverse correlation, with values in one series rising as those in the other decline, and vice versa. A value of 1 shows a perfect direct correlation, and the value 0 means there is no linear relationship. Although Pearson's formula is the most commonly used correlation method, it cannot assess nonlinear associations between variables or those arising from sampled data not subject to a normal distribution. In these cases, nonparametric methods such as Spearman's, the Kendall rank, or a polychronic correlation coefficient can be used to analyze the relationships.

Pearson's correlation is a covariance between two samples divided by product of the standard deviation of those samples. The covariance and standard deviation can be calculated by using Equations (1) and (2), respectively. Then, we can use those equations to obtain the correlation coefficient as shown in Equation (3). The variables x_i and y_i are the specific values in a sample, while $\bar{x}$ and $\bar{y}$ are the mean of each sample. Additionally, variable n is the number of values contained in a sample.

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right) \times \left(y_i - \bar{y}\right)\right] \tag{1}$$

$$SD(x) = \sqrt{\frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}} \tag{2}$$

$$r(x,y) = \frac{cov(x,y)}{SD(x) \times SD(y)} = \frac{\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right) \times \left(y_i - \bar{y}\right)\right]}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \times \sum\left(y_i - \bar{y}\right)^2}} \tag{3}$$

The vectors x and y represent the average eye sizes being analyzed and the expression values of a specific gene, respectively. n, on the other hand, corresponds to the number of strains being selected for this iteration of calculation. Furthermore, since correlation does not imply causation, a correlation analysis is usually followed by a statistical significance test using the p-value, which allows us to assess the level of significance of the correlation calculation [22].

2.3. Parallel Computation in R

Unlike C/C++ or Fortran, R has not been designed to run in parallel. Although its basic packages are highly optimized for sequential programs, it still requires a long time to process huge datasets. Therefore, several packages have been developed, and made available on CRAN, to facilitate parallel processing of R code. The doParallel package, for example, uses a data parallelization scheme to spread computational tasks onto multiple processing cores of a non-clustered architecture with a shared-memory paradigm. doParallel is not a stand-alone package but operates jointly with three other packages: parallel, foreach, and iterators. The parallel package provides the basis for process creation and termination on both UNIX-based and Windows operating systems. In a similar fashion to the parallel package for openMP, the foreach package is used to distribute the data onto multiple CPU cores and consolidate the results from each process by applying an appropriate function [23].

As shown in Figure 1, the doParallel functions makeCluster, registerDoParallel, and stopCluster are used to manage the parallel region creation and termination. The %dopar% enables the loop to be parallelized through foreach, where each CPU core executes a chunk of the loop's iteration space. In fact, foreach and doParallel packages can be executed in both serial and parallel modes by toggling the .inorder parameter, which should be TRUE for serials and FALSE for parallels.

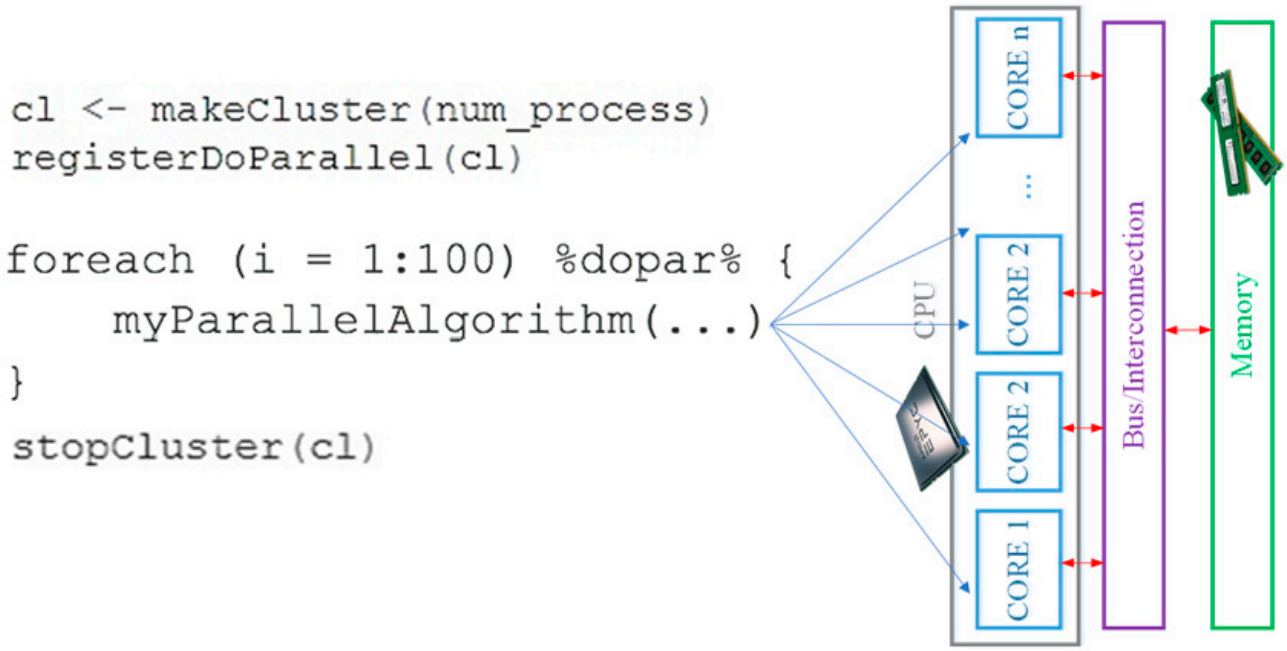

Figure 1. The data parallelization scheme of the R doParallel package on a shared-memory architecture.

2.4. The Computational Approach

The algorithms mapping phenotypes to associated RP genotypes using eye-size and DGRP gene expression data were originally proposed by Nguyen et al. [18]. In fact, a huge barrier towards extending their analysis was the intensive computational requirements of the algorithms. Hence, the main contribution of this research is to use a parallel computing approach to improve performance and reduce the overall execution time.

The main steps of the computations are summarized in Algorithm 1. In addition to the two datasets, it receives bottom and top quantile values as inputs. Those values, in turn, determine the strains representing the largest and the smallest eye-size groups. In fact, small eye sizes reflect the phenotype related to the RP genes while healthy eyes are characterized by large sizes. The quantile values can be adjusted so that the number of strains selected from each extreme eye-size group are balanced.

