A Computational Pipeline for Functional Gene Discovery

Aolani Colon

Rutgers, The State University of New Jersey

Rishabh Hirday

Rutgers, The State University of New Jersey

Ami Patel

Rutgers, The State University of New Jersey

Amrita Poddar

Rutgers, The State University of New Jersey

Emma Tuberty-Vaughan

Rutgers, The State University of New Jersey

Tianyue Fu

Rutgers, The State University of New Jersey

Xin Ai

Rutgers, The State University of New Jersey

Wei Vivian Li

Rutgers, The State University of New Jersey

Li Cai (✉ lcail@rutgers.edu)

Rutgers, The State University of New Jersey

Research Article

Keywords: Computational pipeline, functional gene discovery, RNA-seq, retina development, photoreceptor cell, outer segment, biological process.

Posted Date: September 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-770297/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2021. See the published version at <https://doi.org/10.1038/s41598-021-03041-0>.

A computational pipeline for functional gene discovery

Aolani Colon^{1,3}, Rishabh Hirday^{1,3}, Ami Patel^{1,3}, Amrita Poddar^{1,3}, Emma Tuberty-Vaughan^{1,3}, Tianyue Fu¹, Xin Ai¹, Wei Vivian Li², and Li Cai¹

¹Department of Biomedical Engineering, Rutgers University, 599 Taylor Road, Piscataway, NJ 08854

²Department of Biostatistics and Epidemiology, Rutgers University, 683 Hoes Lane West, Piscataway, NJ 08854

³These authors contributed to this work equally.

Corresponding author:

Li Cai, Ph.D.

Department of Biomedical Engineering

Rutgers University

599 Taylor Road

Piscataway, NJ 08854

lcai@rutgers.edu

848-445-6559

ABSTRACT

Many computational pipelines exist for the detection of differentially expressed genes. However, computational pipelines for functional gene detection are rarely exist. We developed a new computational pipeline for functional gene identification from transcriptome profiling data. Key features of the pipeline include clustering optimization by gap statistics, gene ontology analysis for each cluster, and literature analysis for functional gene discovery. By leveraging this pipeline on RNA-seq datasets of mouse retinal development studies, we identified 14 candidate genes involved in the formation of the photoreceptor outer segment. The expression of top three candidate genes (*Pde8b*, *Laptm4b*, and *Nr1h4*) in the outer segment of the developing mouse retina were experimentally validated by immunohistochemical analysis. This computational pipeline can accurately predict novel functional gene for a specific biological process, e.g., the outer segment development of the photoreceptor cells in the mouse retina. This pipeline is also applicable to functional gene discovery for any other biological processes and in any other organs and tissues.

Keywords:

Computational pipeline; functional gene discovery; RNA-seq, retina development, photoreceptor cell, outer segment, biological process.

INTRODUCTION

Many computational pipelines exist for the detection of differentially expressed genes (DEGs), but not for functional gene detection. Major steps for DEG identification include: i) normalization, ii) dispersion estimates, and iii) differential gene expression based on condition. In order to carry out these functions, linux-based alignment tools, e.g., HISAT ¹ and STAR ², and R packages DESeq2 ³ and EdgeR ⁴ are most commonly used. Few pipelines, however, move data along for further analysis beyond DEGs, such as gene ontology analysis and novel gene discovery ⁵. We have developed a pipeline that covers the entire breadth of analysis for transcriptome profiling data, e.g., RNA-seq data, to facilitate functional gene discovery. Our pipeline (Figure 1) takes raw data in the form of fastq file format and includes modules to align sequencing reads/tags to a reference genome, normalize data based on condition, performs differential gene expression analysis, optimally clusters genes based on expression similarity, and carries out gene ontology enrichment analysis on each generated cluster. The identification of a new functional gene is based on 1) membership in gene cluster with established main biological function, e.g., photoreceptor outer segment (OS) development; and 2) the candidate gene does not have an established function with the main biological functions of the cluster. The functional gene discovery allows for the development of new hypotheses for the investigation the candidate genes. Key packages and methods utilized in this pipeline include Hisat2 ¹, featureCounts ⁶, DESeq2 ³, the gap statistic method ⁷, hierarchical/K-means clustering ^{1,3,8}, and clusterProfiler ⁹. To validate the pipeline, we attempted to identify functional genes from publicly available datasets of mouse retinal development studies at NCBI GEO. We have compiled two datasets to generate a dataset of gene expression profiling from mouse embryonic day 11 (E11) to postnatal day 28 (P28), which cover the full length of retinal development. Our pipeline accurately and effectively identified new functional genes for the development of the photoreceptor OS.

