A method for systematically surveying data visualizations in infectious disease genomic epidemiology

Anamaria Crisan[1], Jennifer L. Gardy[2,3], Tamara Munzner[1,*]

[1] Department of Computer Science, University of British Columbia, Vancouver, British Columbia, CANADA
[2] School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, CANADA
[3] British Columbia Centre for Disease Control, Vancouver, British Columbia, CANADA

Correspondence should be addressed to TM (tmm@cs.ubc.ca)

Abstract

Data visualization is an important tool for exploring and communicating findings from genomic and health datasets. Yet, without a systematic way of understanding the design space of data visualizations, researchers do not have a clear sense of what kind of visualizations are possible, or how to distinguish between good and bad options. We have devised an approach using both literature mining and human-in-the-loop analysis to construct a visualization design space from corpus of scientific research papers. We ascertain *why* and *what* visualizations were created, and *how* they are constructed. We applied our approach to derive a Genomic Epidemiology Visualization Typology (GEViT) and operationalized our results to produce an explorable gallery of the visualization design space containing hundreds of categorized visualizations. We are the first to take such a systematic approach to visualization analysis, which can be applied by future visualization tool developers to areas that extend beyond genomic epidemiology.

Introduction

Cheaper and more accurate genomic sequencing technologies are enabling public health decision makers, from doctors to epidemiologists to researchers to policy makers, to make more informed, near real-time, data-driven decisions toward pathogen diagnosis[1], routine surveillance[2,3], and public health interventions[4]. Yet as pathogen genomic data become more ubiquitous and are combined with other sources of routinely collected public health data, analysts and decision-makers are forced to confront the dimensionality challenges that attend such “big data”, with interpretability of results being chief amongst them.

Data visualization is an emergent solution to address interpretability challenges. It has been shown to improve comprehension of numerical results in medical risk communication[5,6], but that context is much less complex than the heterogeneous datasets used in modern genomic epidemiology, which can include, amongst other things, genomic, patient, clinical, epidemiological, and geographic data elements. While the rise of public health genomics has been met with concrete efforts to visualize ‘omics data[7], including Nextstrain[8] and Microreact[9], few of these visualizations have been tested with target end-users to assess a visualization’s utility and usability in decision-making contexts[10]. What is absent is a notion of a visualization design space – the combinatorial space of visualizations that can be produced using basic graphical primitives (points, lines, areas) and aesthetic properties (position, color, size, and so on) to depict input data – and a way to systematically construct and analyze this design space to inform the design and evaluation of public health genomic data visualizations.

Design spaces are common in number of disciplines, ranging from architecture to computer science, but are absent in bioinformatics research, resulting in missed opportunities. Visualization design spaces could arguably be inferred from the byproducts of search engines such as Google Image Search or PubMed Search, or more complex scholarly literature analysis tools such as Semantic Scholar and SourceData[11]. However, the construction and exploration of a design space from these search results would require extensive additional intellectual investment. Other more explicit attempts to describe a design space exist in the form of web galleries such as SetVis[12],TreeVis[13], Visualizing Health(http://www.vizhealth.org/), or BioVis Explorer[14], but while these are closer to the spirit of our definition of a design space they lack the systematicity of ours and are limited to specific subsets of possible visualizations designs. Thus, there remains the need to enable researchers, bioinformaticians, and other software tool developers to generate broad and explorable visualization design spaces.

Here we propose a systematic approach to constructing a data visualization design space by analyzing figures from the existing public health genomic research literature. Our human-in-the-loop approach blends automated algorithmic with manual curation steps that inject contextual knowledge into the design space construction process. Our approach specifically aims to systematically construct a design space that incorporates information about *why* researchers visualize data, *what* visualizations they use and *how* those visualizations are constructed, and finally to understand *how many* examples of specific data visualizations there are in our dataset. We demonstrate a concrete instantiation of this approach for a specific use case through the generation of a Genomic Epidemiology Visualization Typology (GEViT). We also provide a browsable gallery of categorized visualizations that supports exploration of the GEViT visualization design space. Our findings from GEViT itself have the most direct implications for microbial genomic research, but our approach can be applied more generally to other disciplines. We demonstrate that rigor is both desirable and achievable in data visualization design and evaluation.

