Mapping regional changes in the glycerophosphocholine second messenger lipidome following brain injury using CIRCOS

Graeme P. Taylor*, Hongbin Xu*, Andrew J. Syrett, and Steffany A.L. Bennett

Ottawa Institute of Systems Biology, Neural Regeneration Laboratory, Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, K1H 8M5, Canada.
andrew.syrett@gmail.com,{gtaylo2, hxu081, sbennet}@uottawa.ca
*joint first authors

Abstract. Although it is clear that lipid deregulation is a critical component of many neurodegenerative diseases, including stroke and Alzheimer's disease, the tools to adequately study such changes are still being developed. While acquisition of lipidomic data has been greatly improved with the development of new technologies, the bottleneck of this type of "omic" research still lies in data analysis and post-processing. Individual lipid species need to be identified, which is being addressed by new tools such as VaLID (Visualization and Phospholipid Identification). However, we still lack the representational capacity to identify important species and critical relationships within large datasets that conventional visualization methods obfuscate. Here we present a Circos-based method to display integrative relationships across multi-modal lipidomic datasets, and develop a method to intuitively explore these data for biologically relevant relationships.

Keywords: lipidomics · visualization · mass spectrometry · bioinformatic knowledge visualization

1 Introduction

Advances in genomics have enabled researchers to identify genetic determinants of early onset disease. Direct biochemical investigations have elucidated multiple signalling pathways altered by these mutant gene products. Further combination of genomics with proteomics is allowing researchers to map global gene and protein changes associated with progressive neurodegeneration. Together, these studies have provided remarkable insight into the molecular nature of neurodegenerative disease [1, 2]. Despite these advances, we do not know why discrete subsets of cells are uniquely susceptible to specific neurodegenerative injuries nor can we protect vulnerable populations. We argue that the next major advance in rational therapeutic design will come from tying a cell's metabolome into genomic and proteomic maps of diseases. We argue that a "targeted systems approach" will provide a new understanding into the metabolic regulatory mechanisms that determine why groups of cells become transiently susceptible and finally succumb to genetic and environmental determinants

of diverse neurodegenerative disorders. Such insight will provide new therapeutic avenues designed to protect vulnerable neuronal populations in the face of ongoing insult. This understanding, however, requires capacity to simultaneously visualize, analyse, and disseminate defining spatial and temporal relationships within complex networks from large datasets in a meaningful and intuitive fashion.

Lipidomics represents the sub-field of metabolomics in greatest need of bioinformatic support [3]. Neurodegenerative disorders such as stroke and Alzheimer's disease are associated with deregulation of lipid metabolism [4-6]. Despite extensive research, the etiology and pathophysiology of these diseases at the cellular and molecular levels are poorly understood and effective treatments have been elusive [7]. Our progress in understanding the impact of alterations in lipid metabolism has been largely hampered by lack of adequate technology capable of coping with the structural complexity and diversity of lipids. This has prevented proper studies of lipid dynamics that would allow identification of regionally specific changes in degenerating tissue. The advent of high-throughput systems-level analysis of lipids and their interacting and modifying partners in cells, tissues, and organs, referred to as lipidomics, has renewed interest and advanced our understanding of the roles of membrane lipids and their metabolites in brain injury [8]. The term lipidomics defines "the large-scale analysis of lipid profiles in cells and tissues" [9]. Neurolipidomics focuses on cellular, regional, and systemic lipid homeostasis in the central nervous system, encompassing not only the identification and measurement of individual lipid isoforms but also the mRNA and protein expression profiles of metabolic enzymes and transporters and the protein targets that effect downstream signalling [10]. Lipidomes represent a multitude of different lipid classes (i.e., glycerophospholipids, sphingolipids, sterols, galactolipids/sulfolipids, and archael tetraether lipids) where each class may be composed of a large number of distinct subclasses, molecular species [12]. For example, membrane phospholipids are derivatives of *sn*-glycero-3-phosphoric acid with (a) a 1-*O*-acyl, a 1-*O*-alkyl (ether-linked plasmanyl), or a 1-*O*-alkyl-1'-enyl (vinyl ether-linked plasmenyl) carbon chain, (b) a long-chain fatty acid esterified to the *sn*-2 position, and (c) a polar headgroup composed of a nitrogenous base, a glycerol, or an inositol unit modifying phosphoric acid at the *sn*-3 position. The polar head group defines 20 different phospholipid classes (i.e., phosphatidylserine (PS), phosphatidylethanolamine (PE), and phosphatidylcholine (PC) etc.,). Individual species are distinguished by their particular combination of fatty acyl residues (chain length and degree of unsaturation) and by the nature of their *sn*-1 or *sn*-2 chemical linkages (acyl, alkyl, or alkenyl). PC(*O*-16:0/20:4), for example, defines a lipid species with a phosphocholine polar head group (PC), an ether linkage at the *sn*-1 position (*O*-), 16 carbons at the *sn*-1 and 20 carbons at the *sn*-2 positions, of which the *sn*-1 chain is fully saturated (referred to as :0) whereas the *sn*-2 chain is characterized by four double bonds (indicated by :4). A typical eukaryotic lipidome may contain somewhere from 9,000 to ~100,000 individual molecular lipid species representing the major lipid families with regional and compartmental specificity [11-13].