Since the input data come from two different sources (eye-size and expression level datasets), we will focus on those strain lines that are shared between the two datasets in each case. That is, non-intersecting strain lines will be discarded by the pre-processing step prior to applying the steps of the main algorithm. In addition, after a careful inspection of the expression level dataset, we found that some lines are represented with only one replicate. Those were filtered out as well and were not considered in the subsequent steps.

Once the extreme strains are identified and selected, Algorithm 2 is executed to generate the matrix of all possible replicate combinations for the selected lines. Since each strain has two replicates, the algorithm encodes each replicate combination with a binary number and decides which replicate to choose based on the values of their corresponding binary digit. Based on the number of selected extreme strains (N), the algorithm encodes the numbers from 0 to 2^{N-1}. For example, in the case of 6 selected strains, the total number of possible replicate combinations is 64 (2^6). The first generated number will be 000000, which selects the first replicate of every strain. Next, the sequence 100000 selects the second replicate of the first strain and the first replicate of the other five. The following numbers will be 010000, 110000, 001000, 101000, ... etc. The last combination will be represented by 111111, which means that the second replicate of each strain is selected.

It is clear that Algorithm 3 takes the longest time to run because the number of possible combinations grows exponentially with the number of selected lines, not to mention the huge size of the DGRP expression dataset. The sequential time complexity of the method

can be expressed using (4), where N refers to the number of selected lines representing extreme eye sizes and G represents the number of genes in the gene expression dataset. When dividing this workload across p processes, the parallel execution time can be expressed by (5). The impact of using parallel processing will be investigated and discussed more in the Results section.

$$T_s = 2^N \times G \tag{4}$$

$$T_p = \frac{T_s}{p} \tag{5}$$

Note that, beside the memory allocated for the two datasets, the space requirements of this algorithm include the matrix holding the replicate combinations and the vectors storing the list of candidate genes, as well as their correlation coefficients and their corresponding p-values. The space complexity of the proposed method can be estimated by (6), where L and G represent the number of lines and genes in the DGRP expression dataset, respectively.

$$Space\ Complexity = L \times G + L \times 2 + 2^N \times N + G \times 3 \tag{6}$$

Algorithm 1. Main Algorithm

Input:
Aes—Average eye sizes
Expr—Expression-level matrix
low_quantile—Bottom quantile of eye sizes
high_quantile—Top quantile of eye sizes
C—Correlation threshold value
num_process—Number of parallel processes
Output:
List of candidate genes
Begin
Filterout strains in *Expr* with only one replicate.
Filterout strains in *Expr* with no matching values in *Aes*.
selSizes ⟵ eye sizes in *Aes* less than *low_quantile* or greater than *high_quantile* values
extreme_strains ⟵ strains in *Aes* corresponding to *selSizes*
rep_combs ⟵ **Algorithm2** (*extreme_strains*)
best_rep_comb ⟵ **Algorithm3** (*selSizes*, *Expr*, *rep_combs*, *C*, *num_ process*)
candidate_genes ⟵ **Algorithm4** (*SelSizes*, *Expr*, *rep_combs*, *best_rep_comb*, *C*)
Print *candidate_genes*, their correlation coefficients, and *p*-values.
End

Algorithm 2. Generate replicate combinations

Input:
selStrain—Selected extreme strains.
Output:
*replicate*_comb—Replicate combinations matrix.
Begin
N ⟵ Length(*selStrain*)
for *i* ⟵ 1 **to** 2^N **do**
binary ⟵ DecimalToBinary(*i*)
for every binary digit *d* at position *j* **in** binary **do**
if *d* is 0 **then**
replicate_comb[i] ⟵ first replicate of strain *j*
else
replicate_comb[i] ⟵ second replicate of strain *j*
return *replicate_comb*
End

Algorithm 3. Find the best replicate combination

Input:
selSizes—Average eye sizes of selected strains
Expr—Expression level matrix
replicate_comb—Replicate combinations matrix
C—Correlation threshold value
num_process—Number of parallel processes
Output:
bestCombination—Best replicate combination
Begin
Create a parallel team with *num_process* processes.
Let *scoreVec* and *combVec* be empty vectors.
foreach replicate combination *c[i]* in *replicate_comb* **do**
for every gene *j* **do**
Score ⟵ 0
selExprs[j] ⟵ expression levels of *j* in the selected combination *c[i]*
temp[j] ⟵ correlation value between *selExprs[j]* and *selSizes*
if *temp[j]* < -C OR *temp[j]* > C **then**
Score ⟵ *Score* + 1
end
end
Append *Score* to *ScoreVec*
Append *c[i]* to *combVec*
end
Terminate the parallel session.
bestCombination ⟵ Replicate combination in *combVec* associated with Max(*ScoreVec*)
return *bestCombination*
End

Algorithm 4. Find candidate genes

Input:
selSizes—Average eye sizes of selected strains
Expr—Expression level matrix
best_rep_comb—Best replicate combination
C—Correlation threshold value
Output:
sorted_genes—List of candidate genes with their correlation coefficients and *p*-values
Begin
Let *m* be the number of genes in *Expr*
Let *Results* be a list of length *m*
foreach gene *j* **do**
Results[*j*] ⟵ correlation coefficient and *p*-values of *selSize* and gene expression levels of *j* for *best_rep_comb* in *Expr*
end
Filterout genes in *Results'* with *p*-values less than 0.05
Filterout genes in *Results'* with correlation coefficients not in the range [C, 1] nor [−1, -C]
sorted_genes ⟵ Sort *Results'* in descending order based on the absolute values of correlation *s* coefficients
return *sorted_genes*
End

Among all the generated combinations, Algorithm 3 tries to find the best combination based on their correlation with the average eye-size values. That is, for each replicate combination, the vector of expression levels of one gene is correlated with the vector of average eye-sizes for the selected extreme strains. When the absolute value of the calculated correlation coefficient exceeds the set threshold, it contributes to the total score of that replicate combination. Based on these accumulated scores, the best combination is determined

as the one achieving the highest score. Obviously, as the number of selected extreme lines increases, the number of replicate combinations grows exponentially. Therefore, when implementing Algorithm 3 we utilized the parallel computing constructs in R to cut down the execution time. More discussion will be given in the results section.

The final step of the analysis is handled by Algorithm 4, which calculates the Pearson's correlation coefficients between all genes in the best replicate combination and extreme eye sizes. The algorithm also computes the *p*-value in each case, and those correlations with *p*-values greater than or equal to 0.05 are filtered out of the list of candidate genes. The same is done with genes whose absolute correlation coefficient values don't exceed the set threshold value. Finally, Algorithm 1 should print out the sorted list of candidate genes along with their correlation coefficients and *p*-values.