Results

Our results are divided into two sections, a literature analysis and a visualization analysis. The purpose of the literature analysis was to derive an underlying structure of the document corpus in order to intelligently sample a variety of visualizations. The visualization analysis portion describes the construction of GEViT using iterative open and axial coding techniques and a descriptive quantitative analysis of the visualizations based upon GEViT. That analysis makes use of the visualization theory and terminology succinctly summarized in co-author Munzner's textbook[15]. A detailed overview of our methodology is provided in the Online Methods, and Supplementary Figures S1, S2, and S3. Additionally, we provide all analysis notebooks and datasets online at: `https://github.com/amcrisan/gevitAnalysisRelease`

LITERATURE MININ

Literature mining identified article clusters according to disease pathogen

We assembled a document corpus of 17,974 articles pertaining to infectious disease genomic epidemiology research published in the past 10 years (Figure 1). Using article titles and abstracts we derived topic clusters in an unsupervised manner, and classified articles as either belonging to a named topic cluster, not belonging to a cluster under current parameter settings, or never being clustered under any parameter settings (Figure 2a, also see Online Methods). Articles that never formed part of a cluster were removed from further analysis, leaving 15,315 documents of which 11,416 (75% of the initial document corpus) formed 32 topic clusters (Figure 2b). Clusters were assigned topics via the top two most frequent terms within the cluster, revealing that infectious disease genomic epidemiology literature is primarily structured around pathogens. We validated our results by comparing our automatically derived cluster naming to the distribution of pathogen terms from an external list (Table S1, Figure 2c), and found there to be a strong correspondence between the automatically derived cluster topics and the propensity for pathogen terms to appear within clusters of the same name (for example, the term "*Influenza Virus*" occurs primarily within the "influenza-viru" cluster). Some notable exceptions are *Escherichia coli*, *Helicobacter pylori*, and *Human Immunodeficiency Virus*, which spread across more clusters in addition to having their own defined cluster; they frequently co-occur with other infections. We also found that clusters with more generic names (for example "viru-sequenc", or "geno-sequenc") contain pathogens that likely had too few articles to form their clusters, possibly because they are part of more recent outbreaks (i.e., Zika, Ebola), while pathogens that tend to be more consistently studied (i.e. *Mycobacterium tuberculosis*, *Influenza Virus*) and hence have more articles tend to form their own clusters. While t-SNE based results (see online methods) should be interpreted cautiously with respect to proximity and cluster density, we found the trends in the literature analysis were well matched to domain knowledge. We filtered the corpus by limiting to pathogens with 40 or more articles, resulting in 6,350 articles within 35 pathogen clusters, then further simplified to 18 clusters: a final set of 17 pathogen clusters that had 100 or more documents and one "other" cluster.

Linking pathogens to *a priori* concepts

The findings from the literature mining were at odds with our own *a priori* assumptions that articles would cluster according to more general concepts, for example drug resistance, surveillance, outbreak responses, and so on, which cross-cut all pathogens. We chose to link the data-driven pathogen clusters to these *a priori* concepts because we envision this taxonomy being used by people specifically interested in them. We did so by analyzing bigrams that occurred within and between pathogen topic clusters, and manually annotating those bigrams to map to some *a priori* concept; for example, the bigram “vancomycin resistance” was mapped to concept of “drug resistance” (Table S2). We mapped a total of 23 *a priori* concepts to 404 bigrams, categorized into three groups: genomic concepts (drug resistance, genome, genotype, molecular biology, pathogen characterization, phylogeny, and population diversity); epidemiology concepts (clusters, disease reservoirs, geography, outbreaks (international, community, hospital), surveillance, transmission, vaccine, and vectors), and medical concepts (clinical, cancer, diagnosis, outcome, and treatment). Some bigrams were not mapped to *a priori concepts*, often because they were standard technical writing phrases (e.g. “statistically significant”, “data show”). *A priori* concepts did not occur uniformly across pathogen clusters (Figure S4A) and a variable number of bigrams mapped to individual *a priori* concepts, with 143 bigrams mapped to “drug resistance” and only one bigram mapped to “disease reservoirs” and topic clusters (Figure S4B).

Document sampling was stratified according to pathogen and *a priori* concepts

We then performed two rounds of stratified sampling using pathogens and *a priori* concepts as strata. The sampling resulted in 204 unique articles to which we manually added 17 additional articles that we deemed contained interesting data visualizations (these are clearly tagged in our analysis), for a total of 221 articles (Table S3) from which we extracted a total of 770 figures, including a small number (45) of ‘missed opportunity’ tables.

