metabolites

MDPI

Protocol

A Single Visualization Technique for Displaying Multiple Metabolite–Phenotype Associations

Mir Henglin [1,†], Teemu Niiranen [2,3,†], Jeramie D. Watrous [4], Kim A. Lagerborg [4], Joseph Antonelli [5], Brian L. Claggett [1], Emmanuella J. Demosthenes [1], Beatrice von Jeinsen [6], Olga Demler [7], Ramachandran S. Vasan [6,8], Martin G. Larson [6,8,9], Mohit Jain [4,*,†] and Susan Cheng [1,6,10,*,†]

1 Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
2 National Institute for Health and Welfare, FI 00271 Helsinki, Finland
3 Department of Medicine, Turku University Hospital and University of Turku, FI 20521 Turku, Finland
4 Departments of Medicine & Pharmacology, University of California San Diego, La Jolla, CA 92093, USA
5 Department of Statistics, University of Florida, Gainesville, FL 32611, USA
6 Framingham Heart Study, Framingham, MA 01701, USA
7 Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
8 Preventive Medicine, Department of Medicine, Boston University Medical Center, Boston, MA 02215, USA
9 Biostatistics Department, School of Public Health, Boston University, Boston, MA 02215, USA
10 Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
* Correspondence: mjain@ucsd.edu (M.J.); susan.cheng@cshs.org (S.C.); Tel.: +1-310-423-9680 (M.J.); +1-858-822-5854 (S.C.)
† These authors contributed equally to this work.

Received: 30 April 2019; Accepted: 28 June 2019; Published: 2 July 2019

Abstract: To assist with management and interpretation of human metabolomics data, which are rapidly increasing in quantity and complexity, we need better visualization tools. Using a dataset of several hundred metabolite measures profiled in a cohort of ~1500 individuals sampled from a population-based community study, we performed association analyses with eight demographic and clinical traits and outcomes. We compared frequently used existing graphical approaches with a novel 'rain plot' approach to display the results of these analyses. The 'rain plot' combines features of a raindrop plot and a conventional heatmap to convey results of multiple association analyses. A rain plot can simultaneously indicate effect size, directionality, and statistical significance of associations between metabolites and several traits. This approach enables visual comparison features of all metabolites examined with a given trait. The rain plot extends prior approaches and offers complementary information for data interpretation. Additional work is needed in data visualizations for metabolomics to assist investigators in the process of understanding and convey large-scale analysis results effectively, feasibly, and practically.

Keywords: metabolomics; visualizations; clinical outcomes research; epidemiology

1. Introduction

Advances in metabolomics technologies have enabled the generation of large-scale metabolomics measures in human studies [1]. Accordingly, newer generation visualization tools are needed to assist with the analyses and interpretation of these increasingly high-dimensional and complex data sets. Several resources now offer a variety of techniques for visualizing metabolomics data structure and exploring the inter-relations between individual and groups of metabolites [2–21].

The most commonly used methods for visually analyzing and representing associations between metabolites and outcomes are borrowed from conventional statistics and other biological fields [2]. Methods such as Manhattan plots, bar and scatter plots, and heatmaps are commonly used for visualizing information on the association between metabolites and outcomes. However, creating visualizations that can facilitate the interpretation of multi-level analyses, including information regarding associations among multiple metabolites and multiple outcomes, continues to pose a special challenge. We have therefore developed a visualization technique that expands upon existing approaches to enable the display of results from multiple simultaneous analyses relating metabolites and clinical phenotypes.