3. Results

3.1. Experimental Setup

3.1.1. Datasets

The analysis here is primarily based on the DGRP expression level dataset, which consists of 368 columns representing 185 different *Drosophila* strain lines and 18,140 rows of genes. Each strain line has two replicates except for RAL513 and RAL890, which have one replicate each. Thus, when excluding these lines, the expression dataset has 6,675,520 expression values in total. On the other hand, the eye-size datasets of $Rh1^{G69D}$, *rpr*, and *p53* consist of 173, 202, and 204 strains, respectively. As discussed above, we will focus only on those strains that intersect with the DGRP expression dataset. Table 3 shows the detailed statistics of the three eye-size datasets. For example, $Rh1^{G69D}$ has 170 filtered strains with an arithmetic mean of 21,540.20 $\times$ 10^3 pixels.

Table 3. Statistics of filtered eye-size datasets.

Dataset	Filtered Strains	Mean	Median	Minimum	Maximum	Quartile	
						1st	3rd
$Rh1^{G69D}$	170	21,540.20	21,561.65	14,254.60	27,349.11	19,995.88	23,199.63
rpr	171	12,666.09	12,486.20	79,57.20	16,883.50	11,620.60	13,906.10
p53	172	14,244.73	14,166.00	10,542.20	17834.60	13,160.12	15,278.90

3.1.2. Quantile Thresholds

To investigate the effect of changing the number of selected lines representing the extreme eye sizes in the results, we need to find adequate quantile threshold values for each case. More specifically, we will investigate groups of 16, 18, 20, and 22 lines while trying to maintain a balanced number of strains representing extremely small and large eye sizes within a group. Table 4 lists the threshold values utilized for each group in the three datasets. Figure 2 depicts the cutoff lines for the 16- and 22-line groups using red and yellow colors, respectively, where the eye-size data is sorted in an ascending order. In the case of the 16-line group, for example, the selected strains are those that appear on the left and the right of the left and the right red cutoff lines, respectively.

Table 4. Quantile threshold values for different groups selected lines.

Dataset	Number of Selected Lines	Quantile (%)	
		Bottom	Top
$Rh1^{G69D}$	16	20.90	87.20
	18	21.00	87.00
	20	21.50	85.30
	22	22.10	84.60
rpr	16	21.10	83.80
	18	23.80	80.80
	20	25.10	80.40
	22	25.30	79.90
p53	16	17.80	83.68
	18	18.90	83.60
	20	19.10	82.20
	22	19.20	81.68

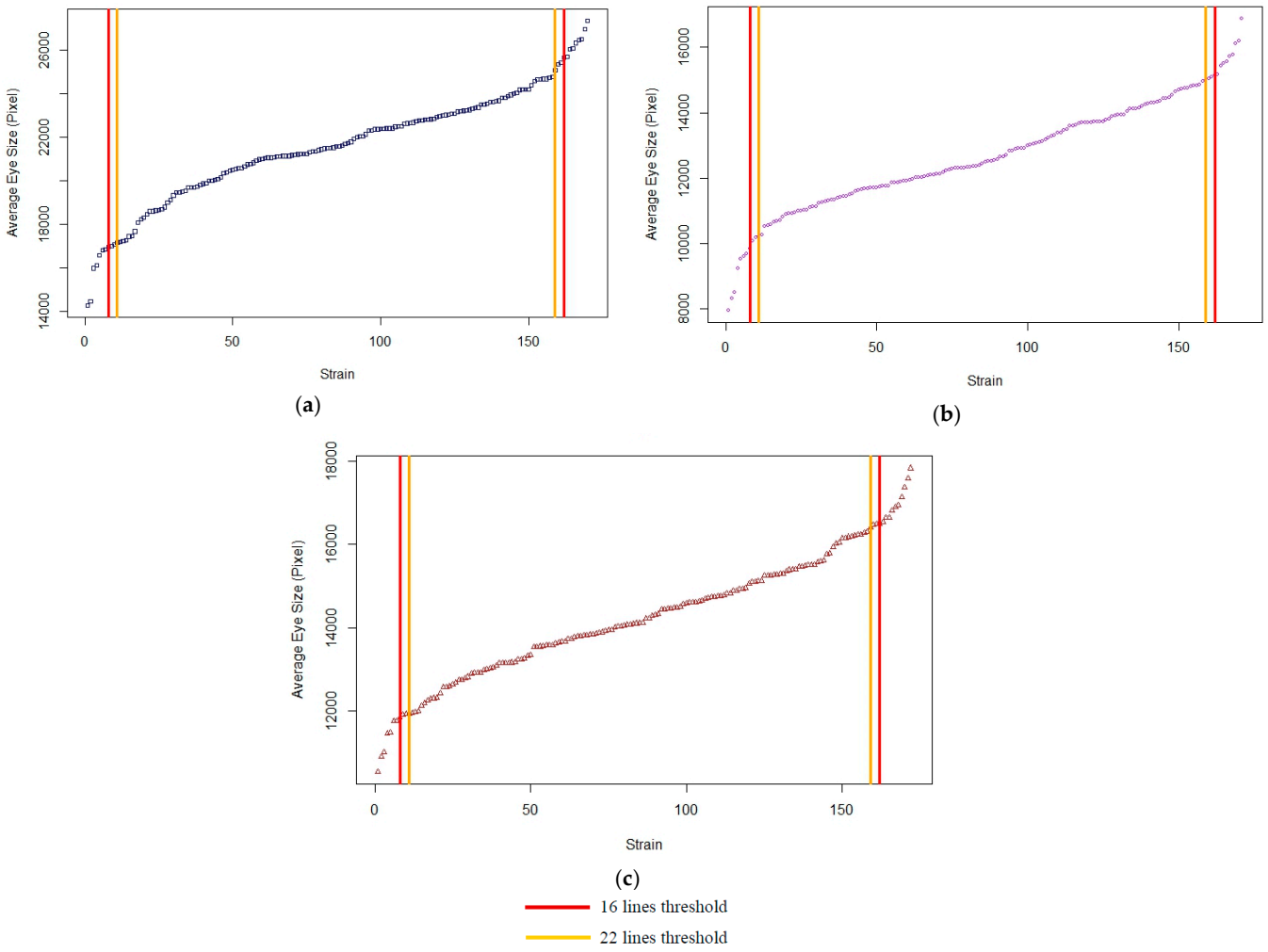

Figure 2. The cut-off thresholds of the average eye sizes for the groups of 16 and 22 extreme strains from the (**a**) *Rh1*^{G69D}, (**b**) *rpr*, and (**c**) *p53* datasets.