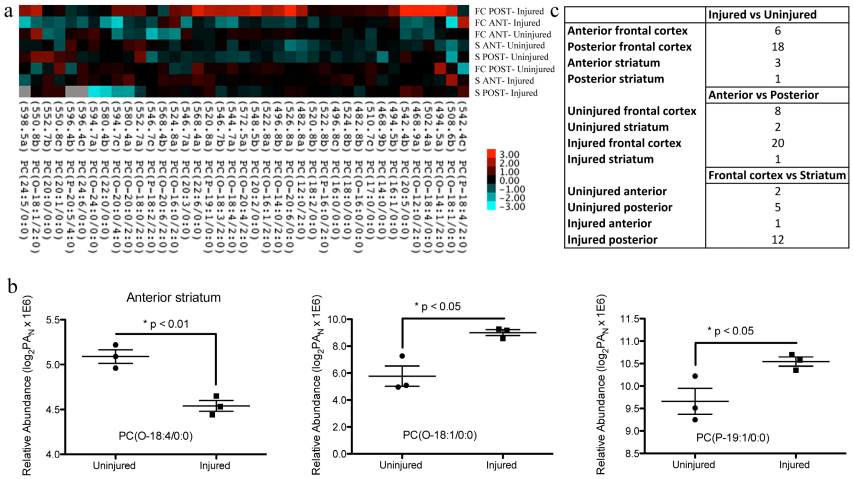
Modern lipidomic methodologies, such as the current workhorse – liquid chromatography electrospray ionization mass spectrometry (LC-ESI-MS), enable simultaneous identification and quantification of hundreds to thousands of lipid molecular

species from a single sample. A typical MS-based lipidomic workflow includes (1) hypothesis generation, (2) sample preparation, (3) MS analysis (4) spectral data processing using lipid-centric tools for identification and quantification of detected lipid species, and (5) interpretation of acquired lipidomic data through statistical data analysis and data visualization [14]. A major roadblock to conveying the importance of membrane metabolism in health and disease is the capacity to visualize changes in these complex networks. The key is to synthesize this broad spectrum of acquired data to better understand the biological systems used to combat diseases. Proper data visualization is a critical aspect of both the analysis and understanding of these data.

Like all other "omics," lipidomic studies of biological systems are rapidly accumulating a wealth of medium to large-scale compositional data. This, in turn, has created new challenges for scientists to provide adequate methods for data exploration and visualization. Conventionally, lipidomic data are graphically presented as heatmaps where individual values contained in a matrix are represented as colors showing relative abundance or levels of expression. Subsequently, lipid species whose levels change significantly are visualized using tables and/or various individual plots. Here, in the example of one of our neurolipidomic datasets quantifying changes in phosphocholine second messengers following murine brain injury, multiple layers of comparisons are made, including striatum vs. cortex (tissue specific), anterior vs. posterior subregions of each anatomical structure (spatial), uninjured vs. injured (health condition), lipid subfamily vs. lipid subfamily (structural and metabolic) etc. Conventionally, this type of lipidomic dataset is visualized using a) a heatmap to display changes of each individual lipid species across all conditions (Fig. 1a), b) multiple individual plots (Fig. 1b) to reveal significantly altered species in all planned comparisons (only 3 out of 79 possible plots are shown), and c) a table to summarize the number of significantly changed species in each planned comparison (Fig.1c). This kind of simple presentation is straightforward and easy for readers to identify differences within each comparison. However, it is highly repetitive and difficult to explore intuitively relational links embedded in the data. Moreover, while the dry nomenclature used to name lipids provides precise structural information, these names are difficult to adapt to function withing existing ontologies used to search literature databases. Therefore, to convey a message to readers effectively, sometimes we need more than just a simple bar graph or a table-based report of lipid changes. This is especially true when lipidomic data involve structural, metabolic, spatial, temporal, tissue-specific, and health condition information. Here, to create clear, meaningful, and integrated data visualizations that give biological insight, without being overwhelmed by the intrinsic complexity of these data, we sought for more creative and intuitive data representation methods than those traditionally employed by lipid biochemists and "lipidcentric" systems biologists.

