Window of Susceptibility for Zika-Induced Microcephaly Identified by Temporal Gene Analysis

Kristin P. Bennett⁴, Elisabeth Brown¹, Hannah De los Santos¹, Matthew Poegel¹, Thomas R. Kiehl⁴, Evan W. Patton³, Deborah L. McGuinness²,³,⁵, Sally Temple⁴, Nathan C. Boles⁴

Affiliations

¹Department of Mathematical Sciences; Rensselaer Polytechnic Institute; Troy, NY 12180, USA

²Institute for Data Applications and Exploration, Rensselaer Polytechnic Institute; Troy, NY 12180, USA

³Department of Computer Science; Rensselaer Polytechnic Institute; Troy, NY 12180, USA

⁴Neural Stem Cell Institute; Rensselaer, NY 12144, USA

⁵Department of Cognitive Science; Rensselaer Polytechnic Institute; Troy, NY 12180, USA

*Correspondence to: bennek@rpi.edu, nathanboles@neuralsci.org

Abstract

Increased understanding of developmental disorders of the brain has shown that genetic mutations, environmental toxins and biological insults typically act during developmental windows of susceptibility. Through analysis of developmental time-course gene expression data derived from human pluripotent stem cells, with disease association, pathway, and protein interaction databases, we identify windows of developmental time that appear most vulnerable to a specific insult, and therefore, the time periods for productive interventions. The results are displayed as interactive Susceptibility Windows Ontological Transcriptome (SWOT) Clocks illustrating disease susceptibility over developmental time. Using this method, we determine the likely windows of susceptibility for multiple neurological disorders, including Zika-induced microcephaly. We find that genes impacted by Zika infection are most active in the earliest stages of neural development, prior to cerebral cortex layer formation.
One Sentence Summary: SWOT Clocks semantically integrate temporal expression data to identify windows of susceptibility during development for disease, including Zika-induced microcephaly.
Development of an organism requires an intricate interplay of cell expansion, differentiation, and the integration of numerous signaling inputs across time and space, with disruptions in the action or timing of these signals potentially having drastic consequences for normal development. Multiple developmental disorders have been linked to exposure of mothers and their fetuses to a wide array of genetic, biological and chemical insults (1-4). Yet, ascertaining the timing and duration of exposure needed to cause harm during human development is a difficult undertaking, due to both ethical and technological constraints. Here, we describe a computational method to predict windows of susceptibility (WOS) to injury from disrupting agents, using publically available human pluripotent stem cell (hPSC) temporal gene expression databases along with a semantic infrastructure for linking and term disambiguation to provide links to a possible disease outcome and to suggest potential mechanisms. The method is generally applicable to time-course expression data derived from hPSC models, which can be analyzed for WOS for a wide range of developmental disorders. As proof of concept, we utilized expression data covering a time-course of human cerebral cortex development from hPSCs (5) and identified putative periods of vulnerability for a variety of diseases that impact corticogenesis, including Zika virus-induced microcephaly.

The RNA-seq expression data used in this analysis was derived from an in vitro model of cerebral cortical development, covering nine distinct developmental time points (days 0, 7, 12, 19, 26, 33, 49, 63, 77) from hPSCs to deep and upper cortical layer production (Fig. 1A) and identifying 14,065 significantly changing transcripts (5). In the present analysis, the data was first standardized to create suitable counts for making accurate comparisons between different sized genes (Supplemental Methods). We then performed a Singular Value Decomposition (SVD) analysis to precisely characterize the temporal signatures (6, 7). SVD decomposes the transcriptome into factors and virtual eigengenes that capture the variation attributed to the genes and time-points; SVD succinctly captures the temporal dynamics of corticogenesis by delineating a unique molecular signature with different waves of gene transcription during cortical development.