2. Methods

For development of these visualization methods, we used a metabolomics dataset comprising >500 bioactive lipids assayed by high resolution liquid chromatography-mass spectrometry (LC-MS) in a subset N = 1447 participants (age 66 ± 9 years, 54% women) of the Framingham Heart Study offspring cohort. LC-MS was performed according to previously described protocols [22]. In brief, plasma samples were prepared and analyzed using a Thermo Vanquish UPLC coupled to a high resolution Thermo QExactive orbitrap mass spectrometer. Metabolites were isolated from plasma using protein precipitation with organic solvent followed by solid phase extraction. Extracted metabolites were underwent chromatographic separation using reverse phase chromatography whereby samples were loaded onto a Phenomenex Kinetex C18 (1.7 µm, 2.1 × 100 mm) column and eluted using a 7 minute linear gradient starting with water:acetonitrile:acetic acid (70:30:0.1) and ending with acetonitrile:isopropanol:acetic acid (50:50:0.02). LC was coupled to a high resolution Orbitrap mass analyzer with electrospray ionization operating in negative ion mode, with full scan data acquisition across a mass range of 225 to 650 m/z. Thermo .raw data files were converted to 32-bit centroid .mzXML using Msconvert (Proteowizard software suite), and resulting .mzXML files were analyzed using Mzmine 2.21, as described [22]. For the present analyses, 16% of analytes had >10% missing values; we replaced missing values for metabolites with 0.25 x the minimum observed value for that metabolite, as reported previously [23]. Metabolite variables were then natural logarithmically transformed and standardized to facilitate cross-metabolite comparisons.

Given the biological importance of lipid metabolites with respect to cardiometabolic disease traits, we performed multivariable regression analyses to examine the relation of each metabolite with several clinical traits and outcomes, in the following order: age, female sex, body mass index, metabolic syndrome [24], prevalent diabetes [25], incident diabetes [25], Framingham Risk Score [26] as a measure of prevalent cardiovascular risk assessed at examination cycle 8, and incident hard cardiovascular disease [27] (Figure 1). We displayed the results of these relational analyses using a variety of techniques, including Manhattan plots (for one outcome at a time, with results ordered by mass-to-charge [m/z] value), bar and scatter plots (for one outcome at a time, with bars representing magnitude and directionality of estimates, and scatter dots representing P values), heatmaps (p's or beta's only), rain plot (beta's, p's, trends across a panel). We created heatmaps and rainplots with metabolites both unclustered and clustered based on hierarchical clustering. Details regarding the coding schema used to develop the rainplot approach and select specific parameters for the type of data displayed, for a given set of outcomes, are provided at: https://github.com/biodatacore/2017.09-rainplots. Detailed instructions and the R code for recreating the graphics in this paper is also provided at the site. All analyses and data visualizations were performed using R v3.4.1.

3. Results

All approaches to visualizing the results of association analyses demonstrated a range in the degree to which different metabolites were related to the different outcomes of interest. The extent and type of information conveyed varied across the visualization techniques, as summarized in Table 1 and detailed below.

Table 1. Dimension of information offered by different visualization methods.

Dimension of Information	Visualization Method			
	Manhattan Plot	**Bar and Scatter Plots**	**Heatmap**	**Rain Plot**
Example	Figure 1 and Figure S1	Figure S3	Figure 1 and Figure S2	Figure 2
Significance of associations with an outcome	X	X		X
Magnitude of associations with an outcome		X	X	X
Directionality of associations with an outcome	X	X	X	X
Clustering			X	X
Significance of associations with multiple outcomes	X		X	X
Magnitude of associations with multiple outcomes				X
Directionality of associations with multiple outcomes				X