Circos is a data visualization system that builds a single image using multiple layers of data [15]. The goal of this system is to enhance readability by presenting complex datasets using a high "data-to-ink ratio" presented as a circular plot. Circos plots were originally developed for displaying complex genomic relationships in a more interpretable fashion [15], but have since been adapted to many fields [16]. In this study, we have expanded on the Circos methodology and applied it to a medium-

scale lipidomic dataset both to better visualize, and to explore intuitively, relational links present in the data. In this paper, we discuss how the Circos visualization tool was used to represent a medium-sized dataset from a unbiased LC-ESI-MS-based lipidomic profile of the small molecular weight glycerophosphocholine second messenger lipidome in the different spatial regions of the striatum and cortex following brain injury in mice.



	Injured vs Uninjured
Anterior frontal cortex	6
Posterior frontal cortex	18
Anterior striatum	3
Posterior striatum	1
	Anterior vs Posterior
Uninjured frontal cortex	8
Uninjured striatum	2
Injured frontal cortex	20
Injured striatum	1
	Frontal cortex vs Striatum
Uninjured anterior	2
Uninjured posterior	5
Injured anterior	1
Injured posterior	12

Figure 1. Representation of lipidomic data using conventional heatmaps and plots. a) Heatmaps displaying relative abundance of each lipid species in the frontal cortex and striatum of the anterior and posterior brain regions under injured and uninjured conditions. b) The levels of three lipid species were significantly altered in the anterior striatum post injury (Student's *t*-test with Welch correction). c) The number of lipid species that showed significant changes in abundance following injury in each of the 12 comparisons.

2 Circos-based Visualization Method

Lipidomic datasets contain complex linkages between spatially/temporally/behaviourally separated samples. Analyzing these datasets via traditional graphing methods is difficult; examining a multitude of graphs comparing individual data points between samples is exhausting, and does not lead to intuitive discovery of underlying relationships amongst the data. The important "trees" of coherent changes in lipid pathway/class/species can be easily lost in the vast "forest" of lipidomic data. Here we developed a means of visualizing a complex lipid dataset that is comprised of multiple layers of information, and we present results in a highly readable format. This approach is conducive to discovering fundamental trends that would be masked by traditional visualization methods.

2.1 Description of Input Data

The lipidomic data presented were obtained by our group from a mouse brain injury model wherein regional differences in lipid metabolism following unilateral injury of one hemisphere were the focus of study. Our goal was to profile glycerophosphocholine second messengers and to quantify relative abundance of the identified lipid molecular species across different brain (a) tissues (striatum vs. cortex), (b) subregions (anterior vs. posterior), (c) conditions (uninjured vs. injured), and (d) second messenger families (lysophosphatidylcholines (LPCs), platelet activating factors (PAFs), *lyso*-PAFs, plasmalogen PAFs, and other).