VISUALIZATION ANALYSIS

Developing GEViT – A Genomic Epidemiology Visualization Typology

Using the analysis set of harvested figures, we used iterative open and axial coding techniques to devise a systematic way to describe how data visualizations are constructed. For analysis, we used whole figures and **did not** split them up into smaller parts. We began by classifying the types of charts in figures, further evolving to also classifying how charts were combined, and finally we also classified how charts were enhanced. We found that these three descriptive axes allowed us to sufficiently describe all visualizations in our dataset (see Online Methods for detailed sufficiency criteria). For each of these descriptive axes we also derived a controlled vocabulary (taxonomy). Collectively, we refer to this result of the descriptive axes and their associated taxonomies as GEViT (Genomic Epidemiology Visualization Typology). Below, we describe each of GEViT's descriptive axes and interleave descriptive statistics to show the distribution of taxonomic codes across these axes to provide an overview of the visualization design space. We also operationalized our analysis to produce a browsable gallery (https://gevit.net) that allows others to explore this GEViT design space through the classified figures (including their captions), where each figure is linked back to the original PubMed articles.

Chart Types in GEViT. We identified seven classes of chart types that form the basis of the data visualizations in our dataset (Figure 3): Common Statistical; Area; Relational; Temporal; Spatial; Tree; and Genomic. We compiled a taxonomy of common chart names to classify specific instances of chart types with each class. When applicable, we also defined special cases of a specific chart; for example, epidemic curves are a special case of bar chart. We also defined one 'Other' category, which included entities that accompanied data visualizations but were not themselves data visualizations, such as tables and images, and miscellaneous visualizations that did not fit elsewhere. In total we observed 23 distinct chart types (plus one miscellaneous category), and found that the most commonly occurring types within data visualizations included Phylogenetic Trees (17.7% of all data visualizations, although some type of tree was present in 23.7% of all visualizations), followed by Tables (9.7%), Bar Charts (8.9%), Genomic Maps (6.9%), Line Charts(6.8%), and Images (5.7%, typically a Gel Image of Pulsed Field Gel Electrophoresis). See Figure S5 for the occurrence of all chart types. The pervasive presence of tables, either alone or in combination with some other chart types, is a notable finding since it indicates missed opportunities for visualization.

Chart Combinations in GEViT. Although the majority of figures were composed of a single chart type (40.1,%), there were distinct and common patterns of combining chart types to create more complex, and often linked, multi-part figures (Figure 4). Composite charts (20.3%) contained multiple chart types that were spatially aligned – for example, a heatmap and tree (dendrogram) that are spatially aligned to indicate both a hierarchical clustering and the underlying data for the clustering. A tree and heatmap can also be visualized independently of each other, but their combined value is evidently relevant for many researchers. Small Multiples (17.3%) showed different aspects of the data through multiple instances of the same chart type. Many Types Linked combinations (13.5%) used multiple different chart types that were visually linked, for example using a common color to denote some property of the data across the different charts, but not spatially aligned (in contrast to Composite charts). Finally, Many Types

General combinations (8.8%) describe a data visualization in which there are multiple chart types, and there does not appear to be any sort of spatial or visual link between them. This situation often arises when authors put many unrelated charts into a single figure due to space restrictions. It was not always straightforward to distinguish between some instances of Many Types Linked and Many Types General, and in such cases we resolved the ambiguity in favor of the latter classification. We also observed instances of Complex Combinations (11.9%) that developed data visualizations using two of the previously describes types of chart combinations. It was notable that trees were mostly commonly combined with other chart types.

Chart Enhancements in GEViT. Lastly, we noted that standard chart types were often enhanced to add metadata through the addition or changing of graphical marks - the basic graphical element corresponding to a data record (*e.g.* a patient), or derived data value (*e.g.* the total number of patients). Basic marks are points, lines, areas, and (perhaps surprisingly) text, which are endowed with aesthetic properties of size, shape, color, and texture that can be modified to encode data (Figure 5a). For example, a phylogenetic tree encodes evolutionary relationships inferred from DNA data (among other sources) as lines of some calculated length that are precisely positioned in space (Figure 5b). By default, the lines of a phylogenetic tree are often black, however those lines can be *re-encoded* to incorporate data from some additional source – for example, coloring lines according to geographic regions. Instead of re-encoding a mark, it is also possible to *add marks* to the base chart type, for example, adding colored point marks to a tree's leaf positions (Figure 5b), or to add linear brackets and text to delineate groups (the most common reason text and lines with bracket shapes are used in our corpus). We did not consider axis text, titles, or data labels to be added marks, subsuming them as constituent parts of the base chart type.