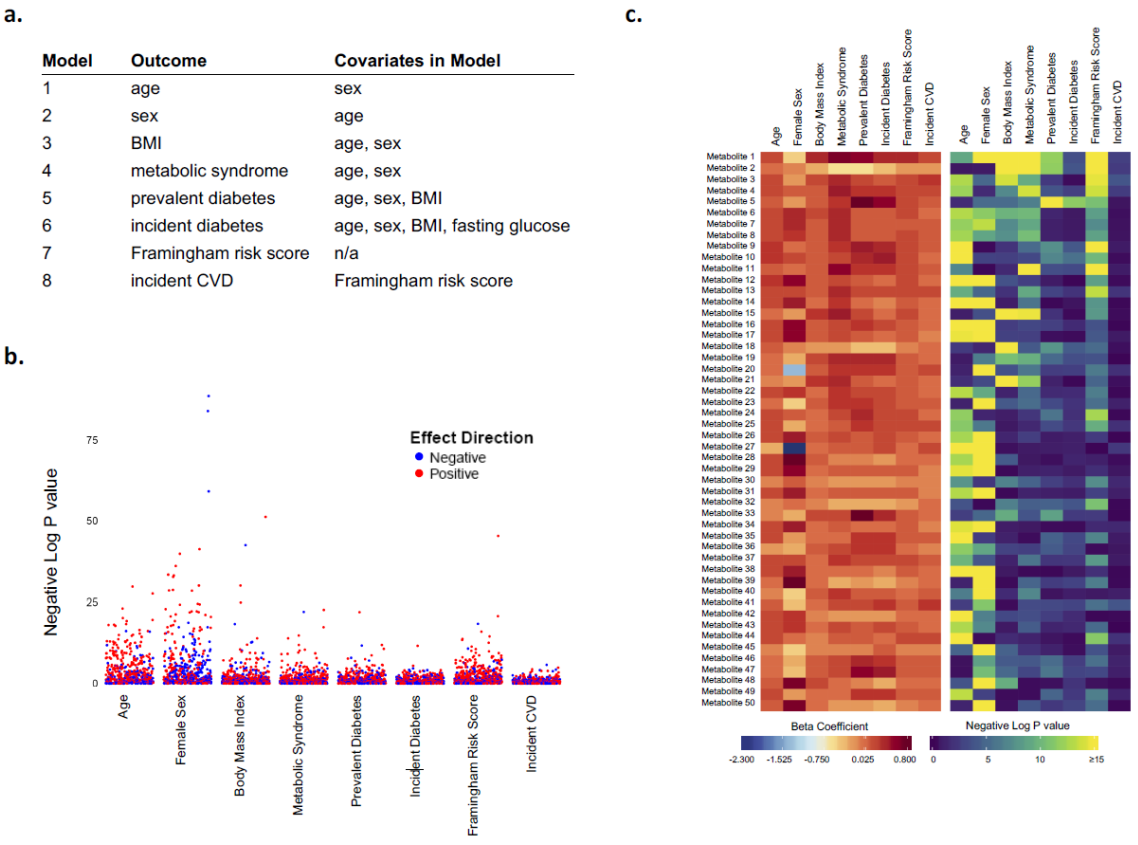

Figure 1. Visualization of complex metabolomics data. For a set of statistical models (**a**) performed in a large human study, for example, a Manhattan plot (**b**) can display the degree to which a wide panel of metabolites is associated with different outcomes although the magnitude of these associations is not conveyed. Pairing of heatmaps can display magnitude as well as directionality and significance for each metabolite association (**c**), although between-metabolite comparisons of associations across all outcomes is not easily discernible.

