

1 **Zika infection of neural progenitor cells perturbs transcription in**
2 **neurodevelopmental pathways.**

3 Lynn Yi^{1,2}, Harold Pimentel³, Lior Pachter^{4*}

4 1 Division of Biology and Biological Engineering, Caltech, Pasadena, CA, USA

5 2 David Geffen School of Medicine, University of California, Los Angeles, CA, 90095, USA

6 3 Department of Genetics, Stanford University, Palo Alto, CA, USA

7 4 Departments of Mathematics, Molecular & Cell Biology and Computer Science, University of
8 California at Berkeley, Berkeley, CA, USA

9

10 *Corresponding author: lpachter@math.berkeley.edu

11

12

13

14

15

16

17

18

19

20

21

22

23

24 **Abstract**

25 **Background**

26 A recent study of the gene expression patterns of Zika virus (ZIKV) infected human neural
27 progenitor cells (hNPCs) revealed transcriptional dysregulation and identified cell-cycle-related
28 pathways that are affected by infection. However deeper exploration of the information present
29 in the RNA-Seq data can be used to further elucidate the manner in which Zika infection of
30 hNPCs affects the transcriptome, refining pathway predictions and revealing isoform-specific
31 dynamics.

32

33 **Methodology/Principal Findings**

34 We analyzed data published by Tang *et al.* using state-of-the-art tools for transcriptome
35 analysis. By accounting for the experimental design and estimation of technical and inferential
36 variance we were able to pinpoint Zika infection affected pathways that highlight Zika's neural
37 tropism. The examination of differential genes reveals cases of isoform divergence.

38

39 **Conclusions/Significance**

40 Transcriptome analysis of Zika infected hNPCs has the potential to identify the molecular
41 signatures of Zika infected neural cells. These signatures may be useful for diagnostics and for
42 the resolution of infection pathways that can be used to harvest specific targets for further study.

43

44

45

46

47

48

49 **Introduction**

50 As infection with Zika virus (ZIKV) is associated with increasing cases of congenital
51 microcephaly and adult Guillain-Barre Syndrome, a characterization of its pathophysiology
52 becomes crucial. A molecular characterization of the effects of infection may help in the
53 development of fetal diagnostics and can accelerate the identification of crucial genes and
54 pathways critical in disease progression. RNA-Sequencing (RNA-Seq) is an effective
55 technology for probing the transcriptome and has been applied to study the effects of ZIKV
56 infection of human neuroprogenitor cells (hNPCs) [1].

57 While initial analyses of the data have been used to conduct a general survey of
58 transcriptome changes upon infection [1-3], they [1,2] used a method, Cufflinks/Cuffdiff [4], that
59 fail to take advantage of the experimental design used in Tang et. al [1]. They [1,2,3] also do not
60 examine transcriptome dynamics at the isoform level.

61 We apply the recently developed kallisto [5] and sleuth [6] programs to improve the
62 accuracy of quantification and to extract information from the data that was previously
63 inaccessible. We find that sleuth's improved control of false discovery rate results in differential
64 transcript and gene lists that are much more specific and that are significantly enriched in
65 neurodevelopmental pathways. They reveal ZIKV's neural tropism and the host's response to
66 viral infection. Furthermore, we demonstrate that the combination of accurate kallisto
67 quantification, assessment of inferential variance and the sleuth response error model allows for
68 the detection of post infection isoform-specific changes that were missed in previous analyses.

69 The sleuth Shiny app drives a freely available website that allows for reproducibility of
70 our analyses, and provides tools for interacting with the data. This makes the dataset useful for
71 analysis by infectious disease experts who may not have bioinformatics expertise.

72

73

74 **Methods**

75 We ran kallisto and sleuth on a total of eight samples of ZIKV infected and mock infected
76 hNPCs (GEO: Series GSE78711). The runs were performed on a laptop and can be repeated
77 using the provided scripts at <http://www.github.com/pachterlab/zika/>. We used kallisto to
78 pseudo-align the RNA-seq reads, building an index using the ENSEMBL GRC38 release 85
79 *Homo sapiens* transcriptome and using default parameters (kmer size = 31, fragment length =
80 187 and sd = 70 for the single end reads), quantifying transcript abundances, and performing
81 100 bootstraps per sample. To identify differentially transcripts and genes we first modified
82 sleuth to be able to take advantage of the technical replicates performed by Tang et. al [1]. This
83 was done by replacing an estimate of inferential variance from an average of bootstrap
84 estimated variances to a weighting based on the number of fragments in each replicate. The
85 response error model of sleuth was then used to identify statistically significant differential
86 genes and transcripts.