3.1.3. Hardware & Software Specifications

To implement the set of proposed algorithms, the R scripting language version 4.2.1 was used. In addition, several external libraries were utilized to analyze the data and support parallel computations. For example, cor() and cor.test() functions from the stats package were used to perform the correlation analysis step. The doParallel (ver. 1.0.17) package was used to facilitate parallelism on loops through the foreach (ver. 1.5.2) package. The original version of the code, initially developed by Nguyen et al. [18], used the highly optimized sapply and lapply sequential functions from the R base package.

The experiments were initially deployed on a laptop, then on the Diamond server of the department of computer science, Purdue University Fort Wayne. The laptop runs 64-bit Windows 8.1 on an Intel i7-6700HQ (sixth generation) CPU with a base clock speed of 2.6 GHz and 8 GB of RAM. The Diamond server runs the Oracle Linux operating system version 8.7 on AMD EPYC 7452 CPU with a base clock speed of approximately 2.3 GHz and 62 GB of RAM. While the laptop has four cores with eight physical threads, the Diamond server has thirty-two cores and can manage up to sixty-four threads.

3.2. Execution Time Analysis

In this set of experiments, the execution time of each algorithm was tested on the laptop for sixteen strain lines of the Rh^{1G69D} eye-size dataset. Nguyen's original code was compared with the enhanced parallel implementation. Table 5 shows the measured execution time for each algorithm separately. The results show that there is a clear overhead when executing Algorithm 2 in parallel, and recommend the used of the sequential sapply function in the following experiments. Similarly, we recommend running Algorithm 4 for eight processes, since the increase in speed achieved when moving from four to eight processes was not significant and further increases in the number of processes may cause overhead. Algorithm 3, on the other hand, showed an outstanding speedup since it was originally the most time consuming one. In fact, the number of possible replicate combinations increases drastically with increasing the number of selected lines. For example, using 16 lines in this case, 65,536 (2^{16}) different correlation calculations are needed for each gene. With 22 lines, the number of possible combinations reaches 2^{22} or 4,194,304.

Table 5. The execution time of each algorithm (in seconds) on a laptop for 16 strains of the Rh^{1G69D} eye-size dataset.

	Nguyen's	**Ours**			
		Serial	**2 Processes**	**4 Processes**	**8 Processes**
Algorithm 1	3.763	21.210	25.710	23.130	24.810
Algorithm 2	862.640	901.010	322.110	215.600	166.960
Algorithm 3	71.342	76.362	23.960	17.868	17.562

Using the Diamond server, we were able to repeat the same experiment using the recommended parameters for Algorithm 2 and Algorithm 4 while increasing the number of processes running Algorithm 3. Figure 3 depicts the average execution time for ten runs using sixteen lines from the Rh^{1G69D} eye-size dataset. The results clearly show the outstanding performance of the proposed parallel solution compared to Nguyen's version, even when only two threads were used. Furthermore, the results show that the best performance was achieved with 32 processes, with a roughly 95% decrease in execution time when compared to the sequential version. Using more than 32 processes showed a slight overhead that affected the execution time. To further investigate the optimal setting for Algorithm 3, we repeated the same set of experiments considering more lines in the extreme eye-size groups. The results shown in Figure 4 confirm that the execution time when using 32 processes is consistently lower than other settings regardless of the number of selected lines.

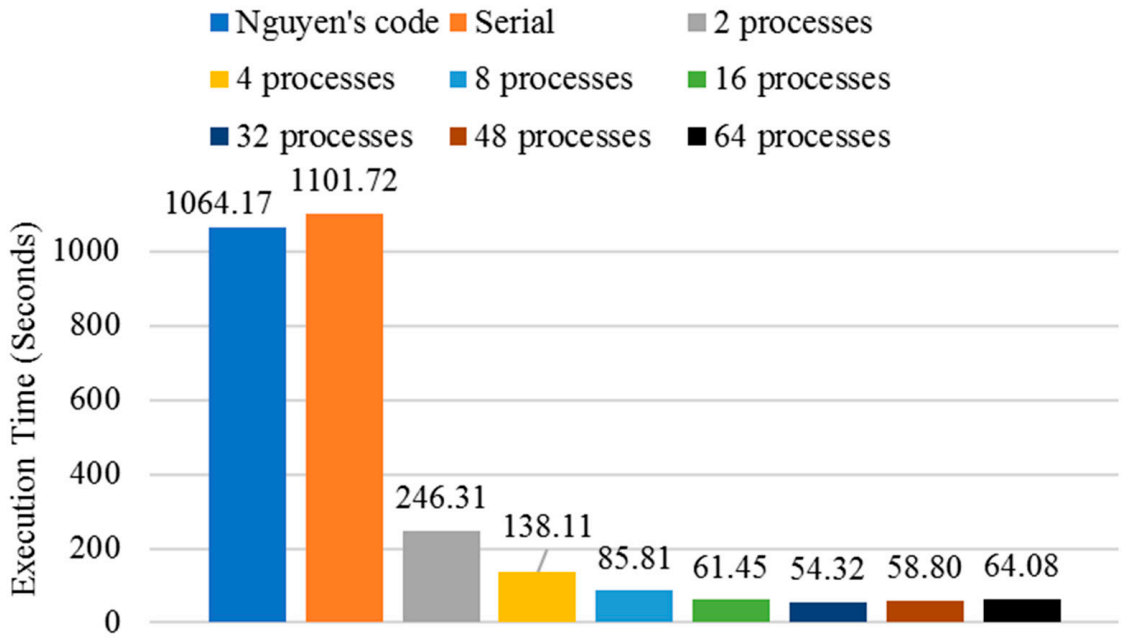

Figure 3. The sequential vs. parallel program execution times on the Diamond server with varying numbers of processes for 16 strains of the Rh^{1G69D} eye-size dataset.