In order to understand more fully how gene expression changed over time, the SVD analysis was then coupled with Fuzzy C-Means clustering. The means of the six clusters for each of the days of the analysis shown in the six inset plots appear in a clear sequence (Fig. 1B). These clusters have a temporal ordering directly corresponding to the five stages of corticogenesis previously identified (5), plus one additional stage previously undescribed that includes about 1/7 of the genes. This stage was found to be enriched for processes related to Neuroectoderm using enrichment analysis by DAVID (8). This resulted
in the following stages: Pluripotency (PP), Neuroectoderm (NE), Neural Differentiation (ND), Cortical Specification (CS), Deep Layers (DL), and Upper Layers (UL) (Fig. 1B). The results of this unbiased analysis form a circular display: the temporal transcriptomic “Clock” (Fig. 1B). The molecular signatures represented by this clock allow us to understand the temporal pattern of expression of each gene during corticogenesis.

It is now possible to further interrogate this temporal dataset by querying against publically available protein interaction, disease and signaling pathway datasets. To facilitate this process, we developed a semantic approach that supports integration and linking of diverse data utilizing ontologies from online resources (Fig. 1C). We further created a novel visualization called a SWOT (Susceptibility Windows Ontological Transcriptome) Clock that combines the associative visual of chord diagrams and the simplicity of a time-based heat map and leverages domain ontologies to facilitate linking of the datasets. The viewer is able to simultaneously analyze the connections between transcriptomes and the experimental time series data, facilitated using the SWOT Clock web tool (https://semnext.tw.rpi.edu/swotclock/). The chords in the center of the diagram illustrate the connections between the transcriptomes based on the STRING database, while the heat map depicts the RNA-Seq expression data measured over the nine time points with day 0 in the inner ring and day 77 in the outer ring. The SVD analysis orders the genes clockwise showing the wave of expression moving clockwise from PP to UL (Fig. 2A-F). The interactive web-based visualization allows the viewer to engage with the Clock by highlighting different connections to get more information about the entity or link using the ontology, and to show only connections for particular stages or transcriptomes that are of interest. Alongside the SWOT Clock, the tool also provides an enrichment analysis calculating the Log Odds Ratio and p-value for each stage in a given Clock. Using this information, we define the dominant stage as the stage with positive logs and lowest p-value. This allows the user to quickly identify stages that are significantly enriched or depleted for a disease, pathway or other transcriptome set.

Using this application in combination with a literature based disease-gene association database (http://diseases.jensenlab.org/Search), we performed enrichment analysis to determine whether each stage of cortical development was enriched or depleted for a neurodevelopmental disorder. This was done by computing the Log Odds Ratio (LOR) for each cluster and disease, as well as the p-values corresponding to the disease and stage. If the LOR is negative, the cluster is likely to be depleted for genes associated with that certain disease, with more negative numbers indicating a stronger depletion.
A positive LOR indicates that the cluster is likely to be enriched for genes associated with that disease, with more positive numbers showing a stronger enrichment. We then calculated 2-sided p-values using the LOR test to evaluate the statistical significance of these results at an FDR≤ 0.1, allowing for the identification of dominant stages as potential WOS to biological or environmental agents in the development of a disease (Table S1).

From this analysis, the enrichment and depletion of disease genes for each stage of development becomes evident. For example, Tauopathy and Autism were both significantly enriched for UL (Figure 2A-B). Loss of *Mapt* (Tau) in mice has been demonstrated to affect the development of the UL due to defects in migration of layer II/III neurons (9). Additionally, a previous study determined exposure to air pollution of pregnant women in the 3rd trimester as the WOS for increasing autism risk for their fetus, both of which is in agreement with our predictions(10). Microcephaly and holoprosencephaly, two conditions that affect early brain formation and corticogenesis, were enriched for NE and ND stages, respectively (Fig. 2C-D). Compellingly, holoprosencephaly’s WOS has already been established to be during the 5th and 6th weeks of pregnancy(11) matching our prediction (Fig. 1A & 2C). In line with the predicted WOS for microcephaly, multiple studies have linked impaired proliferation of early neural progenitors to the pathogenesis of this disease(11). Additional analysis can indicate which signaling pathways likely contribute to disease pathogenesis. For instance, using the KEGG pathways database, we found statistically significant enrichment for the ‘Spliceosome’ and ‘Cell Cycle’ during the NE stage and overlapping with the WOS for microcephaly, further supporting a role for disruptions in splicing and cell cycle control leading to microcephaly (11-13) (Fig. 2E-F).