RESULTS

Sequence Alignment, Tag Counting, and Differential Gene Expression Analysis

We tested our computational pipeline using two previously published RNA-seq datasets (*Gene Expression Omnibus*, GEO accession # GSE74660 ¹⁰ and GSE101986 ¹¹ under NCBI Short-Read Archive (SRA) # SRX1411331 to SRX1411346 and SRX3044650 to SRX3044673). The datasets (single end read format) were originally generated from mouse developing retinas using Illumina's Genome Analyzer IIX. All data samples were obtained via fastq-dump from the SRA toolkit (<http://ncbi.github.io/sra-tools>). The combined dataset contains 12 time points with triplicate samples for each time point.

Each sample was then aligned to mouse genome (mm10) using HISAT2 ¹, which resulted in > 97% seq tag mapping rate. The resulting SAM files were converted into sorted bam files using samtools ¹². A count matrix of sequencing tags/reads with 33,487 genes ([Supplemental Table S1](#)) was generated using featureCounts ⁶. DESeq2 ³ was used to identify significantly differentially expressed genes (DEGs, p-value < 0.05) across the embryonic and postnatal development stage. Samples from E11 to E16 were considered as replicates for the embryonic stage, while samples from P0 to P28 were considered as replicates for the postnatal stage. The top ranked 3000 DEGs ([Supplemental Table S2](#)) were normalized using transcript per million (TPM) method ³ and selected as the most significant genes in retinal development for further analysis.

Clustering Analysis and Optimal Cluster Number Selection by Gap Statistics

We next performed clustering analysis on the top 3000 DEGs. A predominant clustering algorithm for temporal gene expression data is the hierarchical method ¹³. In this method, the number of clusters is determined by the distinct number of branches on the dendrogram produced. However, this approach is highly subjective and inconsistent, hence we instead used the Gap statistics ⁷ to determine the optimal number of clusters within the dataset. Gap statistics produced consistent results with varying input parameters and identified that the optimal number of gene clusters in this dataset was 33. Thus, the top 3000 DEGs were grouped into 33 distinct clusters with various members of similar expression pattern ([Supplemental Table S3](#), [Figure S1](#)).

Identification of Gene Clusters Involved in Photoreceptor Outer Segment Development

For each of the 33 gene clusters, gene ontology analysis was performed using clusterProfiler ⁹ to determine the enrichment of genes involved in a specific biological process, molecular function, and cellular component ([Supplemental Table S4](#)). We further analyzed the clusters with gene expression level in rising trend, e.g., clusters #2, 4, and 8 containing 120, 171, and 177 genes, respectively. The expression pattern of the three clusters parallels with the development of the photoreceptor outer segment (OS), which is known to begin around postnatal day 0 (P0) and peak between P10 and P14 ¹⁰. Thus, these three clusters were designated as OS gene clusters. Each gene from the three clusters were cross-referenced with a list of 126 manually curated known genes important for the development of the photoreceptor outer segment (OS) ¹⁴ ([Supplemental Table S5](#)). This resulted in 24 DEGs in clusters # 2, 4, 8 matched with the 126 known OS genes ([Figure 2A-C](#)).

