Abstract: A Method for Systematically Surveying Data Visualizations in Infectious Disease Genomic Epidemiology

Background. Stakeholders within public health can use the results of genomic analyses to establish practice guidelines and enact policies. Yet, these stakeholders vary in their abilities to interpret genomic findings and contextualize the results with other sources of data. Data visualization is an emergent solution to address interpretability challenges, but absent is a systematic and robust method to help identify the appropriate visualization to use in different contexts. Methods. We have developed a systematic method for generating an explorable visualization design space, which catalogues visualizations existing within the infectious disease genomic epidemiology literature. Our method uses an automated literature analysis phase to establish why data were visualized, followed by a manual visualization analysis phase to establish what data were visualized and how. The literature analysis phase queried PubMed and used an unsupervised cluster analysis on article titles and abstracts to discover topic clusters that suggested why data were visualized. In order to ensure that we had a variety of data visualizations for further analysis, we sampled articles from across topic clusters and then extracted their figures. We then applied open and axial coding techniques, from qualitative research methods, to the sampled figures in order to iteratively derive taxonomic codes that described elements of each data visualization, thus enabling us to compare visualizations. Results. We applied our method to a document corpus of approximately 18,000 articles, from which we sampled 204 articles for analysis. We added 17 articles manually for a final 221 articles that yielded 801 figures and 49 missed opportunity tables. These figures served as inputs to the visualization analysis phase and resulted in taxonomic codes along three descriptive axes of visualization design: chart types within the visualization, chart combinations, and chart enhancements. We refer to the collective complement of derived taxonomic codes as GEViT (Genomic Epidemiology Visualization Typology). To operationalize GEViT and the results of the literature analysis we have created a browsable image gallery (http://gevit.net), that allows an individual to explore the myriad of complex types of data visualizations (i.e. the visualization design space). Our analysis of the visualization design space through GEViT also revealed a number of data visualization challenges within infectious disease genomic epidemiology that future bioinformatics work should address. Conclusions. Generating explorable visualization design spaces can help stakeholders and bioinformaticians design and evaluate data visualizations for different contexts. By consistently codifying a visualization’s elements it also becomes possible to meaningfully test visualization efficacy beyond an individual’s intuition and preferences.