It is also possible to add more complex types of marks, which are specific instances of the basic marks types presented in Figure 5a. Connection marks are a specific instance of line marks that *connect* two other marks. Containment marks are a specific instance of area marks that enclose other marks. Finally, a glyph is a complex mark that could itself be a type of chart, but that is smaller than the base chart type and embedded within it (in contrast, we define that composite chart types have the same frame size and one chart is not embedded within the other). The only glyph we identified within our dataset was a pie chart, which was often added to geographic maps or node-link graphs (Figure 5b) to denote proportion variability in the data.

We differentiate between the instances when chart enhancements are added consistently, or just as one-off marks. When the addition or re-encoding of marks is applied consistently to the base chart type, for example re-encoding all or many lines in a tree, or adding points to all or many leaf nodes, we defined these as structured enhancements. Adding one-off marks, even if they are driven by the data or the addition of some arbitrary ink, was considered to be an annotation and defined as an unstructured enhancement. It was not always easy to differentiate between structured and unstructured enhancements, and in such cases we resolved ambiguities by choosing structured enhancement when analyzing figures.

In our dataset we observed that most figures were enhanced (83.8% of all chart types), typically through the addition of lines, points, or text (59.6%) while re-encoding of marks was less common (45.6%). The use of text as a graphical mark with aesthetic properties that can be manipulated to convey information was common in our dataset, either by adding text marks to a base chart type, or re-encoding of text labels by manipulating the font face. The text itself ranged from the very simple case of a single letter or number, to a full word, to a complex concatenated string of metadata such as specimen ID, location, and year. Annotations were also less common (33.6%), and were most commonly an arrow to text, or a containment mark that highlighted only a single group.

Discussion

Data visualization is an increasingly important analytic tool for exploring and communicating results from large genomic and health datasets, but efforts to harness its potential power are impeded when visualization creators make *ad hoc* choices rather than systematically consider visualization design alternatives. While we found some instances of quite impressive and well thought out data visualizations, the systematic nature of our GEViT design space construction allowed us to assess the considerable variability of visualization design quality and revealed the unexplored potential within the design space. GEViT presents a higher level of abstraction than the existing grammar of graphics proposed by Wilkinson[16] and famously instantiated by Wickham[17] in the R tidyverse, yet is developed in the same spirit of standardizing, generalizing, and simplifying the construction of data visualizations from individual components. We found this high level of abstraction to be useful for exploring design spaces, while lower level abstractions are needed for implementation. Software tools designed with awareness of the visualization design space for genomic epidemiology could better support figure creators to make reasoned and informed choices and to avoid the *ad hoc* random walk through the set of possibilities. Compared to the robust and systematic use of statistical techniques in genomic epidemiology, there is far to go before genomic epidemiology data visualization becomes truly mature.

Delineating a design space, as we have done through GEViT, is just a first step; the obvious next step is to provide robust guidance on good or bad practice in a way that is more targeted to the genomic epidemiology than the existing general visualization literature. Even this first step of establishing the design space shows gaps that require attention and provides design alternatives against which future researchers and practitioners could test and calibrate any new solutions. We emphasize the importance of using empirical studies of visualizations, with multiple design alternatives, in order to triangulate optimal design patterns for different contexts and tasks.

Two notable findings pertain to missed opportunities involving text: the pervasive use of tables (often combined with other chart types) where visualization could have been used but was not, and the practice of encoding information with aesthetic properties such as color and size applied to long text string labels. The visualization literature discourages the use of text as a mark type because reading text imposes cognitive load, whereas the goal of using aesthetic properties to encode information is to support purely perceptual processing[15]. We suspect that the widespread use of text marks in this hybrid way stems from an incomplete knowledge of the design space and the lack of tools to support the visualization of complex and heterogenous data.
Showing raw data through text also compounds another notable tendency of these visualizations to show all data records, which limits their scalability. An under-explored alternative would be to visually summarize the data at multiple levels of detail. Another finding was the pervasiveness of phylogenetic trees. Although few researchers in genomic epidemiology would consider this

finding surprising, we note that our own prior work suggested that phylogenetic tree visualizations have unclear utility for clinical and public health stakeholders[18]. Perhaps the convention of showing them routinely in a genomics research context has prevented the community from seeing the forest for the trees, so to speak. Further innovation in visualization design may result in different default choices.