We speculate that genes in these three clusters are involved in the OS development. Thus, each gene from these clusters was further analyzed for its function in the retina using PubMed literature searches. The genes that had no literature hits related to the retina were then predicted as OS gene candidates. This resulted in 5 genes from cluster #2 (*Kif9*, *Pde8b*, *Parcg*, *Lrrc46*, *Ndr3*), 6 genes from cluster #4 (*Laptm4b*, *Rtn4rl2*, *Tmem8*, *Slc17a7*, *Drd4*, *Loxl4*), and 3 genes from cluster #8 (*Tmem255b*, *Napb*, *Nr1h4*), a total number of 14 candidate genes.

Experimental Verification of Computational Identified Novel Genes

The top candidates, one from each of the three clusters (*Pde8b*, *Laptm4b*, and *Nr1h4* from clusters # 2, 4, 8, respectively), were further experimentally validated to confirm their expression in the OS of the developing mouse retina. Immunohistochemical analysis was performed on tissue sections of mouse retina at P14 using antibodies against *Pde8b*, *Laptm4b*, *Nr1h4*. Antibody against *Rhodopsin* was used as a positive control to stain the OS of rod photoreceptors. Developmental stage P14 was chosen for the two following reasons: 1) RNA-seq analysis reveals the expression of these three genes at the highest level at this stage, and 2) the mouse OS development reaches the peak level around P14¹⁰. Results show that signals from *Pde8b*, *Laptm4b* and *Nr1h4* antibody staining were detected in the OS and overlap with the signals from the known OS marker *Rhodopsin* (Figure 2D). These results confirm the expression of our computationally predicted genes in the retina photoreceptors and validate the accuracy and effectiveness of the computational pipeline for novel functional gene discovery.

DISCUSSION

In this study, we have developed an effective and accurate computational pipeline that utilizes transcriptome profiling data for the identification of novel functional genes involved in a specific biological process, e.g., the photoreceptor OS development. This pipeline integrates bioinformatic tools and literature analysis on the existing knowledge for an accurate identification of novel functional genes. Our computational pipeline can also be applied to identify novel functional genes in other biological processes and in other tissue/organs. The expression of the candidate genes in tissue samples was further experimentally validated, demonstrates the effectiveness and accuracy of the pipeline. However, the explicit roles of these genes in the biologic systems, e.g., the outer segment of the retina, are yet to be further experimentally investigated. Gain- and loss-of-function analysis, e.g., gene knockout or RNAi-mediated gene silencing experiments, should be conducted to

explore the specific function of these genes. A limitation to this pipeline is that it requires manual curation of information on each individual gene within the cluster and it can be time consuming. Future direction of this pipeline should improve the manual curation process of gene information by leveraging automated text mining and machine learning technologies¹⁵ to expedite the literature review for novel functional gene discovery.

METHODS

Development of the Computational Pipeline

The workflow of the computational pipeline is described in the following steps (Figure 1):

① Sequence alignment

For mapping next-generation sequencing data (e.g., fastq files) to a reference genome, we used a fast and sensitive alignment program, HISAT2¹. Reference genome with the annotation file were obtained from the NCBI and UCSC Genome Browser. The resulting SAM files were converted into sorted bam files using samtools¹². Sequence tag counts were performed using featureCounts⁶ to generate a tag count matrix.

② Detection of DEGs

DESeq2³ was used to identify DEGs in the transcriptome profiling data. The count matrix generated from the sequence alignment was used as the input for DESeq2. DESeq2 identifies DEGs by comparing an experimental condition to a control (e.g., embryonic time points for this study) to generate a list of the DEGs (Geneid) with associated values in baseMean, log2FoldChange, lfcSE, p-value, p-adj. In this study, we aim to identify functional genes for OS development, which is prominent during postnatal states. Thus, all postnatal time points were tested against embryonic time points to identify DEGs between these two developmental periods. The top 3000 DEGs identified through DESeq2 analysis were then further analyzed.