2.2 Presentation of Lipidomic Dataset

The visualization approach specifically shows: (a) relative lipid abundance, (b) comparisons of lipid abundance within specific metabolic families between related brain tissues/regions/conditions, and (c) direction of change for each comparison. The overall plot is separated into four major quadrants (Fig. 2), each delineating a different set of brain coordinates; the top half of the plot represents anterior brain regions (and, conversely, the lower half for posterior regions), and the left half of the plot represents the uninjured (contralateral) brain and the right half the injured (ipsilateral) brain. Abundances of each lipid species across n=3 subjects/condition are presented as two heatmaps in concentric rings. These data are median-centered to allow comparison of a lipid species across regions and conditions. The two rings represent the lipid abundance identified in the (a) temporal cortex (outer ring), and (b) striatum (inner ring). Comparisons are made between adjacent quadrants and within quadrants and presented as links (connecting lines). Links are only shown for comparisons of relative lipid abundance that are statistically significantly different. Statistics are multiple t-tests with Welch's correction and using a false detection rate (FDR) of 1% [17] The outer links show comparisons with significant difference in the temporal cortex between adjacent quadrants (regions/conditions), the middle links signify temporal cortex and striatum differences (within the same region/condition), and the inner links display differences in lipid abundance between regions/conditions in striatum. Each of these links is colour-coded for the comparison being made (i.e., contralateral anterior vs. contralateral posterior = light purple, contralateral anterior vs. ipsilateral anterior = dark grey, ipsilateral anterior vs. ipsilateral posterior = rose, ipsilateral posterior vs. contralateral posterior = dark green, temporal cortex vs. striatum = purple). These links also show the direction of change for abundance, indicating the lower lipid abundance with a pointed end and the higher abundance with a broad end.

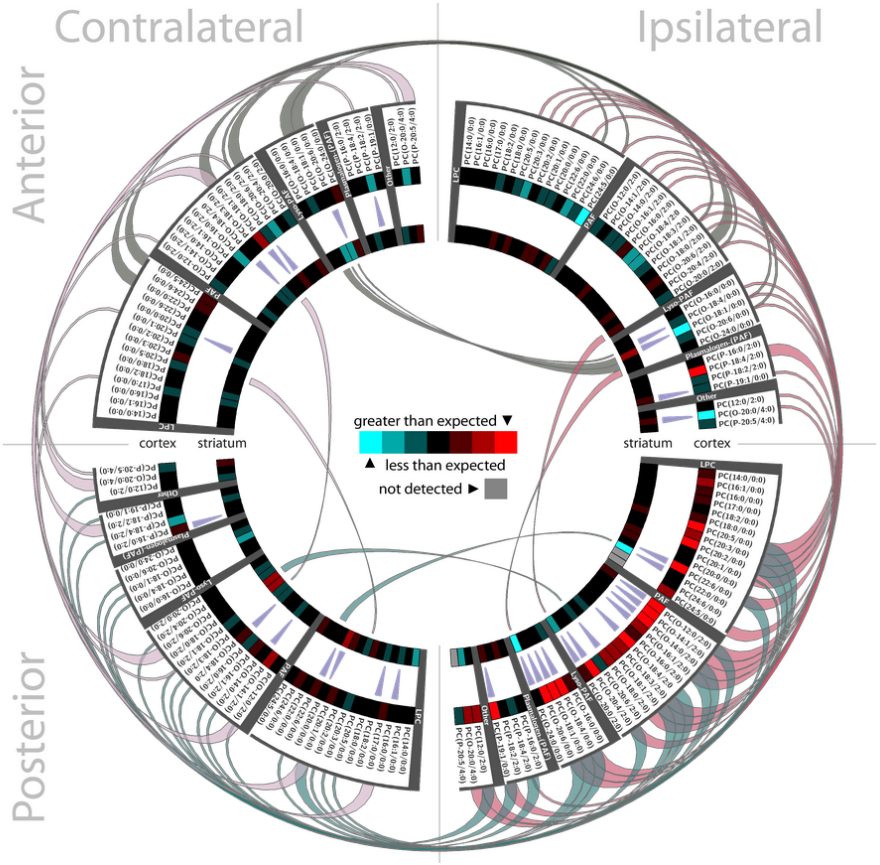


Figure 2. Circos visualization of a lipidomic dataset, showing multiple layers of information including (a) lipid identifications, separated by class, (b) relative lipid abundance, by species and tissue type (temporal cortex and striatum), (c) significantly different lipid abundance in i) temporal cortex in different brain regions/conditions (outer links), ii) striatum in different brain regions/conditions (inner links), iii) temporal cortex and striatum in similar brain regions/conditions (middle links), (d) direction of change in lipid abundances (narrow or broad tips to links)

This visualization method allows for the identification of integrative trends in the data, and can lead to hypotheses regarding the underlying mechanisms responsible for the change (e.g., changes in specific lipid classes, changes in lipid remodelling pathways). In the example diagram (Fig. 2), it can be noted that (a) the cortex consistently has more changes in lipid abundance than striatum, (b) the posterior region of the ipsilateral cortex has consistently higher lipid abundance of all classes of lipids than other positions and regions.