87 Table 1a: Experimental design and inferential variance estimation weights

Sample	Accession Number	Condition	Seq method	Seq machine	Reads	Fragments / weights
Mock1-1	SRR3191542	mock	paired-end	MiSeq	15855554	7927777
Mock2-1	SRR3191543	mock	paired-end	MiSeq	14782152	7391076
ZIKV1-1	SRR3191544	zika	paired-end	MiSeq	14723054	7361527
ZIKV2-1	SRR3191545	zika	paired-end	MiSeq	15242694	7621347
Mock1-2	SRR3194428	mock	single-end	NextSeq	72983243	72983243
Mock2-2	SRR3194429	mock	single-end	NextSeq	94729809	94729809
ZIKV1-2	SRR3194430	zika	single-end	NextSeq	71055823	71055823
ZIKV-2-2	SRR3194431	zika	single-end	NextSeq	66528035	66528035

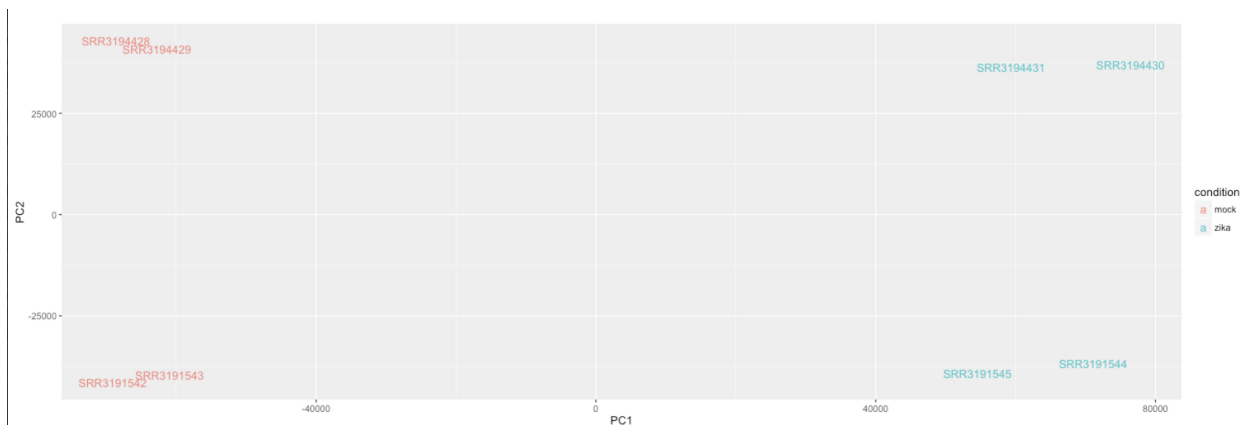
88

89

90 Results

91 We detected 4610 transcripts across 3646 genes that are differentially expressed
92 between ZIKV and mock infected samples (false discovery rate of 0.05) (Fig 1: principle
93 component analysis. S1 Table: differentially expressed transcripts, sorted by significance level).
94 2895 of the 3646 differentially expressed genes were also reported in Tang et. al [1], but they
95 report an additional 3969 genes that we failed to find containing a significant transcript (they
96 found a total of 6864 significant genes), whose 18423 transcripts have an average qual = 0.55.
97 Furthermore, we found 751 differentially expressed genes corresponding to 5426 transcripts not
98 detected by Cufflinks.

99



100

101 Figure 1: Principle component analysis shows that the primary contributor to variance is whether
102 the sample is ZIKV-infected or mock-infected. The secondary component is method of
103 sequencing, i.e. paired-ends or single-end.

104

105 The statistics and figure above, along with interactive data visualization tools, can be
106 found via Sleuth's Shiny app: <http://lair.berkeley.edu/tang16/>.

107

108

109 **Zika induced isoform divergence.**

110 Differentially regulated genes may be missed in gene-level differential analysis for
111 several reasons. Noise in the measurement of a highly expressed transcript can mask
112 expression changes in lowly expressed transcripts. In the case of isoform switching, the
113 upregulation in one isoform can “cancel out” the effects of downregulation in another. We
114 identified 108 genes that undergo isoform divergence as a result of infection, where isoform
115 divergence is defined as a gene containing one or more transcripts that are significantly
116 upregulated and at least one other transcript that is significantly downregulated (see S2 Table of
117 isoform diverging transcripts with statistics). Of these 108, 57 genes were not considered
118 differentially expressed genes by Cuffdiff analysis, corresponding to 150 transcripts.

