Network-based bioinformatics analysis of spatio-temporal RNAseq data reveals transcriptional programs underpinning normal and aberrant retinal development

Devi Krishna Priya Karunakaran[1#], Sahar Al Seesi[2#], Abdul Rouf Banday[1], Marybeth Baumgartner[1], Anouk Olthof[1,3], Christopher Lemoine[1], Ion Mandoiu[2], Rahul N Kanadia[1*]

[1]Dept. of Physiology and Neurobiology, University of Connecticut, CT 06269, USA
[2]Dept. of Computer Science and Engineering, University of Connecticut, CT 06269, USA
[3]Utrecht University, 3508 TC Utrecht, The Netherlands
[#]These authors contributed equally to this work.
[*]Author to whom the correspondence should be addressed: Rahul.kanadia@uconn.edu

Abstract. Capturing transcriptome changes across developing tissue, through analysis of whole transcriptome sequencing analysis, enables us is to find patterns in gene expression across time with the expectation that it will reveal transition in biological processes. We developed a bioinformatics pipeline that analyzes samples in the comparison and categorizes genes in bins based on their gene differential expression and alternative splicing status. We integrated our binning results with downstream analysis that intelligently query gene annotation databases (DAVID and GeneMania) to discover molecular programs being employed during retinal development. We applied out pipeline to two retinal RNA-Seq dataset. Through our integrated analysis, we were able to leverage molecular signatures to predict potential phenotypes from gene expression changes long before they manifest and can be extended to other tissues and disease pathogenesis.

1 Introduction

The retina has been the most accessible part of the developing central nervous system with a wealth of information on detailed birth order of its cell types and on many genes involved in executing specific programs such as cell cycle regulation, cell fate determination, and neuronal differentiation. However, a comprehensive gene regulatory network is still not achieved as gene-centric approach can only go so far. Another concern is that at any given time the retina consists of different cell types with varied transcriptomes, which renders finding meaning from co-transcriptionally regulated genes difficult. We wanted to investigate whether higher depth of the captured transcriptome through RNAseq with minimal cross-compartment (nucleus-cytoplasm) normalization could resolve this issue.

Here we report analysis of RNAseq data from cytoplasmic and nuclear transcriptome of the developing retina. We show that combinatorial use of RNAseq with our custom bioinformatics strategy revealed the precise order of gene activation and transitions in processes during retinal development. Transition in gene expression was validated and resolved at the isoform level through our custom microarray. Importantly, we show proof of principle by extending our methodology to analyze RNAseq data from P21-Nrl-WT and KO retinae. Our approach which focuses on understanding the temporal progression in gene expression during normal/aberrant development can be extended to development and disease progression of other tissues.

2 Pipeline description

Samples are analyzed in pairs, and genes were classified based on their expression levels (expressed vs. not expressed), differential gene expression calls, and the number of expressed isoforms for the gene. Gene expression is estimated using IsoEM [6]. For gene differential expression, two methods are used, GFOLD [3] and Fisher's exact test with house-keeping gene normalization as suggested by Bullard et al.[2]. Gapdh was used as the housekeeping gene for this analysis. Genes are called differentially expressed if they showed ≥2 fold expression in one sample by both methods. Genes are categorized in bins based on their gene differential expression and the number of expressed isoforms. Genes belonging to each bin were analyzed for enrichment of individual GOterms using the Database for Annotation, Visualization and Integrated Discovery, DAVID [4][5]. Default parameters (≤0.05 Benjamini score) were used for all analyses. The gene lists enriching for GOterms were run through GeneMANIA to identify potential partners [7] .

3 Results

We applied the pipeline on fractionated RNA-Seq of normal retina at two different developmental time points, embryonic day (E) 16 and postnatal day (P) 0. Through our integrated analysis, we captured high-resolution transcription kinetics and by RNA deep sequencing of cytoplasmic and nuclear extracts of the developing retina, which revealed sets of co-transcriptionally regulated genes. Specifically, we found de novo transcription of genes whose transcripts were exclusively found in the nuclear transcriptome. Our downstream analysis, using DAVID and GeneMania showed that we used the co-transcriptionally regulated genes to identify the precise order of biological processes and the expression kinetics of the underlying genes across embryonic day (E) 16 and postnatal day (P) 0. Interestingly the genes exclusively found in the P0 nuclear transcriptome enriched for functions that are known to be executed during postnatal development. This finding showed that the P0 nuclear transcriptome is temporally ahead of that of its cytoplasm. We extended our strategy to perform static comparison between P21-Nrl-wildtype (WT) and P21-Nrl-knockout (KO) data [1]. While co-transcriptionally regulated genes were identified, we did not find enrichment for any biological processes. However, and showed that only through temporal analysis between data from P0 retina to either P21-Nrl-WT or P21-Nrl-KO showed, the molecular signatures predicting that the Nrl-KO retina would have problems with vasodilation was revealed. Indeed, histological manifestation of vasodilation has been reported at a later time point (P60). Thus, our transcriptome analysis platform can leverage molecular signatures to predict potential phenotypes from gene expression changes long before they manifest and can be extended to other tissues and disease pathogenesis.

References

[1] MJ . Brooks, HK. Rajasimha, JE. Roger, A. Swaroop, **Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl(-/-) retinal transcriptomes**. *Mol Vis* 17:3034-3054, 2011.

[2] JH. Bullard, E. Purdom, KD. Hansen, S. Dudoit, **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments,** *BMC Bioinformatics* 11:94, 2010

[3] J. Feng, C. A. Meyer, Q. Wang, J. S. Liu, X. S. Liu3, Y. Zhang, **GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data**, *Bioinformatics,* 28: 2782-2788, 2012.

[4] W. Huang da, BT. Sherman, RA. Lempicki RA. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Res* 37:1-13, 2009.

[5] W. Huang, BT. Sherman, RA. Lempicki, **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 4:44-57, 2009.

[6] M. Nicolae and S. Mangul and I.I. Mandoiu and A. Zelikovsky, **Estimation of alternative splicing isoform frequencies from RNA-Seq data**, *Algorithms for Molecular Biology* 6:9, 2011.

[7] D. Warde-Farley, SL. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, CT. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, GD. Bader, Q. Morris. **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic Acids Res* 38:W214-220, 2010.