2.3 Use as an Intuitive Exploratory Tool for Lipidomic Dataset

Specific lipid changes can be further investigated through a simple and intuitive web-based interface. When accessed using any JavaScript compatible browser, either a single lipid species or single lipid class can be highlighted by "mousing-over" any informative aspect of that lipid or class (e.g., a lipid name, a specific linkage or heatmap). This highlighting retains the colours of the species/class of interest, and all species and classes not of interest are greyed out.

Also presented is a "sidebar" that displays pertinent information on the specifically highlighted lipid, including: (a) the predicted identity (based on work performed by the CIHR Training Program in Neurodegenerative Lipidomics (CTPNL) [18]), (b) mass/charge, (c) structure of the predicted identity (or most likely structure for lipids with double bonds) (Fig. 3). For highlighted classes the sidebar will instead show a space filling model depiction of a lipid molecule representational of the class. In practice, this can be used to identify which specific lipid species change across regions, positions, tissues, and conditions, and how. In the provided example (Fig. 2), there is a trend suggesting that posterior ipsilateral cortex has the highest lipid abundance. However, by highlighting each lipid species in turn, two specific species can be identified that run counter to this trend (Fig. 3): PC(20:1/0:0) in the posterior ipsilateral cortex has a statistically significant higher abundance than the anterior ipsilateral and a lower abundance than the posterior contralateral cortex (Student's *t*-test with Welch correction, FDR=1% , $p<0.05$,rose-coloured connector, Fig. 3), whereas PC(*P*-18:4/2:0) in the posterior ipsilateral cortex has a significantly lower abundance than the anterior ipsilateral and a higher abundance than the posterior contralateral cortex (Student's *t*-test with Welch correction, FDR=1% , $p<0.05$, rose-coloured connector, Fig. 3). This web-interface also makes for a dynamic and vibrant tool for live presentation of the dataset. High-resolution images of each positional state are also created for use as static figures for print publications.

2.4 Developmental process

The command-line nature of Circos allows automation of multi-step tasks to be performed. The appropriate Circos template was developed, as per the Circos tutorials, and all calculated data and parameters were output into data files for reading by Circos. Custom colours were designated for each component of the heatmaps and links. A separate Circos "chromosome" (for a total of four chromosomes comprising the plot) represents each condition/position. The chromosome was divided into equal lengths to position each class label and lipid species, here a total of 44 "positions." To develop the highlighting ability of the Circos plot, each element of the Circos diagram was tagged with a "URL" tag indicating its relative location in the chromosome. Selective configuration files were generated as input files into the main Circos.conf file – these files "grey out" all components of the full chromosome (e.g., the heatmap, the outer, middle, and inner linkages) except those at a specific position (starting at "position 1"). A simple Visual Basic program was written to generate a daughter

configuration file for each chromosome position by incrementing the position settings in the configuration file, resulting in 45 individual configuration files (one for each of the 44 positions and a base image with no greyed out values). A batch file ran Circos to generate a Circos plot using each of these configuration files individually, resulting in a total of 45 images being created. The Windows XP command-line program ImageMagick 6.8.7 was used to properly generate transparent backgrounds for these 45 images. An "overlay" image was created using Adobe Illustrator CS5 to display labels for brain regions and positions, and a centre legend for the heatmaps. Structural images for the "sidebar" were manually generated for each lipid species by obtaining appropriate high-resolution lipid structure files from VaLID (Visualization and Phospholipid Identification) [18]. These were processed in ChemDraw Ultra 12.0. Adobe Photoshop CS5 was used to construct the final composite sidebar images.