We have presented an approach to systematically develop an explorable visualization design space through a human-in-the-analysis-loop model that exploits the strengths of both automatic processing for speed and low effort, and manual curation where human judgment is harnessed to integrate data-driven insights with human expertise. The exploratory rather than confirmatory nature of our study is both its strength and its primary limitation. While we have made all of our intermediate analysis outputs available in the spirit of transparency, the qualitative manual analysis phase are unlikely to yield identical results if undertaken by a different researcher. Although our approach will surely benefit from ongoing innovations in image recognition, machine learning, and natural language processing, we argue that attempting to fully automate the entire process would be premature. Developing a faster process that still provides a way to include a human in the analysis loop will be fruitful future work for us.

There are many other ways that our resulting design space could be explored, and for brevity we have only touched upon a few selected findings. Nevertheless, these results have allowed us to appreciate the expressiveness of visualization designs in infectious disease genomic epidemiology. Our results provide guidance to both software tool developers, including bioinformaticians, and to researchers engaged with creating their own visualizations: we provide a concrete terminology for describing data visualizations, and a source of inspiration through the exploration of a design space. Most importantly, our work demonstrates that it is possible to think systematically and rigorously about data visualizations and that there exist open, complex, interesting, and impactful problems in visualization design and analysis.

Online Methods

See Online Methods Document

Acknowledgements

The authors appreciate discussions with the InfoVis Group at the University of British Columbia and the Bedford lab at the Fred Hutch Institute. The authors would also like to acknowledge their funding sources. AC is supported by a CIHR Vanier Scholarship, JG by the Canada Research Chairs Program and a Michael Smith Foundation for Health Research Scholar award, and TM by the NSERC Discovery Program.

Author Contributions

AC, JG, and TM devised and interpreted the analysis and jointly wrote the paper.

Competing Interests Statements

The authors declare no competing interests.

References

1. Pankhurst, L. J. *et al.* Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: A prospective study. *Lancet Respir. Med.* **4,** 49–58 (2016).
2. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8,** 97 (2016).
3. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–32 (2016).
4. Bradley, P. *et al.* Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat. Commun.* **6,** 10063 (2015).
5. Zipkin, D. A. *et al.* Evidence-based risk communication: A systematic review. *Annals of Internal Medicine* **161,** 270–280 (2014).
6. Ancker, J. S. & Kaufman, D. Rethinking Health Numeracy: A Multidisciplinary Literature Review. *J. Am. Med. Informatics Assoc.* **14,** 713–721 (2007).
7. Gehlenborg, N. *et al.* Visualization of omics data for systems biology. *Nat. Methods* **7,**

S56--68 (2010).

8. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *bioRxiv* (2017).
9. Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* (2016). doi:10.1099/mgen.0.000093
10. Carroll, L. N. *et al.* Visualization and analytics tools for infectious disease epidemiology: A systematic review. *J. Biomed. Inform.* **51,** 287–298 (2014).
11. Liechti, R. *et al.* SourceData: A semantic platform for curating and searching figures. *Nature Methods* **14,** 1021–1022 (2017).
12. Alsallakh, B. *et al.* Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges. in *Eurographics conference on Visualization (EuroVis)– State of The Art Reports* 1–21 (2014). doi:10.2312/eurovisstar.20141170
13. Schulz, H. J. Treevis.net: A tree visualization reference. *IEEE Comput. Graph. Appl.* **31,** 11–15 (2011).
14. Kerren, A., Kucher, K., Li, Y.-F. & Schreiber, F. BioVis Explorer: A visual guide for biological data visualization techniques. *PLoS One* **12,** e0187341 (2017).
15. Munzner, T. *Visualization Analysis and Design*. (CRC Press, 2014).
16. Wilkinson, L. The grammar of graphics. *Wiley Interdisciplinary Reviews: Computational Statistics* **2,** 673–677 (2010).
17. Wickham, H. A layered grammar of graphics. *J. Comput. Graph. Stat.* **19,** 3–28 (2010).
18. Crisan, A., McKee, G., Munzner, T. & Gardy, J. L. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ* **2018,** (2018).
19. Meirelles, I. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. (Rockport Publishers, 2013).
20. Bertin, J. *Semiology of graphics: diagrams, networks, maps*. *Components* (1983). doi:10.1037/023518

FIGURE LEGENDS

Figure 1 Summary of literature analysis steps and document sampling.

Figure 2 Summary of literature analysis results. a) Documents were classified according to whether they were part of a cluster (green), unclustered under current parameter settings (purple), or never formed part of cluster (orange). The 32 cluster boundaries were automatically determined and are shown as light grey ovals. **b)** Clustered documents and their topics, which are automatically assigned based upon top two terms with the cluster. **c)** Verification of cluster topics against an external list of pathogens. The small multiples show the distribution across the

clusters of the pathogen named in the panel header, for the 35 pathogens with 40 or more matching documents.