③ Normalization

For clustering analysis, the dataset containing the top 3000 DEGs was normalized by converting the count matrix to a transcript-per-million (TPM) matrix using an algorithm developed by Michael Love³. For robustness, the median TPM was used across all replicates to represent gene expression level at that time point. The TPM values were then transformed into z scores by subtracting the mean of the genes and dividing by the standard deviation.

④ Determination of the optimal number of gene clusters/expression patterns

To identify the optimal number of clusters or gene expression patterns within the dataset, we used gap statistics method ⁷. This technique uses the output of any clustering algorithm and compares the change in within-cluster dispersion with that expected under an appropriate reference null distribution ⁷. After testing the data in a varying number of clusters up to 100 possible clusters (kmax = 100), an optimal number of clusters can be determined. Hierarchical clustering was performed on the top normalized genes. The clusters were created using the Euclidean distance and the complete linkage method. These parameters were chosen as they resulted in more well-balanced clusters compared to the other methods. Temporal plots were further utilized to visualize the expression of the genes in each cluster.

⑤ Gene ontology analysis

Genes in each cluster were further analyzed to determine the enrichment on the biological process (BP), cellular component (CC), and molecular function (MF) via clusterProfiler ⁹. The enrichment of gene ontology (GO) terms of interest in each cluster were selected for novel functional gene discovery.

⑥ Predication of new functional genes

To predict a new functional gene, we performed literature search for each gene in those clusters involved in a particular term in the BP, CC, and MF of interest. The predication of a novel functional gene is based on 1) its membership in a cluster with established biological function, e.g., photoreceptor OS development, and 2) lack of documented function from literature search. For example, *Pde8b* is a member in cluster #2, while literature search using PubMed resulted in zero hits with the terms: “*Pde8b*” and “retina” or “photoreceptor”.

⑦ Experimental verification of computationally predicted new functional genes

Techniques including immunohistochemistry, qRT-PCR, and in situ hybridization are commonly used for the detection of gene expression. In this report, CD1 mouse retinas at P14 (n=3) were harvested and processed following a previously established protocol ¹⁶. Briefly, retinas were dissected from the animals and fixed with 4% (w/v) PFA for 1 hr. Following fixation, samples were washed three times with PBS for 10 min each, soaked in 15% sucrose overnight and then 30% (w/v) sucrose at 4 °C until the retinas sank, embedded in cryopreserving media (Tissue Tek OCT compound), and stored at -20 °C. Cryosectioning was performed with a Cryotome (Thermo Scientific, Waltham, MA). Sample sections (12 μm thickness) were air dried for 10 min and stored at -20 °C. Immunofluorescence staining was performed on sections as described previously ¹⁶ using antibodies against *Pde8b*, *Laptm4b*, *Nr1h4*, and *Rhodopsin*. Cell nuclei were counter stained with DAPI to highlight the retinal structure. Images of the immunostained sections were captured using a Zeiss Axio Imager M1

(AxioVision 4.8 software) and a Zeiss LSM 900 (Zeiss ZEN blue software).

Declarations

The authors declare no competing interests. Data and materials will become freely available upon acceptance for publication.

Funding

This work was supported by funding from the National Eye Institute of the National Institutes of Health under award number NIH EY031439-02

Authors' contributions

Conceptualization, L.C.; Methodology, A.C., R.H., A.P., A.P., E.T.V., F.U., X.A.; writing—original draft, L.C., A.C., R.H., A.P., A.P., E.T.V., F.U., X.A.; writing—review and editing, L.C., A.C., R.H., A.P., A.P., E.T.V., F.U., X.A., W.V.L; funding acquisition, L.C. The authors read and approved the final manuscript.

Acknowledgements

The authors thank the members of the Cai Laboratory for helpful suggestions on the project.