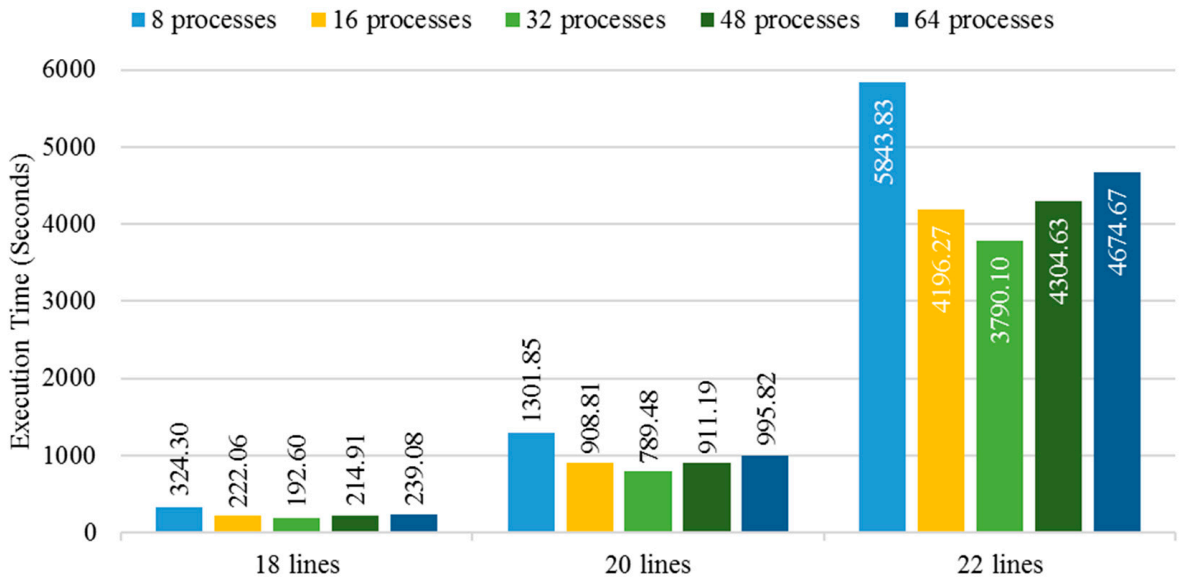

Figure 4. The parallel program execution time on the Diamond server with varying numbers of processes for different strain groups of the Rh^{1G69D} eye-size dataset.

3.3. Suspected Candidate Genes

Here, we are going to focus only on the top ten candidate modifiers that exhibit the highest correlation values with the extreme eye sizes for each of the three eye-size datasets. The complete list of genes will be provided as supplementary material (Tables S1–S12). The top candidate RP genes for the Rh^{1G69D} dataset using 16, 18, 20, and 22 strain lines are listed in Table 6. Tables 7 and 8 show the top genes for the *rpr* and *p53* datasets, respectively. The suspect genes that carry over at least three groups of selected lines are highlighted in blue. Those are potentially significant and worth further investigation. For example, the gene FBgn0032847 appeared in all four groups of the Rh^{1G69D} dataset and has the highest correlation values in both the 20 and 22 groups. The same is true for the genes FBgn0027601 and FBgn0053017 in the *rpr* dataset and FBgn0004373 in the *p53* dataset. Furthermore, the

gene FBgn0037770 was found to be a common suspect gene between $Rh1^{G69D}$ and *p53*, and FBgn0015024 existed in both lists of *rpr* and *p53*.

Table 6. The top 10 candidate modifiers for the $Rh1^{G69D}$ eye-size dataset with various selected lines.

No. of Selected Lines	Gene ID	Correlation Coefficient	*p*-Value
16	FBgn0027378	−0.863661390	$1.624926332 \times 10^{-5}$
	FBgn0263005	−0.856324086	$2.297889577 \times 10^{-5}$
	XLOC_006268	−0.849992594	$3.053251949 \times 10^{-5}$
	FBgn0032847	−0.840543849	$4.560179069 \times 10^{-5}$
	FBgn0033782	0.840068819	$4.649931586 \times 10^{-5}$
	FBgn0036761	−0.834562369	$5.802852891 \text{ x } 10^{-5}$
	FBgn0037531	−0.828535797	$7.329098346 \times 10^{-5}$
	FBgn0086679	−0.826073369	$8.042371477 \times 10^{-5}$
	FBgn0031233	−0.814773350	$1.210237900 \times 10^{-4}$
	FBgn0038639	−0.810804945	$1.388117979 \times 10^{-4}$
18	FBgn0085376	0.859283205	$4.923670750 \times 10^{-6}$
	FBgn0033782	0.854529969	$6.322511262 \times 10^{-6}$
	FBgn0037531	−0.850683118	$7.692063225 \times 10^{-6}$
	FBgn0003345	0.849498533	$8.161927414 \times 10^{-6}$
	FBgn0036299	−0.824116242	$2.609375328 \times 10^{-5}$
	FBgn0086679	−0.820352399	$3.052250605 \times 10^{-5}$
	FBgn0032847	−0.809207602	$4.758311930 \times 10^{-5}$
	FBgn0037016	0.804435051	$5.705515305 \times 10^{-5}$
	FBgn0263005	−0.799150447	$6.936945025 \times 10^{-5}$
	FBgn0263659	0.798357906	$7.139778718 \times 10^{-5}$
20	FBgn0032847	−0.835162086	$4.614128240 \times 10^{-6}$
	FBgn0086679	−0.810201591	$1.489680678 \times 10^{-5}$
	XLOC_004892	0.808596485	$1.596926917 \times 10^{-5}$
	FBgn0037531	−0.798402859	$2.447648614 \times 10^{-5}$
	FBgn0037770	−0.795138465	$2.792299682 \times 10^{-5}$
	FBgn0033087	−0.782129177	$4.616635661 \times 10^{-5}$
	FBgn0038039	−0.777530252	$5.470964221 \times 10^{-5}$
	FBgn0261703	0.777148268	$5.547684002 \times 10^{-5}$
	FBgn0263602	−0.776940029	$5.589897080 \times 10^{-5}$
	FBgn0023513	−0.77665287	$5.648562369 \times 10^{-5}$
22	FBgn0032847	−0.819701175	$3.041416193 \times 10^{-6}$
	FBgn0003345	0.806371734	$5.849733373 \times 10^{-6}$
	FBgn0039125	−0.798479606	$8.420737122 \times 10^{-6}$
	FBgn0027378	−0.798116202	$8.559914937 \times 10^{-6}$
	FBgn0037770	−0.775234661	$2.259984672 \times 10^{-5}$
	FBgn0036299	−0.773718232	$2.400671440 \times 10^{-5}$
	FBgn0038039	−0.766653185	$3.162045710 \times 10^{-5}$
	FBgn0030817	−0.759197474	$4.186542502 \times 10^{-5}$
	FBgn0033087	−0.758610329	$4.278306102 \times 10^{-5}$
	FBgn0263602	−0.757552951	$4.447992560 \times 10^{-5}$

Table 7. The top 10 candidate modifiers for the *rpr* eye-size dataset with various selected lines.