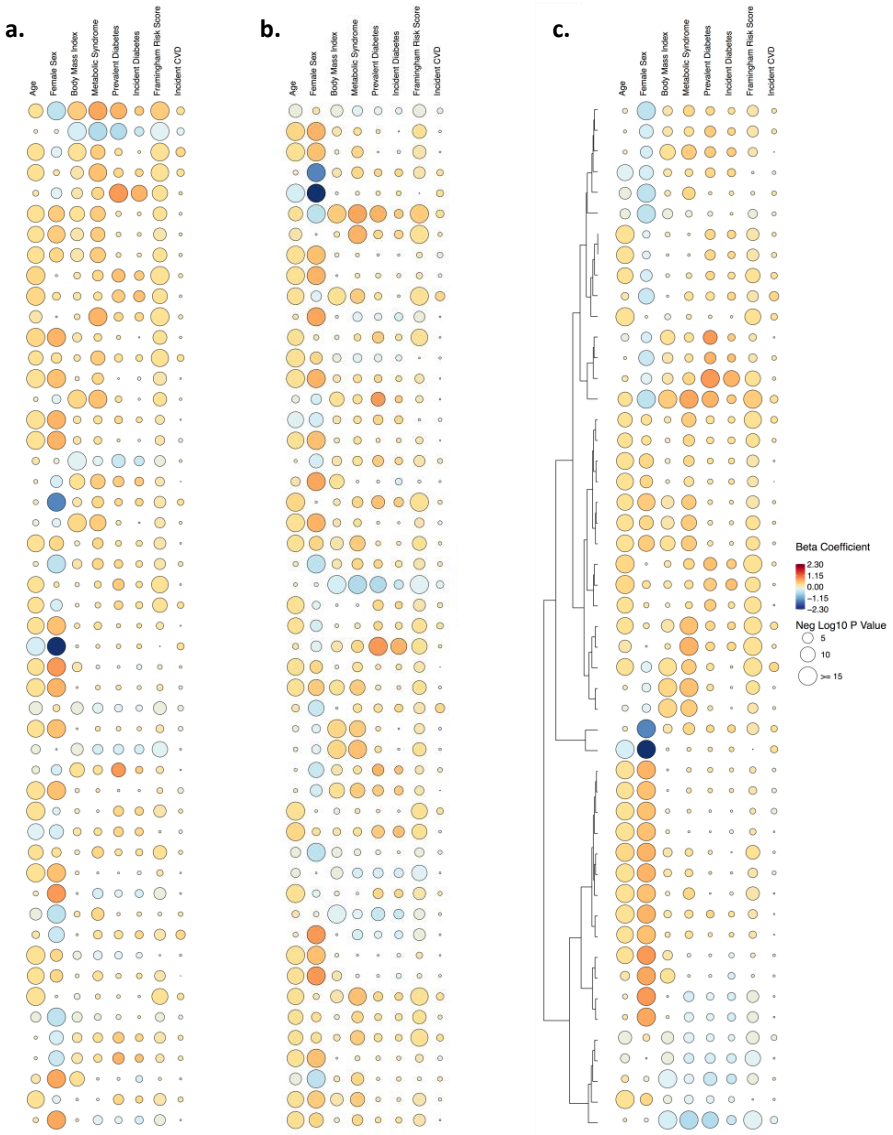

Figure 2. Rain plots. Results are ordered top to bottom by smallest to largest P value (**a**), mass-to-charge ratio (**b**), and by clustering (**c**).

Manhattan plots display P values for each model run and highlight statistically significant associations of all metabolites with each outcome across all the metabolites analyzed and ordered from left to right along the x axis by mass-to-charge ratio (*m*/*z*) (Figure 1 and Figure S1). The marked significance of metabolite associations with sex as well as BMI, compared with the other outcomes, is more clearly demonstrated when the results for all outcomes are displayed in a faceted plot with a shared y axis. Although colored dots can provide information regarding the directionality of associations, along with corresponding P value, the magnitude of each association is not easily conveyed. This issue can be addressed using parallel heatmaps, a commonly applied visualization approach for high-dimensional 'omics' data, wherein one plot depicts magnitude of effect and directionality (e.g., beta coefficients) while the other depicts corresponding statistical significance (e.g., P value) (Figure 1 and Figure S2). Although such an approach allows for global visualization of data, discerning clear patterns within or between metabolites can be difficult, especially across multiple heatmap plots and with increasing data set size. In turn, the bar and scatter plots display both the directionality and magnitude of associations, along with P values, for each outcome (Figure S3). The extent to which the same metabolite is positively or negatively associated with different outcomes is not as easily

discernible. However, the overall trend of generally more positive or more negative associations observed between a given outcome (e.g., age and Framingham Risk Score) and a large panel of metabolites is clearly displayed. Exceptions to such trends are also highlighted.

For studies that involve multiple staged experiments or statistical models, several options exist for visualizing analysis results. Our objective was to overcome limitations associated with discerning clear patterns within or between metabolites, especially across multiple plots and outcomes, and with increasing data set size. We therefore combined the visualization concepts offered by the conventional heatmap and previously reported raindrop plot methods; [28] the latter should not be mistaken for another similarly named method used to visualize collections of likelihoods and distributions [29]. As seen in the adapted 'rain plot' shown in Figure 2, directionality and magnitude of estimates for the top 50 metabolites associated with the selected outcomes are displayed using a color fill scale and the corresponding significance level is represented by size of the circle (i.e., rain 'droplet'). The metabolites can also be ordered top to bottom by smallest to largest P value, mass-to-charge ratio, and by clustering.