REFERENCES

- 1 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357-360 (2015).
- 2 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
- 3 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).
- 4 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
- 5 Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature communications* **10**, 1-11 (2019).
- 6 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
- 7 Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**, 411-423 (2001).
- 8 Müllner, D. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software* **53**, 1-18 (2013).
- 9 Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* **16**, 284-287 (2012).
- 10 Kim, J. W. *et al.* Recruitment of Rod Photoreceptors from Short-Wavelength-Sensitive Cones during the Evolution of Nocturnal Vision in Mammals. *Dev Cell* **37**, 520-532, doi:10.1016/j.devcel.2016.05.023 (2016).
- 11 Brooks, M. J. *et al.* Improved Retinal Organoid Differentiation by Modulating Signaling Pathways Revealed by Comparative Transcriptome Analyses with Development In Vivo. *Stem Cell Reports* **13**, 891-905, doi:10.1016/j.stemcr.2019.09.009 (2019).
- 12 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
- 13 Nascimento, M. *et al.* Independent Component Analysis (ICA) based-clustering of temporal RNA-seq data. *PloS one* **12**, e0181195 (2017).
- 14 Collin, G. B. *et al.* Mouse Models of Inherited Retinal Degeneration with Photoreceptor Cell Loss. *Cells* **9**, doi:10.3390/cells9040931 (2020).
- 15 Gulo, C. A., Rúbio, T. R., Tabassum, S. & Prado, S. G. in *2015 Imperial College Computing Student Workshop (ICCSW 2015)*. (Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik).
- 16 Li, Y. *et al.* Top2b is involved in the formation of outer segment and synapse during late-stage photoreceptor differentiation by controlling key genes of photoreceptor transcriptional regulatory network. *J Neurosci Res* **95**, 1951-1964, doi:10.1002/jnr.24037 (2017).

Figures and Figure Legends

Figure 1. Outline of computational pipeline for functional gene discovery.

Schematic diagram illustrates the steps in the computational pipeline for functional gene discovery in retinal outer segment development as an example: ① Sequence alignment, ② Detection of DEGs, ③ Normalization, ④ Determination of the optimal number of gene clusters/expression patterns, ⑤ Gene ontology analysis, ⑥ Predication of new functional genes, ⑦ Experimental verification of computationally predicted new functional genes.

Figure 2. Gene expression profile and gene enrichment associated with OS development.

Plots of gene expression level as Z-score over developmental stages and gene enrichment for cluster #2 (A), #4 (B) and #8 (C). Each line represents the expression profile of a gene. Gray lines represent known OS genes, while colored lines

represent predicted novel functional OS genes. **(D)** Representative photomicrographs of mouse retina sections at postnatal day 14 (P14) immunostained with antibodies against Pde8b, Laptm4b, Nr1h4, and known photoreceptor markers for rods (Rhodopsin). OS = outer segment, ONL = outer nuclear layer, OPL = outer plexiform layer, INL = inner nuclear layer, IPL = inner plexiform layer, and GCL = ganglion cell layer. Scale bar = 50 μ m.

SUPPLEMENTARY INFORMATION

Figure S1. Plots of 33 clusters from the top 3000 normalized DEGs. The x-axis represents the developmental stages from embryonic day 11 (E11) to postnatal day 28 (P28). The y-axis represents the z-score, a normalized level of gene expression.

Supplementary Tables

Table S1: Count Matrix associated with RNA seq Data
See attached file “Table S1”.

Table S2: Differential Expression Analysis Results
See attached file “Table S2”.

Cluster	Size
1	41
2	120
3	89
4	171
5	37
6	35
7	109
8	177
9	429
10	167
11	36
12	168
13	343
14	71
15	22
16	73
17	48
18	50
19	90
20	230
21	40
22	107
23	29
24	43
25	9
26	19
27	23
28	31
29	8
30	5
31	12
32	15
33	5

Table S3: Size of the optimized 33 clusters for the top 3000 genes.