The potential role of Zika virus as a causative agent in microcephaly constitutes an international health crisis (4, 14). Several recent studies have examined the role of Zika infection in the pathogenesis of microcephaly, including those using hPSC models (15). As already demonstrated, the WOS for microcephaly is in the NE stage of corticogenesis. To further investigate the potential effects of Zika infection and the WOS for Zika related microcephaly, we retrieved recently published gene expression data obtained from human cortical progenitors derived from pluripotent stem cells infected with the Zika virus (15). We identified 1431 significantly changing genes, with 539 being up-regulated and 892 down-regulated (Supplemental Methods, Table S2). We then constructed SWOT Clocks for both the down-regulated and up-regulated genes (Fig. 3A&B).
When examining the down-regulated subset, the WOS for Zika virus infection was identified as the Neuroectoderm stage, similar to the microcephaly SWOT Clock. Furthermore, this subset of genes were enriched for the cell cycle KEGG pathway, similar to the WOS genes for microcephaly (Fig. 3C). However, this subset was not enriched for the spliceosome KEGG pathway, illustrating that the Zika virus may not depend on altering splicing to induce microcephaly. When considering the up-regulated subset, a different picture emerges, with the SWOT Clock indicating a role for these genes in the Pluripotency stage, potentially affecting the metabolism pathways of these cells. Additionally, the up-regulated list is enriched for GO categories from our ontology dealing with the endoplasmic reticulum and its response to stress, which is a known effect of Zika infection. Overall, this data analysis establishes a very early WOS for Zika-induced microcephaly, which is in agreement with a retrospective study of Zika associated microcephaly (16).

To gain insight into how Zika could lead to microcephaly, we examined the intersection between known microcephaly genes changing during development and the Zika affected genes and surprisingly did not find any overlap. We then expanded our microcephaly list to include 1st degree neighbors in our interaction network and found significant overlap. We constructed a SWOT Clock (Fig. 3D) that indicates that Zika viral infection upregulates genes connected to microcephaly genes that are expressed in the Pluripotency stage while down regulating genes that are normally highly expressed early in the Neuroectoderm stage. Thus, we find that microcephaly-related genes impacted by Zika are most active in the earliest stages of neural development, the foundation of building the cerebral cortex. A network analysis identified three distinct communities and processes underlying the function of intersection genes (Fig. 3E, Supplemental Methods), revealing subsets of genes affected by Zika virus infection, yet not previously implicated in microcephaly, as likely candidates for further study (Table S3). One mis-regulated candidate, MYC, has previously been implicated in brain development (17) and is a highly connected node in our network (Fig. 3E) making it an attractive target for a role in the pathogenesis of Zika induced microcephaly. Overall, this analysis highlights the effectiveness of our method in identifying WOS and guiding studies into possible agents of disease.
Acknowledgments

The authors would like to thank the following funding sources: Rainwater Foundation, NSF 1331023.


17. A. Wey, P. S. Knoepfler, c-myc and N-myc promote active stem cell metabolism and cycling as architects of the developing brain. *Oncotarget* **1**, 120 (Jun, 2010).
Figure Legends

Figure 1. SVD reveals a developmental clock in expression data.
Singular Value Decomposition (SVD) was used to analyze gene expression data derived from RNA-seq over a developmental time-course covering corticogenesis. (A) Diagram illustrating how each time point in the protocol relates to human development. (B) Corticogenesis clock formed by genes plotted in space spanned by first two left singular vectors colored by their corresponding Fuzzy C-Means clusters which match the stages of corticogenesis. The average cluster profiles from Fuzzy C-Means are shown placed at their position in the clock. (C) Flow diagram of work flow for establishing the SWOT clocks for each disease using expression data and integrating publically available databases using an ontological semantic enhanced analysis.