Figure 3 Chart Types in GEViT. We used common names for chart types and also separated them into seven main classes and also one Other class. Special cases of chart types were defined only when there were multiple instance of the same specific chart across our dataset. Chart types with an asterisk mark (*) indicate that they are included in the analysis through manually added articles.

Figure 4 Chart Combinations in GEViT. The six combination types differ based on the number of chart types, the number of charts, and the approach to linking them together.

Figure 5 Chart Enhancements in GEViT. a) Our characterization of marks and their associated aesthetics properties is based on longstanding conventions in the visualization literature[15,19] with roots in Bertin's *Semiology of Graphics*[20]. Illustrative examples are shown for **b)** a tree and **c)** node-link chart types

Figure 6. GEViT Gallery. A screen shot of the resulting GEViT gallery, available online at: http://gevit.net. Images in the GEViT gallery are intentionally blurred for this publication. The GEViT gallery provides links back to the original source publication and presents the images under fair use copyright terms.

Figure 1 Summary of literature analysis steps and document sampling.

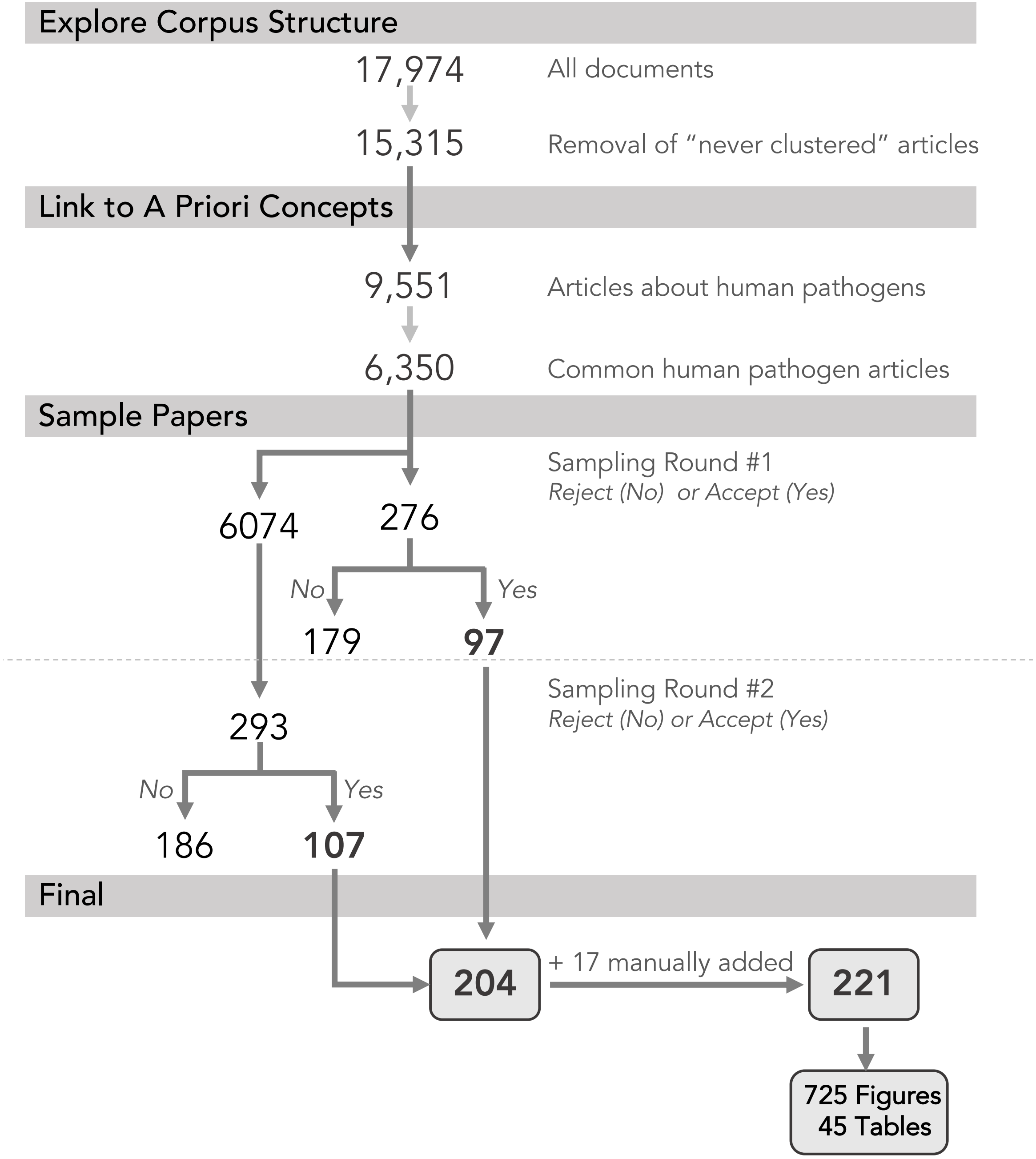


Figure 2 Summary of literature analysis results. a) Documents were classified according to whether they were part of a cluster (green), unclustered under current parameter settings (purple), or never formed part of cluster (orange). The 32 cluster boundaries were automatically determined and are shown as light grey ovals. **b)** Clustered documents and their topics, which are automatically assigned based upon top two terms with the cluster. **c)** Verification of cluster topics against an external list of pathogens. The small multiples show the distribution across the clusters of the pathogen named in the panel header, for the 35 pathogens with 40 or more matching documents.

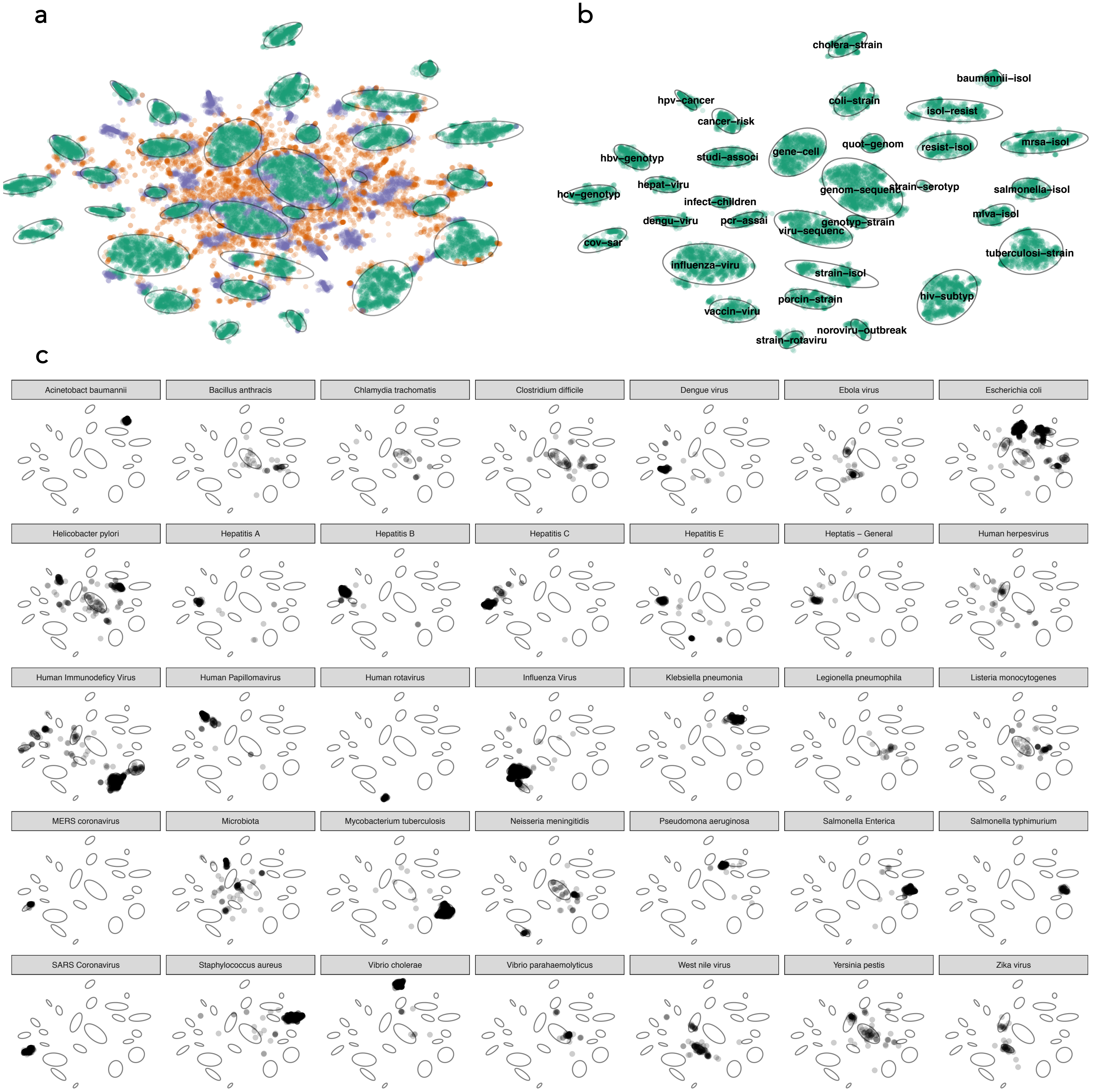


Figure 3 Chart Types in GEViT. We used common names for chart types and also separated them into seven main classes and also one Other class. Special cases of chart types were defined only when there were multiple instance of the same specific chart across our dataset. Chart types with an asterisks mark (*) indicate that they are included in the analysis through manually added articles.