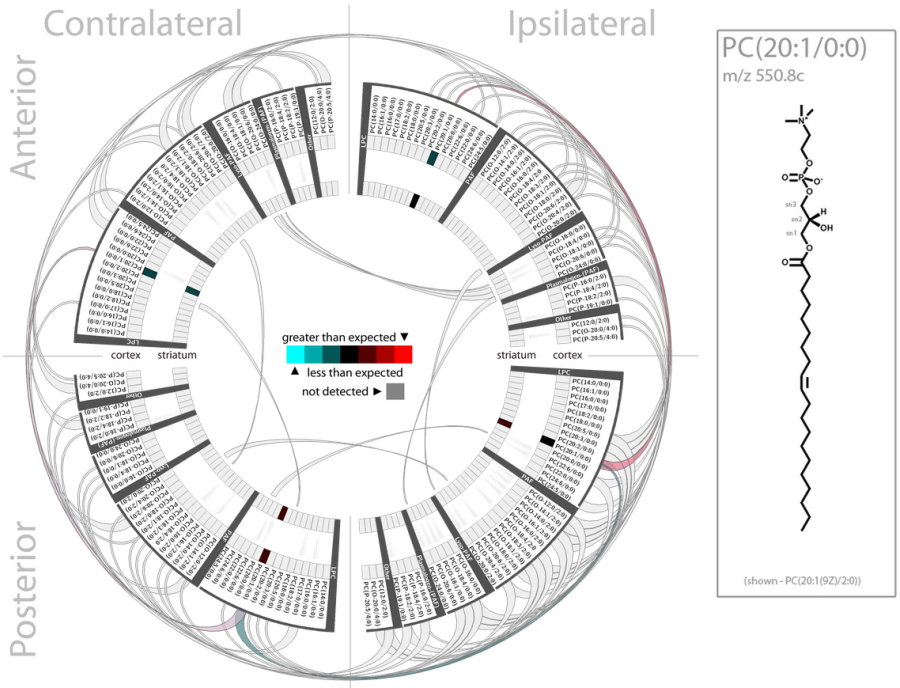


Figure 3. Highlighted lipid species PC(20:1/0:0). This view emphasizes the changes in this lipid species, and allows trends to be identified across the different comparisons being made. Information about this specific species is displayed in the sidebar on the right-hand side.

Circos was set to automatically output an HTML image map using the described URL tags for eah chromosome position. A Visual Basic program was written to replace these URL position tags in the HTML file with JavaScript to enable rollover function to display the individually created "greyed out" image by position. Thus, when a user performs a rollover on any aspect of a position (heatmap, link, label) the MouseRollover displays the image file for that position, along with the associated "sidebar" image for each position. The included MouseOut code displays the base

image when deselecting a position. The background of the webpage contains the "overlay" image, which is displayed underneath transparent dynamic Circos images.

2.5 Scalability and limitations

In this paper we present 44 different lipid species and their values with respect to three different binary variables. Thus, we display 352 unique lipid values using a Circos-based methodology. In mammalian cells, there may be several thousand unique phospholipid species. However, these species are organized into classes and subclasses. For example, we routinely identify 499 different species of *phosphocholines* (a subclass within the *glycerophospholipid* class) with different expression patterns in human and murine tissue and plasma. By focusing on a single subclass at a time, fewer species require concurrent analysis, and biologically important changes in the lipidome can still be investigated (e.g., pathway analysis). The Circos-based visualization approach we present here can be successfully scaled up to analyze entire subclasses, but may require some minor alterations with respect to specifications.

For example, using an artificial dataset, we investigated visualizing 190 lipids under the same visualization paradigm of three binary variables, which requires the display of 1520 unique lipid values. At this number the width of each individual heat map is decreased, causing reduced legibility of the plot. This legibility loss can be partially ameliorated, while maintaining the total information conveyed, by removing the lipid names from the Circos diagram (i.e., reducing the number of explicit specifications – lipid names will still be displayed in the sidebar image when a species of interest is highlighted). Unfortunately, we agree that the directionality of link changes is completely lost with this higher number of species. One solution, to maintain both the amount of information conveyed and the legibility of the plot, is to show an enlarged version of the selected lipid values in the centre of the plot. This is currently outside the scope of our visualization approach but could be implemented at a later date. An alternative solution is to reduce the number of binary variables displayed using two Circos diagrams, and displaying two diagrams on a single webpage/figure. Table 1 lists the estimated number of lipid species that can be effectively shown using each approach.

Table 1: Number of lipid species able to be visualized using a Circos-based approach. Shown in this paper are three binary variables without modification, highlighted in grey in the table.