119 An analysis on these 108 isoform diverging genes using Reactome pathway analysis [7]
120 show pathway enrichment in neuronal system (specifically transmission across chemical
121 synapses and protein-protein interactions at the synapses), developmental biology (specifically
122 axon guidance), immune system, DNA repair, chromatin modifying enzymes, gene expression
123 (rRNA and transcriptional regulation), metabolism, signal transduction, transmembrane
124 transport and vesicle-mediated transport.

125 One of these 57 isoform diverging genes not picked up by Cufflink is NRCAM, neuronal
126 cell adhesion molecule, which according to Gene Cards is putatively involved in neuron-neuron
127 adhesion and axonal cone growth. Another is CHRNA7, cholinergic receptor nicotinic alpha 7
128 subunit. [8] See Fig 2a and 2b for plots of the changes in their transcripts levels.

129

130



131

132

133 Figure 2a and b: Examples of genes with divergent isoforms, NRCAM and CHRNA7, viewed

134 with our Shiny app. For a specific gene, it displays each transcript and the abundances

135 corresponding to each sample.

136

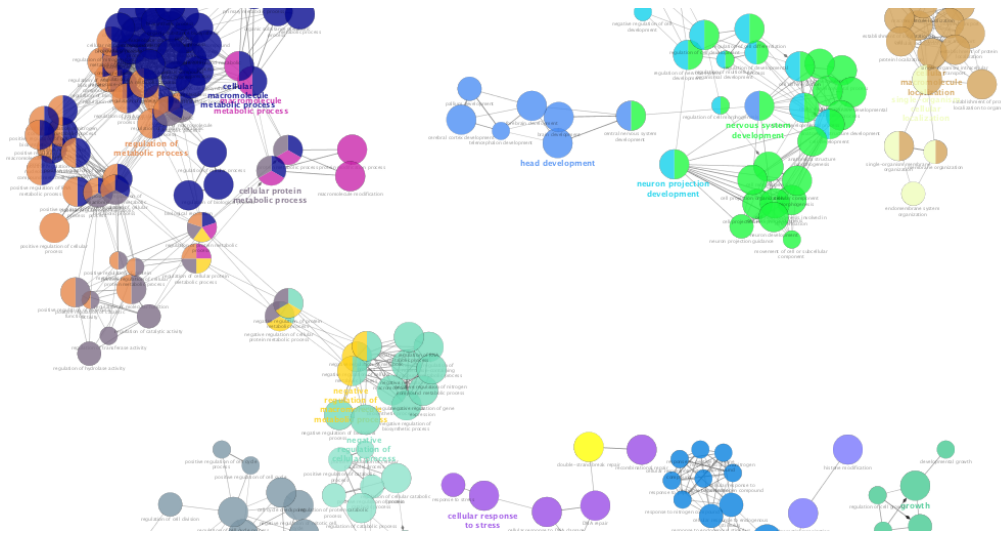
137

138

139

140 **A gene ontology (GO) analysis of sleuth-discovered genes showcase neural and head**
141 **development networks.**

142 We analyzed the set of 3656 genes with differentially regulated transcripts with a gene
143 ontology tool, ClueGO a plugin for Cytoscape [9, 10], over the Biological Processes ontology
144 network, using GO Term Fusion. We set the network specificity to global (GO tree interval: 1-4),
145 using pathways with a minimum of 50 genes and kappa score of 0.5. The enriched nodes of
146 particular interest include neuron projection guidance (pval = 2.7E-3 vs >0.05 with Cuffdiff),
147 cerebral cortex development (1.6E-7 vs >0.05), neuron development (9.9E-6 vs 3.9E-4), neuron
148 projection development (1.8E-6 vs 5.0E-5), nervous system development (3.0E-10 vs 1.0E-9),
149 central nervous system development (6.9E-9 vs 1.0E-4), brain development (2.8E-9 vs 8.0E-4),
150 forebrain development (1.9E-7 vs 4.1E-2), telecephalon development (2.7E-5 vs 5.2E-3), head
151 development (pval = 1.3E-6 vs 3.2E-4), and cellular response to stress (9.4E-26 vs 7.3E-22).
152



153
154 Figure 3: A subgraph of the network resulting from ClueGO analysis of the differentially
155 regulated genes discovered by kallisto and sleuth.

156
157

158 A GO analysis results tables and cytoscape JSON file (see Supplementary materials) that can
159 be used to render the network in cytoscape.