Figure 2. SWOT Clocks for multiple neurological disorders.
(A-D) SWOT clocks for four neurological disorders are presented. Each clock shows the connections between genes involved in each disease while simultaneously illustrating each gene’s expression in developmental time and calculating the most enriched stage and likely WOS. SWOT clocks for the KEGG pathways (E) Spliceosome and (F) Cell Cycle demonstrate a similar profile to the SWOT clock for Microcephaly. Legend: Pluripotency (PP); Neuroectoderm (NE), Neural Differentiation (ND), Cortical Specification (CS), Deep Layers (DL), Upper Layers (UL).

Figure 3. Genes down-regulated by Zika infection of cortical progenitors reveal a SWOT clock similar to microcephaly.
Expression data for Zika infected cortical progenitors were re-analyzed using more stringent criteria to establish significantly changing genes and SWOT Clocks of the (A) down-regulated and (B) up-regulated genes were generated. The down-regulated SWOT Clock showed a similar profile to the Microcephaly clock. Using goseq, enrichments in KEGG pathways for either up-regulated or down-regulated significantly changing genes were calculated. A SWOT Clock for (C) the Cell Cycle KEGG pathway genes enriched for the down-regulated genes was generated. (D) Intersection of Microcephaly genes plus 1st degree neighbors with the Zika genes filtered through the developmental time-course data. Only nodes with a betweeness centrality above 0.02 are displayed for clarity’s sake. Legend: Pluripotency (PP); Neuroectoderm (NE), Neural Differentiation (ND), Cortical Specification (CS), Deep Layers (DL), Upper
Layers (UL). Network analysis of Intersection genes shows distinct three communities contribute to the gene intersection.
Figure 1

A

hPSC Cortical Differentiation

Days post Induction

0 7 12 19 26 33 49 63 77

PP NE ND CS DL UL Maturation

1st Trimester 2nd Trimester 3rd Trimester

Weeks post Conception

Human Gestation

B

Pluripotency (PP)
Neuroectoderm (NE)
Neural Differentiation (ND)
Cortical Specification (CS)
Deep Layers (DL)
Upper Layers (UL)

C

SVD and Fuzzy Clustering
Scale and Normalize Data
Developmental Temporal Expression Data (Cortecon)
Stage Enrichments and Visualization in SWOT Clock
Ontological Semantic Enhanced Analysis
KEGG Pathways
STRING PPI
ZIKAVirus Regulated Genes in Cortical Progenitors
Literature Based Gene-Disease Association dB

Maturation

0 7 12 19 26 33 63 77 49

Weeks post Conception

Days post Induction

Human Gestation

Components 1 and 2

ZIKAVirus Regulated Genes in Cortical Progenitors
Stage Enrichments and Visualization in SWOT Clock
Ontological Semantic Enhanced Analysis
SVD and Fuzzy Clustering
Scale and Normalize Data
Developmental Temporal Expression Data (Cortecon)
Figure 2

A  Holoprosencephaly
   ND p<0.01

B  Autism
   CS p<0.05; UL p<0.01

C  Tauopathy
   UL p<0.05

D  Microcephaly
   NE p<0.01

E  Spliceosome
   NE p<0.01

F  Cell cycle
   NE p<0.01
Figure 3

A  Zika Down-regulated Genes

B  Zika Up-regulated Genes

C  Cell Cycle

D  Intersection of significantly changing genes in ZIKV infection and Microcephaly 1st degree neighbors

E  Intersection of significantly changing genes in ZIKV infection and Microcephaly 1st degree neighbors

Betweenness Centrality (Node Size)

0.02  0.18

DNA Replication

Cell Division

Positive Regulation of Transcription from RNA Pol-II promoter