4. Discussion

A 'rain plot' approach combines the information from paired heatmaps into a single plot that emphasizes two types of information: (i) between-outcome comparisons, such as the extent to which most metabolites in this panel are associated with certain outcomes (e.g., age, sex) more than others; and, (ii) between-metabolite comparisons, such as the extent to which certain metabolites are associated with an aggregate measure of clinical cardiovascular disease risk (e.g., Framingham risk score) with or without concurrent relations to major component risk factors (e.g., diabetes risk for Metabolites 3 and 4, compared to for Metabolites 1 and 2). Between-metabolite comparisons, in particular, can facilitate identification of potentially important biological differences underlying the observed results of relating multiple metabolites to multiple phenotypes.

The rain plot visualization emphasizes two types of comparisons: (1) between-outcome results, and (2) between-metabolite results. For between-outcome comparisons, visually scanning for vertical patterns of large sized or deeply colored droplets (i.e., droplet 'streams') serves to highlight those outcomes that are the most broadly associated with a given panel of metabolites. As shown in Figure 2, this particular panel of bioactive lipids appears more globally associated with older age, sex, and Framingham Risk Score. A special feature of the rain plot is its emphasis on potentially important between-metabolite comparisons. For instance, certain metabolites (i.e., 20 and 27) are very strongly associated with sex (Figures 1 and 2). For each of these metabolite rows, a visual scan from left to right clarifies the relatively lower degrees of association for these metabolites with other outcomes of interest. The plot also visually clarifies interesting findings between the top-most prioritized metabolites. For instance, Metabolite 1 is positively associated with older age, male sex, and greater metabolic as well as cardiovascular disease risk. Conversely, Metabolite 2 is associated with lower metabolic risk, but is not significantly associated with either age or sex (Figure 2). The plot also highlights a finding for Metabolite 3 that distinguishes this analyte from Metabolites 1 and 2: while similarly associated with both greater Framingham Risk Score and risk for incident cardiovascular disease events, Metabolite 3 is not associated with prevalent or incident diabetes (Figure 2). In effect, Metabolite 3 appears associated with both prevalent and future cardiovascular risk through a biological pathway that is likely distinct from diabetes risk. Similar between-metabolite comparisons are possible across the entire plot.

As analytical chemistry methods continue to mature, resulting in larger and more complex metabolomics data, there is a growing need for ways to visually understand and interpret the relations of these high-dimensional data with multiple outcomes of interest. Using a dataset of metabolite measures performed in a population scale cohort, we compared several existing techniques that are commonly used for visualizing associations between metabolites and clinical outcomes. To improve these prior methods, we developed and demonstrate the potential utility of a rain plot approach to maximally render the multiple types of information that can be derived from the observed relationships between a panel of metabolites and a set of clinical traits and outcomes. We anticipate that this

approach may be further extended and applied to alternate study designs using different types of molecular phenotyping data—as part of the ongoing effort to effectively, efficiently, and feasibly convey the results of large-scale, high-dimensional data analyses [30,31].

Supplementary Materials: The following are available online at http://www.mdpi.com/2218-1989/9/7/128/s1, Figure S1: Manhattan plots; Figure S2: Parallel heat maps; Figure S3: Bar and scatter plots.

Author Contributions: Conceptualization, M.H., T.N., J.D.W., K.A.L., B.L.C., M.J. and S.C.; Funding acquisition, T.N., K.A.L., J.A., O.D., R.S.V., M.G.L., M.J. and S.C.; Investigation, M.H., T.N., J.D.W. and K.A.L.; Methodology, M.H., T.N., J.D.W., K.A.L., J.A., B.L.C., E.J.D., B.v.J., O.D., R.S.V., M.G.L. and S.C.; Resources, J.A., R.S.V. and M.G.L.; Supervision, B.L.C. and S.C.; Writing—original draft, M.H., T.N. and S.C.; Writing—review & editing, M.H., B.L.C., M.J. and S.C.

Funding: This project was supported in part by the National Institutes of Health grants T32-ES007142 (JA), K01-HL135342 (OD), K01-DK116917 (JDW), N01-HC-25195 (RSV, MGL), HHSN268201500001I (RSV, MGL), R01-HL134168 (SC, MJ), R01-HL143227 (SC, MJ), R01-ES027595 (MJ, SC), the Emil Aaltonen Foundation (TN), the Finnish Medical Foundation (TN), the Paavo Nurmi Foundation (TN), and the Frontiers of Innovation Scholars Program (KAL).