No. of Selected Lines	Gene ID	Correlation Coefficient	*p*-Value
16	FBgn0053017	0.878155178	$7.701469269 \times 10^{-6}$
	FBgn0032225	0.864824355	$1.535297878 \times 10^{-5}$
	FBgn0033244	−0.863862383	$1.609129485 \times 10^{-5}$
	FBgn0015513	−0.863578519	$1.631477307 \times 10^{-5}$
	FBgn0027601	−0.863380084	$1.647253890 \times 10^{-5}$
	FBgn0004620	0.859876472	$1.947654735 \times 10^{-5}$
	XLOC_003703	0.857112086	$2.215964129 \times 10^{-5}$
	FBgn0052451	−0.854030715	$2.550919501 \times 10^{-5}$
	FBgn0030394	0.853756622	$2.582664724 \times 10^{-5}$
	FBgn0015024	−0.853296712	$2.636675031 \times 10^{-5}$
18	FBgn0030394	0.863043231	$4.013975720 \times 10^{-6}$
	FBgn0052451	−0.861824973	$4.291403231 \times 10^{-6}$
	FBgn0053017	0.857064474	$5.539380529 \times 10^{-6}$
	FBgn0015513	−0.854016815	$6.492125538 \times 10^{-6}$
	XLOC_001754	0.85284074	$6.895646685 \times 10^{-6}$
	FBgn0027601	−0.851538255	$7.367464694 \times 10^{-6}$
	FBgn0032225	0.848190573	$8.709105665 \times 10^{-6}$
	FBgn0040508	0.846060487	$9.667524618 \times 10^{-6}$
	FBgn0051523	−0.839435144	$1.324811250 \times 10^{-5}$
	XLOC_003703	0.838478248	$1.384887686 \times 10^{-5}$
20	FBgn0052451	−0.865408608	$8.358927794 \times 10^{-7}$
	FBgn0027601	−0.856270053	$1.457998848 \times 10^{-6}$
	XLOC_003703	0.846002475	$2.608233863 \times 10^{-6}$
	FBgn0037223	0.842028538	$3.230636630 \times 10^{-6}$
	FBgn0053017	0.841496599	$3.323055691 \times 10^{-6}$
	FBgn0033244	−0.834443888	$4.784948638 \times 10^{-6}$
	FBgn0051523	−0.834340071	$4.810090828 \times 10^{-6}$
	XLOC_004120	0.832821352	$5.191263979 \times 10^{-6}$
	XLOC_006378	0.832584454	$5.253030551 \times 10^{-6}$
	FBgn0039491	0.828453174	$6.437918905 \times 10^{-6}$
22	FBgn0027601	−0.848794066	$5.948318658 \times 10^{-7}$
	FBgn0053017	0.828428119	$1.924641707 \times 10^{-6}$
	FBgn0004620	0.819129408	$3.131318980 \times 10^{-6}$
	FBgn0015513	−0.816721873	$3.536085022 \times 10^{-6}$
	FBgn0036874	0.815971236	$3.671371942 \times 10^{-6}$
	FBgn0039491	0.815006786	$3.851872618 \times 10^{-6}$
	FBgn0033244	−0.812433525	$4.372265832 \times 10^{-6}$
	FBgn0036017	0.80726159	$5.608563439 \times 10^{-6}$
	FBgn0085692	0.806424546	$5.835169750 \times 10^{-6}$
	XLOC_003128	0.804285096	$6.451350215 \times 10^{-6}$

Table 8. The top 10 candidate modifiers for the *p53* eye-size dataset with various selected lines.

No. of Selected Lines	Gene ID	Correlation Coefficient	*p*-Value
16	FBgn0263110	0.914150185	$7.326821639 \times 10^{-7}$
	FBgn0030089	0.912212484	$8.520806952 \times 10^{-7}$
	FBgn0051804	−0.893134917	$3.203841392 \times 10^{-6}$
	XLOC_002940	−0.880664278	$6.703825055 \times 10^{-6}$
	FBgn0262148	−0.84019034	$4.626832053 \times 10^{-5}$
	FBgn0004373	0.824802374	$8.432604659 \times 10^{-5}$
	FBgn0029952	−0.824030001	$8.677364856 \times 10^{-5}$
	XLOC_006034	−0.816964908	$1.120425958 \times 10^{-4}$
	FBgn0034624	−0.813349043	$1.271757728 \times 10^{-4}$
	FBgn0026369	0.809052147	$1.473328296 \times 10^{-4}$
18	FBgn0030089	0.90872731	$1.812918682 \times 10^{-7}$
	FBgn0051804	−0.884756312	$1.083361127 \times 10^{-6}$
	XLOC_002940	−0.857875241	$5.307141864 \times 10^{-6}$
	FBgn0004373	0.823245148	$2.706658753 \times 10^{-5}$
	FBgn0263598	0.820901298	$2.983927843 \times 10^{-5}$
	FBgn0085478	0.813059623	$4.094743159 \times 10^{-5}$
	FBgn0005632	0.811169207	$4.409799798 \times 10^{-5}$
	XLOC_001981	−0.810258107	$4.568867256 \times 10^{-5}$
	FBgn0029952	−0.808005811	$4.983200012 \times 10^{-5}$
	FBgn0015024	0.800553196	$6.589932229 \times 10^{-5}$
20	XLOC_002940	−0.838696327	$3.848574574 \times 10^{-6}$
	FBgn0004373	0.830822887	$5.732735463 \times 10^{-6}$
	FBgn0050039	−0.814340058	$1.241477643 \times 10^{-5}$
	FBgn0259146	−0.810730528	$1.455735618 \times 10^{-5}$
	FBgn0005649	0.805892929	$1.792731166 \times 10^{-5}$
	FBgn0037770	0.799502932	$2.340117744 \times 10^{-5}$
	FBgn0027338	0.799283244	$2.361258995 \times 10^{-5}$
	FBgn0037327	0.792106903	$3.149179899 \times 10^{-5}$
	XLOC_003332	−0.790428741	$3.363187698 \times 10^{-5}$
	FBgn0085478	0.789154503	$3.533971171 \times 10^{-5}$
22	XLOC_002940	−0.857556268	$3.401980564 \times 10^{-7}$
	FBgn0030089	0.84421251	$7.858326004 \times 10^{-7}$
	FBgn0085478	0.802630148	$6.966511250 \times 10^{-6}$
	FBgn0029976	0.799946333	$7.878992278 \times 10^{-6}$
	FBgn0005632	0.799262002	$8.127819248 \times 10^{-6}$
	FBgn0004373	0.789236473	$1.265143068 \times 10^{-5}$
	FBgn0029952	−0.788090936	$1.328762771 \times 10^{-5}$
	FBgn0263598	0.786198336	$1.440028907 \times 10^{-5}$
	FBgn0021760	0.78617631	$1.441370482 \times 10^{-5}$
	XLOC_004713	−0.784049252	$1.576201243 \times 10^{-5}$

4. Discussion

We will focus our discussion here on top correlated RP genes that appeared at least three times in different line groups or across different eye-size datasets. Based on this criteria, 16 genes have been identified and are listed in Table 9. The table shows the actual gene names, symbols, and their human orthologs (the human genes that share the same functionality as the *Drosophila* genes). Some of these genes have potential connections to apoptosis/disease, which are briefly described in the last column.