Display paradigm	**# of binary variables**		
	1	2	3
Without modification (e.g., Fig 2)	176	88	44
Modified to increase visibility	760	380	190
Without modification, two separate Circos diagrams	352	176	88
Modified to increase visibility, two separate Circos diagrams	1520	760	380

3 Conclusion

In this study, we present a method of applying Circos style visualization to a medium-scale lipidomic dataset capable of dynamically and interactively conveying regional, metabolic, and disease-specific information simultaneously. Currently, the only time-consuming aspects to this procedure are the initial set up of the Circos plot, and the manual generation of structures in the sidebar image. Aside from the latter, the visualization approach taken here can be easily scaled up and will allow complex comparisons to be clearly visualized and informatively explored within large lipidomic datasets. Because conventional methods may miss fundamentally important trends hiding in these complex datasets, selecting an appropriate visualization method for such datasets is crucial. Dynamically generated Circos figures provide for both an integrative look and an intuitive method of exploring lipidomics.

4 Acknowledgements

This resource was funded by the Canadian Institutes of Health Research (CIHR) MOP 89999 and a Strategic Training Initiative in Health Research (STIHR) CIHR/ Training Program in Neurodegenerative Lipidomics (CTPNL) and Institute of Aging TGF 96121 to SALB. HX receives CIHR Training Program in Neurodegenerative Lipidomics (CTPNL)/Institute of Aging postdoctoral fellowship and a MITACS Accelerate postdoctoral fellowship. AS received studentships from the Heart and Stroke Foundation of Canada and CTPNL

5 References

1. Noorbakhsh, F., Overall, C.M., Power, C.: Deciphering complex mechanisms in neurodegenerative diseases: the advent of systems biology. Trends in neurosciences 32, 88-100 (2009)
2. Lausted, C., Lee, I., Zhou, Y., Qin, S., Sung, J., Price, N.D., Hood, L., Wang, K.: Systems Approach to Neurodegenerative Disease Biomarker Discovery. Annual review of pharmacology and toxicology (2013)
3. Niemela, P.S., Castillo, S., Sysi-Aho, M., Oresic, M.: Bioinformatics and computational methods for lipidomics. Journal of chromatography. B, Analytical technologies in the biomedical and life sciences 877, 2855-2862 (2009)
4. Reitz, C.: Dyslipidemia and the risk of Alzheimer's disease. Current atherosclerosis reports 15, 307 (2013)
5. Kang, J., Rivest, S.: Lipid metabolism and neuroinflammation in Alzheimer's disease: a role for liver X receptors. Endocrine reviews 33, 715-746 (2012)
6. Hayashi, H.: Lipid metabolism and glial lipoproteins in the central nervous system. Biological & pharmaceutical bulletin 34, 453-461 (2011)
7. Citron, M.: Alzheimer's disease: treatments in discovery and development. Nat Neurosci 5 Suppl, 1055-1057 (2002)

8. Wenk, M.R.: The emerging field of lipidomics. Nature reviews. Drug discovery 4, 594-610 (2005)
9. Piomelli, D., Astarita, G., Rapaka, R.: A neuroscientist's guide to lipidomics. Nat Rev Neurosci 8, 743-754 (2007)
10. Bennett, S.A.L., Valenzuela, N., Xu, H., Franko, B., Fai, S., Figeys, D.: Using neurolipidomics to identify phospholipid mediators of synaptic (dys)function in Alzheimer's Disease. Frontiers in physiology 4, 168 (2013)
11. van Meer, G.: Cellular lipidomics. The EMBO journal 24, 3159-3165 (2005)
12. Yetukuri, L., Ekroos, K., Vidal-Puig, A., Oresic, M.: Informatics and computational strategies for the study of lipids. Molecular bioSystems 4, 121-127 (2008)
13. Shevchenko, A., Simons, K.: Lipidomics: coming to grips with lipid diversity. Nature reviews. Molecular cell biology 11, 593-598 (2010)
14. Xu, H., Valenzuela, N., Fai, S., Figeys, D., Bennett, S.A.L.: Targeted lipidomics - advances in profiling lysophosphocholine and platelet-activating factor second messengers. The FEBS journal 280, 5652-5667 (2013)
15. Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A.: Circos: An information aesthetic for comparative genomics. Genome Research (2009)
16. Krzywinski, Martin I, http://circos.ca/intro/published_images/
17. Benjamini, Y., Hochberg, Y.: On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Ed Beh Stat 25, 60-83 (2000)
18. Blanchard, A.P., McDowell, G.S.V., Valenzuela, N., Xu, H.B., Gelbard, S., Bertrand, M., Slater, G.W., Figeys, D., Fai, S., Bennett, S.A.L.: Visualization and Phospholipid Identification (VaLID): online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass. Bioinformatics 29, 284-285 (2013)