Conflicts of Interest: The authors declare no relevant conflicts of interest.

References

1. Misra, B.B.; van der Hooft, J.J. Updates in metabolomics tools and resources: 2014–2015. *Electrophoresis* **2016**, *37*, 86–110. [CrossRef] [PubMed]
2. Chia, P.L.; Gedye, C.; Boutros, P.C.; Wheatley-Price, P.; John, T. Current and Evolving Methods to Visualize Biological Data in Cancer Research. *J. Natl. Cancer Inst.* **2016**, *108*. [CrossRef] [PubMed]
3. Wang, R.; Perez-Riverol, Y.; Hermjakob, H.; Vizcaino, J.A. Open source libraries and frameworks for biological data visualisation: A guide for developers. *Proteomics* **2015**, *15*, 1356–1374. [CrossRef] [PubMed]
4. Sugimoto, M. Metabolomic pathway visualization tool outsourcing editing function. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2015**, *2015*, 7659–7662. [CrossRef] [PubMed]
5. Xia, J.; Sinelnikov, I.V.; Han, B.; Wishart, D.S. MetaboAnalyst 3.0–making metabolomics more meaningful. *Nucleic Acids Res.* **2015**, *43*, W251–W257. [CrossRef] [PubMed]
6. Clasquin, M.F.; Melamud, E.; Rabinowitz, J.D. LC-MS data processing with MAVEN: A metabolomic analysis and visualization engine. *Curr. Protoc. Bioinform.* **2012**, *14*. Unit14 11. [CrossRef]
7. Grapov, D.; Wanichthanarak, K.; Fiehn, O. MetaMapR: Pathway independent metabolomic network analysis incorporating unknowns. *Bioinformatics* **2015**, *31*, 2757–2760. [CrossRef]
8. Matheus, N.; Hansen, S.; Rozet, E.; Peixoto, P.; Maquoi, E.; Lambert, V.; Noel, A.; Frederich, M.; Mottet, D.; de Tullio, P. An easy, convenient cell and tissue extraction protocol for nuclear magnetic resonance metabolomics. *Phytochem. Anal.* **2014**, *25*, 342–349. [CrossRef]
9. Grace, S.C.; Embry, S.; Luo, H. Haystack, a web-based tool for metabolomics research. *BMC Bioinform.* **2014**, *15* (Suppl. 11), S12. [CrossRef]
10. Eichner, J.; Rosenbaum, L.; Wrzodek, C.; Haring, H.U.; Zell, A.; Lehmann, R. Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **2014**, *966*, 77–82. [CrossRef]
11. Xia, J.; Wishart, D.S. Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. *Curr. Protoc. Bioinform.* **2011**, *14*. Unit 14 10. [CrossRef] [PubMed]
12. Chagoyen, M.; Pazos, F. Tools for the functional interpretation of metabolomic experiments. *Brief Bioinform.* **2013**, *14*, 737–744. [CrossRef] [PubMed]
13. Mak, T.D.; Laiakis, E.C.; Goudarzi, M.; Fornace, A.J., Jr. MetaboLyzer: A novel statistical workflow for analyzing Postprocessed LC-MS metabolomics data. *Anal. Chem.* **2014**, *86*, 506–513. [CrossRef] [PubMed]
14. Kuo, T.C.; Tian, T.F.; Tseng, Y.J. 3Omics: A web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst. Biol.* **2013**, *7*, 64. [CrossRef] [PubMed]
15. Karnovsky, A.; Weymouth, T.; Hull, T.; Tarcea, V.G.; Scardoni, G.; Laudanna, C.; Sartor, M.A.; Stringer, K.A.; Jagadish, H.V.; Burant, C.; et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics* **2012**, *28*, 373–380. [CrossRef]