Cluster	Term	Count	P-Value
2	Visual perception*	12	1.98E-12
	Sensory perception of light stimulus*	12	2.66E-12
	Retina homeostasis*	8	5.70E-11
	Photoreceptor cell maintenance*	7	5.10E-10
	Detection of light stimulus*	7	1.47E-09
3	Photoreceptor cell maintenance*	5	2.18E-07
	Retina homeostasis*	5	6.64E-07
	Regulation of non-motile cilium assembly	3	1.96E-05
	Neural retina development*	4	8.06E-05
	Retina morphogenesis in camera-type eye*	4	8.48E-05
4	Visual perception*	20	1.84E-21
	Sensory perception of light stimulus*	20	3.04E-21
	Detection of light stimulus*	9	1.26E-11
	Detection of visible light*	6	2.25E-08
	Detection of abiotic stimulus	9	5.96E-08
6	Lens development in camera-type eye*	7	3.66E-07
	Axon guidance	8	5.28E-05
	Neuron projection guidance	8	5.43E-05
	Regulation of cellular response to growth factor stimulus	8	1.07E-04
	Positive regulation of protein kinase activity	10	1.24E-04
7	Synaptic vesicle cycle	7	1.24E-05
	Positive regulation of synaptic vesicle recycling	3	1.59E-05
	Synaptic vesicle transport	4	2.37E-05
	Vesicle-mediated transport in synapse	7	2.65E-05
	Synaptic vesicle localization	4	5.93E-05
8	Visual perception*	7	1.11E-05
	Sensory perception of light stimulus*	7	1.31E-05
	Detection of visible light*	4	1.41E-05
	Organic hydroxy compound metabolic process	11	2.27E-05
	Negative regulation of interleukin-6 production	4	9.63E-05
9	Potassium ion transport	26	6.16E-16
	Regulation of membrane potential	34	1.19E-15
	Monovalent inorganic cation transport	34	4.18E-15
	Neurotransmitter transport	25	1.42E-12
	Drug transport	19	7.36E-11
10	Lens development in camera-type eye*	7	3.66E-07
	Axon guidance	8	5.28E-05
	Neuron projection guidance	8	5.43E-05
	Regulation of cellular response to growth factor stimulus	8	1.07E-04
	Positive regulation of protein kinase activity	10	1.24E-04
13	Substrate adhesion-dependent cell spreading	10	5.97E-07
	Nucleobase biosynthetic process	5	1.19E-06
	rRNA transcription	6	1.73E-06
	Nucleotide biosynthetic process	13	1.74E-06
	ncRNA processing	17	1.87E-06
16	Positive regulation of cell projection organization	7	8.73E-05
	Photoreceptor cell maintenance*	3	1.13E-04
20	DNA repair	23	2.43E-11
	Double-strand break repair	15	1.14E-09
	DNA recombination	15	2.98E-08
	Double-strand break repair via homologous recombination	10	1.89E-07
	Recombinational repair	10	1.89E-07
22	Embryonic organ development	20	9.37E-15
	Appendage morphogenesis	13	3.31E-13
	Limb morphogenesis	13	3.31E-13
	Embryonic limb morphogenesis	12	6.61E-13
	Embryonic appendage morphogenesis	12	6.61E-13
23	Anatomical structure maturation	5	5.22E-06
	Developmental maturation	5	1.27E-05
	Eye development*	5	3.73E-05
	Visual system development*	5	3.86E-05
	Sensory system development*	5	4.00E-05
26	Forebrain development	5	3.25E-06
	Central nervous system neuron differentiation	4	1.08E-05
	Negative regulation of neuron differentiation	4	1.91E-05
	Cerebral cortex GABAergic interneuron differentiation	2	2.14E-05
	GABAergic neuron differentiation	2	3.53E-05

Table S4: Gene ontology terms enriched in clusters of the top 3000 genes.

Table S5. List of genes associated with retinal outer segment development.
See attached file “Table S5”.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS2DiffExpData.csv](#)
- [TableS1RawCountMatrixncbiRefGene.txt](#)
- [TableS50Sgenes.xlsx](#)