Table 9. Suspected candidate modifiers of Retinitis pigmentosa shared between different groups and/or datasets.

Gene ID	Shared Datasets/Lines	Gene Symbol	Gene Name	Human Ortho.	Link to RP
FBgn0032847	$Rh1^{G69D}$: 16, 18, 20, 22	CG10756	*TBP-associated factor 13*	*TAF13*; *SUPT3H*	Unknown
FBgn0037531	$Rh1^{G69D}$: 16, 18, 20	CG10445	N/A	*TTF2*; *HLTF*	Unknown
FBgn0086679	$Rh1^{G69D}$: 16, 18, 20	CG9770	*pink*	*HPS5*; *TECPR2*	Eye expression and primary function.
FBgn0027601	*rpr*: 16, 18, 20, 22	CG9009	*pudgy*	*ACSF2/ACSF3*	Fatty acid metabolism influences mitochondrial function and cell death.
FBgn0053017	*rpr*: 16, 18, 20, 22	CG33017	N/A	*GPATCH8*	Unknown
FBgn0052451	*rpr*: 16, 18, 20	CG32451	*secretory pathway calcium atpase*	*ATP2C1/ATP2C2*	Calcium influx can be a trigger for apoptosis. Loss in humans is associated with various diseases, including some atrophy/degeneration.
XLOC_003703	*rpr*: 16, 18, 20	N/A	N/A	N/A	Unknown
FBgn0015513	*rpr*: 16, 18, 22	CG10379	*myoblast city*	*DOCK1/DOCK2/ DOCK5*	Associated in (DOCK2) with immunodeficiency 40 (OMIM 616433). More distant orthologue (DOCK3) associated with neurodevelopmental disorder with autophagy and degenerative axons.
FBgn0033244	*rpr*: 16, 20, 22	CG8726	N/A	*PXK*	Loss in humans associated with susceptibility to lupus.
FBgn0004373	*p53*: 16, 18, 20, 22	CG7004	*four-wheel drive*	*PI4KB*	Connection to deafness and to insulin signaling in human/rodents.
XLOC_002940	*p53*: 16, 18, 20, 22	N/A	N/A	N/A	Unknown
FBgn0030089	*p53*: 16, 18, 22	CG9113	*adaptor protein complex 1, gamma subunit*	*AP1G1/AP1G2*	Associated with USRISR, a neurodevelopmental disorder (AP1G1) (OMIM 619548).
FBgn0029952	*p53*: 16, 18, 22	CG12689	N/A	N/A	Unknown
FBgn0085478	*p53*: 18, 20, 22	CG34449	*zinc finger DHHC-type containing 8*	*ZDHHC5/ ZDHHC8*	Linked to learning and memory (neuronal function) in mouse models.
FBgn0037770	$Rh1^{G69D}$: 20, 22 *p53*: 20	CG5358	*arginine methyltransferase 4*	*CARM1*; *METTL27/7B/7A*; *PRMT9/3/7/6/8/2/1*; *NDUFAF5*; *ALKBH8*; *BUD23*; *ATPSCKMT*; *GSTCD*; *TRMT9B*; *ANTKMT*	Unknown
FBgn0015024	*rpr*: 16 *p53*: 18	CG2028	*casein kinase Iα*	*Hsap\CSNK1A1, Hsap\CSNK1A1L*	a biomarker for Alzheimer's Disease

This list of candidate modifiers is completely distinct from the results discussed in both Chow's [4] and Amstutz's [17] due to the different approaches implemented for the DGE analysis. However, when compared with Nguyen's top ten candidate genes [18], two intersecting genes were identified: *MORF-related gene 15* (*MRG15*) and *p* (*pink*). However, the relationship of these candidate genes with the degenerative retinal disease cannot be validated until further in-depth biological lab experiments.

5. Conclusions

In this study, we have improved the algorithm for finding candidate RP genes originally proposed by Nguyen [18]. The method starts with identifying extreme eye-size strains and generating an exhaustive list of their possible replicate combinations. The best replicate combination is then identified as the one maximizing the correlation between

the eye-size phenotype and the expression profiles of all genes. With 2^N possible replicate combinations considered for each gene in the dataset, the sequential implementation of the method was found to be time consuming, which limited the number of lines (N) that could be considered and hence the scope of the analysis.

Therefore, we have implemented a parallel version of the method using the foreach and the doParallel R packages. This enabled the program to run faster than the original version and decreased the execution time by up to 95% when using 32 processes. Moreover, beside the DGRP gene expression dataset, three eye-size datasets, $Rh1^{G69D}$, *rpr*, and *p53*, were also considered. Several experiments have been conducted on groups of 16, 18, 20, and 22 extreme lines. In our analysis, we focused on the top ten candidate genes, giving a particular importance to those genes which emerged over multiple groups of selected lines and/or datasets. As a result, sixteen candidate genes were identified and need to be further validated through a follow up biological study to prove their association with the RP disease.

Future directions for this work could be extending the focus of analysis beyond the top ten candidate genes. In addition, given access to more powerful computational resources, the experiments could include groups with more than 22 extreme lines. Another extension could be applying a standard DGE analysis tool, such as DESeq2, to the expression data for the best replicate combination. One can also consider rewriting the code using other languages such as C/C++ and use GPUs for a more efficient parallel implementation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/computation11060118/s1, Table S1: The list of all valid genes for 16-lines of Rh1G69D, Table S2: The list of all valid genes for 18-lines of Rh1G69D, Table S3: The list of all valid genes for 20-lines of Rh1G69D, Table S4: The list of all valid genes for 22-lines of Rh1G69D, Table S5: The list of all valid genes for 16-lines of rpr, Table S6: The list of all valid genes for 18-lines of rpr, Table S7: The list of all valid genes for 20-lines of rpr, Table S8: The list of all valid genes for 22-lines of rpr, Table S9: The list of all valid genes for 16-lines of p53, Table S10: The list of all valid genes for 18-lines of p53, Table S11: The list of all valid genes for 20-lines of p53, Table S12: The list of all valid genes for 22-lines of p53.