16. Cheng, S.; Mueller, K. The Data Context Map: Fusing Data and Attributes into a Unified Display. *IEEE Trans Vis Comput. Graph* **2016**, *22*, 121–130. [CrossRef] [PubMed]
17. Cheng, S.; Xu, W.; Mueller, K. ColorMap(ND): A Data-Driven Approach and Tool for Mapping Multivariate Data to Color. *IEEE Trans Vis Comput. Graph* **2019**, *25*, 1361–1377. [CrossRef]
18. Cheng, S.; Zhong, W.; Isaacs, K.E.; Mueller, K. Visualizing the Topology and Data Traffic of Multi-Dimensional Torus Interconnect Networks. *IEEE Access* **2018**, *6*, 57191–57204. [CrossRef]
19. Shenghui, C.; Mueller, K. Improving the fidelity of contextual data layouts using a Generalized Barycentric Coordinates framework. In Proceedings of the 2015 IEEE Pacific Visualization Symposium (PacificVis), Hangzhou, China, 14–17 April 2015; pp. 295–302.
20. DaGoo. Available online: http://www.dagoo.work (accessed on 13 June 2019).
21. Johnson, C.H.; Ivanisevic, J.; Siuzdak, G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 451–459. [CrossRef]
22. Watrous, J.D.; Henglin, M.; Claggett, B.; Lehmann, K.A.; Larson, M.G.; Cheng, S.; Jain, M. Visualization, Quantification, and Alignment of Spectral Drift in Population Scale Untargeted Metabolomics Data. *Anal. Chem.* **2017**, *89*, 1399–1404. [CrossRef]
23. Huan, T.; Li, L. Counting missing values in a metabolite-intensity data set for measuring the analytical performance of a metabolomics platform. *Anal Chem.* **2015**, *87*, 1306–1313. [CrossRef] [PubMed]
24. Grundy, S.M.; Cleeman, J.I.; Daniels, S.R.; Donato, K.A.; Eckel, R.H.; Franklin, B.A.; Gordon, D.J.; Krauss, R.M.; Savage, P.J.; Smith, S.C., Jr.; et al. Diagnosis and management of the metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. *Circulation* **2005**, *112*, 2735–2752. [CrossRef] [PubMed]
25. Abraham, T.M.; Pencina, K.M.; Pencina, M.J.; Fox, C.S. Trends in Diabetes Incidence: The Framingham Heart Study. *Diabetes Care* **2015**, *38*, 482–487. [CrossRef] [PubMed]
26. Wilson, P.W.; D'Agostino, R.B.; Levy, D.; Belanger, A.M.; Silbershatz, H.; Kannel, W.B. Prediction of coronary heart disease using risk factor categories. *Circulation* **1998**, *97*, 1837–1847. [CrossRef] [PubMed]
27. Niiranen, T.J.; Lyass, A.; Larson, M.G.; Hamburg, N.M.; Benjamin, E.J.; Mitchell, G.F.; Vasan, R.S. Prevalence, Correlates, and Prognosis of Healthy Vascular Aging in a Western Community-Dwelling Cohort. *Framingham Heart Study* **2017**, *70*, 267–274. [CrossRef] [PubMed]
28. Raindrop Plot. Available online: http://mc-3.ca/raindrop-plot (accessed on 1 July 2019).
29. Barrowman, N.J.; Myers, R.A. Raindrop Plots. *Am. Stat.* **2003**, *57*, 268–274. [CrossRef]
30. Gehlenborg, N.; O'Donoghue, S.I.; Baliga, N.S.; Goesmann, A.; Hibbs, M.A.; Kitano, H.; Kohlbacher, O.; Neuweger, H.; Schneider, R.; Tenenbaum, D.; et al. Visualization of omics data for systems biology. *Nat. Methods* **2010**, *7*, S56–S68. [CrossRef] [PubMed]
31. O'Donoghue, S.I.; Gavin, A.C.; Gehlenborg, N.; Goodsell, D.S.; Heriche, J.K.; Nielsen, C.B.; North, C.; Olson, A.J.; Procter, J.B.; Shattuck, D.W.; et al. Visualizing biological data-now and in the future. *Nat. Methods* **2010**, *7*, S2–S4. [CrossRef] [PubMed]