160

161 **Discussion**

162 RNA-Seq can provide rapid and highly resolved probing of infection response, and initial
163 studies of Zika infection dynamics highlight neurally active isoforms, genes and pathways that
164 may play an important role in disease etiology. However the simplicity of RNA-Seq library prep
165 and cDNA sequencing belies the complexity of analysis. We have shown that a careful analysis
166 of previously published data can reveal novel targets with higher confidence, and in the process
167 rendering a valuable dataset usable by the community of Zika researchers.

168

169 The kallisto and sleuth tools we have used in our analysis are particularly powerful when
170 coupled with the interactive sleuth Shiny application, and our publicly available server providing
171 access to our analysis contains numerous interactive plots and analyses that cannot be
172 reproduced in a static publication. This highlights the utility and importance of data sharing [11],
173 and we hope that our analysis, aside from its usefulness for the Zika scientific community, can
174 also serve as a blueprint for future data sharing efforts.

175

176 sleuth is a fast and accurate pipeline for analyzing RNA-Seq data that allows for testing
177 at the isoform level. The alignment and quantification pipeline is feasible and compatible with a
178 standard desktop computer. The interactive Sleuth application, made publically available, allows
179 for informative data visualization, including those of library prep fragment size distributions,
180 principle component analysis, and gene and transcript expression changes. We invite the
181 scientific community studying Zika to utilize this toolkit.

182

183 **References**

- 184 [1] Tang H, Hammack C, Ogden SC, Wen Z, Qian X, Li Y, Yao B, Shin J, Zhang F, Lee EM,
185 Christian KM, Didier RA, Jin P, Song H, Ming GL, Zika Virus Infects Human Cortial Neuro
186 Progenitors and Attenuate Their Growth. *Cell Stem Cell*. 2016 May 5;18(5):587-90.
- 187 [2] Rolfe AJ, Bosco DB, Wang J, Nowakowski RS, Fan J, Ren, Bioinformatic analysis reveals
188 the expression of unique transcriptomic signatures in Zika virus infected human neural stem
189 cells, *Cell Biosci*. 2016 6:42, doi: 10.1186/s13578-016-0110-x.
- 190 [3] Wang Z, Ma'ayan A, An open RNA-seq data analysis pipeline tutorial with an example of
191 reprocessing data from a recent Zika virus study. *F1000Research* 2016, 5:1574,
192 doi:10.5256/f1000research.9804.r14924.
- 193 [4] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg,
194 J.L. Rinn and L. Pachter, Differential gene and transcript expression analysis of RNA-seq
195 experiments with TopHat and Cufflinks, *Nature Protocols*, 7 (2012), 562--578.
196
- 197 [5] Bray NL, Pimentel H, Melsted P and Pachter L, Near-optimal probabilistic RNA-seq
198 quantification, *Nature Biotechnology* 34, 525–527 (2016), doi:10.1038/nbt.3519.
199
- 200 [6] Pimentel HJ, Bray N, Puente S, Melsted P, Pachter L, Differential analysis of RNA-Seq
201 incorporating quantification uncertainty, *bioRxiv* 058164; doi: <http://dx.doi.org/10.1101/058164>.
202
- 203 [7] Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie
204 M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky
205 V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P. The Reactome pathway
206 knowledgebase. *Nucleic Acids Res*. 2014 Jan;42(Database issue):D472-7. doi:
207 10.1093/nar/gkt1102. Epub 2013 Nov 15.
208
- 209 [8] Stelzer G, Rosen R, Plaschkes I, Lieder I, Zimmerman S, Twik M, Fishilevich S, Nudel R, Kohn A,
210 Mazor Y, Kaplan S, Iny Stein T, Warshawsky D, Guan-Golan Y, Rappaport N, Safran M, and Lancet
211 D. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis*, *Current*
212 *Protocols in Bioinformatics* (2016), *in press*
213
- 214 [9] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, et al. ClueGO: a Cytoscape plug-
215 in to decipher functionally grouped gene ontology and pathway annotation networks.
216 *Bioinformatics*. 2009, 25(8):1091-3.
217
- 218 [10] Bindea G, Galon J, Mlecnik B, CluePedia Cytoscape plugin: pathway insights using
219 integrated experimental and in silico data. *Bioinformatics*. 2013, 29(5):661-3.
220
- 221 [11] Longo D, Drazen J, Data Sharing. *N Engl J Med* 2016; 374:276-277, [January 21, 2016](#),
222 DOI: 10.1056/NEJMe1516564.
223
224