Author Contributions: Conceptualization, A.K.; Data curation, C.M.; Formal analysis, R.P.; Methodology, A.K.; Project administration, A.K.; Resources, R.P.; Software, C.M.; Visualization, C.M.; Writing—original draft, A.K.; Writing—review & editing, C.M. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Genomic sequence and gene expression data for the DGRP is available at http://dgrp.gnets.ncsu.edu/ (accessed on 4 June 2023). Gene expression data was initially published in Huang et al. 2015. Eye size data is available as a supplementary file from Show et al. 2016. Code can be provided upon request.

Acknowledgments: This research was part of a Master thesis research in Computer Science at Purdue University Fort Wayne. We would like to thank the department for facilitating the access to the Diamond Server to run the parallel experiments during the period January to May 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hartong, D.T.; Berson, E.L.; Dryja, T.P. Retinitis pigmentosa. *Lancet* **2006**, *368*, 1795–1809. [CrossRef] [PubMed]
2. Huang, W.; Massouras, A.; Inoue, Y.; Peiffer, J.; Ràmia, M.; Tarone, A.M.; Turlapati, L.; Zichner, T.; Zhu, D.; Lyman, R.F.; et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res.* **2014**, *24*, 1193–1208. [CrossRef] [PubMed]
3. Mackay, T.F.C.; Richards, S.; Stone, E.A.; Barbadilla, A.; Ayroles, J.F.; Zhu, D.; Casillas, S.; Han, Y.; Magwire, M.M.; Cridland, J.M.; et al. The Drosophila melanogaster Genetic Reference Panel. *Nature* **2012**, *482*, 173–178. [CrossRef] [PubMed]
4. Chow, C.Y.; Kelsey, K.J.P.; Wolfner, M.F.; Clark, A.G. Candidate genetic modifiers of retinitis pigmentosa identified by exploiting natural variation in Drosophila. *Hum. Mol. Genet.* **2016**, *25*, 651–659. [CrossRef] [PubMed]

5. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szcześniak, M.W.; Gaffney, D.J.; Elo, L.L.; Zhang, X.; et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **2016**, *17*, 13. [CrossRef]
6. Karademir, D.; Todorova, V.; Ebner, L.J.A.; Samardzija, M.; Grimm, C. Single-cell RNA sequencing of the retina in a model of retinitis pigmentosa reveals early responses to degeneration in rods and cones. *BMC Biol.* **2022**, *20*, 86. [CrossRef] [PubMed]
7. Li, J.; Du, W.; Xu, N.; Tao, T.; Tang, X.; Huang, L. RNA-Seq Analysis for Exploring the Pathogenesis of Retinitis Pigmentosa in P23H Knock-In Mice. *Ophthalmic Res.* **2021**, *64*, 798–810. [CrossRef]
8. Robinson, M.D.; Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **2007**, *23*, 2881–2887. [CrossRef]
9. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef]
10. Hardcastle, T.J.; Kelly, K.A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform.* **2010**, *11*, 422. [CrossRef]
11. Leng, N.; Dawson, J.A.; Thomson, J.A.; Ruotti, V.; Rissman, A.I.; Smits, B.M.G.; Haag, J.D.; Gould, M.N.; Stewart, R.M.; Kendziorski, C. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **2013**, *29*, 1035–1043. [CrossRef] [PubMed]
12. Tarazona, S.; García-Alcalde, F.; Dopazo, J.; Ferrer, A.; Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Res.* **2011**, *21*, 2213–2223. [CrossRef] [PubMed]
13. Li, J.; Tibshirani, R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **2013**, *22*, 519–536. [CrossRef] [PubMed]
14. Soneson, C.; Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **2013**, *14*, 91. [CrossRef] [PubMed]
15. Liang, P.; Pardee, A.B. Analysing differential gene expression in cancer. *Nat. Rev. Cancer* **2003**, *3*, 869–876. [CrossRef] [PubMed]
16. Rodriguez-Esteban, R.; Jiang, X. Differential gene expression in disease: A comparison between high-throughput studies and the literature. *BMC Med. Genom.* **2017**, *10*, 59. [CrossRef]
17. Amstutz, J.; Khalifa, A.; Palu, R.; Jahan, K. Cluster-Based Analysis of Retinitis Pigmentosa Modifiers Using Drosophila Eye Size and Gene Expression Data. *Genes* **2022**, *13*, 386. [CrossRef]
18. Nguyen, T.; Khalifa, A.; Palu, R. Identifying Genes Related to Retinitis Pigmentosa in Drosophila melanogaster Using Eye Size and Gene Expression Data. *BioMedInformatics* **2022**, *2*, 625–636. [CrossRef]
19. Huang, W.; Carbone, M.A.; Magwire, M.M.; Peiffer, J.A.; Lyman, R.F.; Stone, E.A.; Anholt, R.R.H.; Mackay, T.F.C. Genetic basis of transcriptome diversity in Drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E6010–E6019. [CrossRef]
20. Palu, R.A.S.; Ong, E.; Stevens, K.; Chung, S.; Owings, K.G.; Goodman, A.G.; Chow, C.Y. Natural Genetic Variation Screen in Drosophila Identifies Wnt Signaling, Mitochondrial Metabolism, and Redox Homeostasis Genes as Modifiers of Apoptosis. *G3 Genes Genomes Genet.* **2019**, *9*, 3995–4005. [CrossRef]
21. Posnien, N.; Hopfen, C.; Hilbrant, M.; Ramos-Womack, M.; Murat, S.; Schönauer, A.; Herbert, S.L.; Nunes, M.D.S.; Arif, S.; Breuker, C.J.; et al. Evolution of eye morphology and rhodopsin expression in the Drosophila melanogaster species subgroup. *PLoS ONE* **2012**, *7*, e37346. [CrossRef]
22. Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur. J. Epidemiol.* **2016**, *31*, 337–350. [CrossRef] [PubMed]
23. Weston, S.; Calaway, R. Getting Started with doParallel and Foreach. 2022. Available online: https://CRAN.R-project.org/package=doParallel (accessed on 6 January 2023).