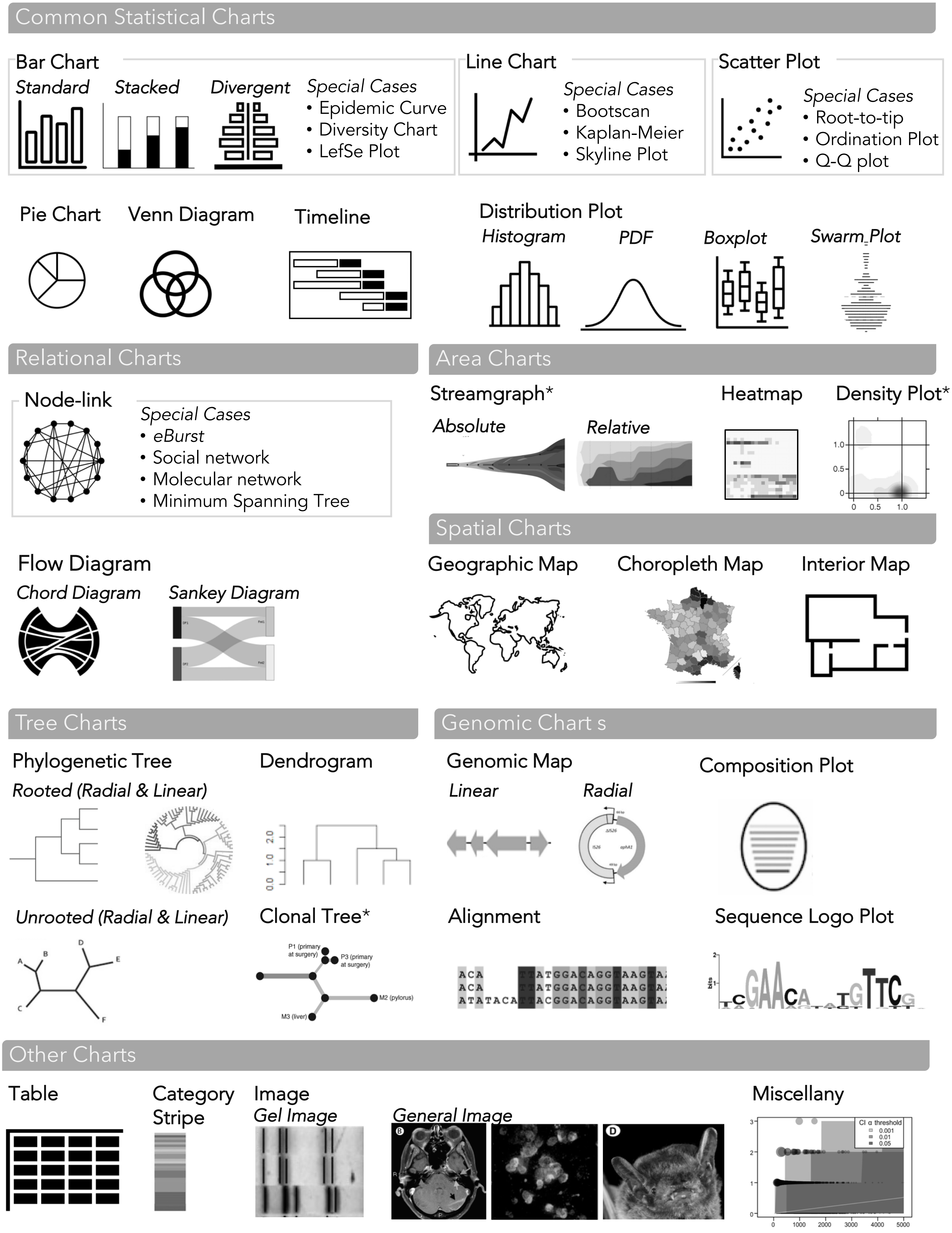
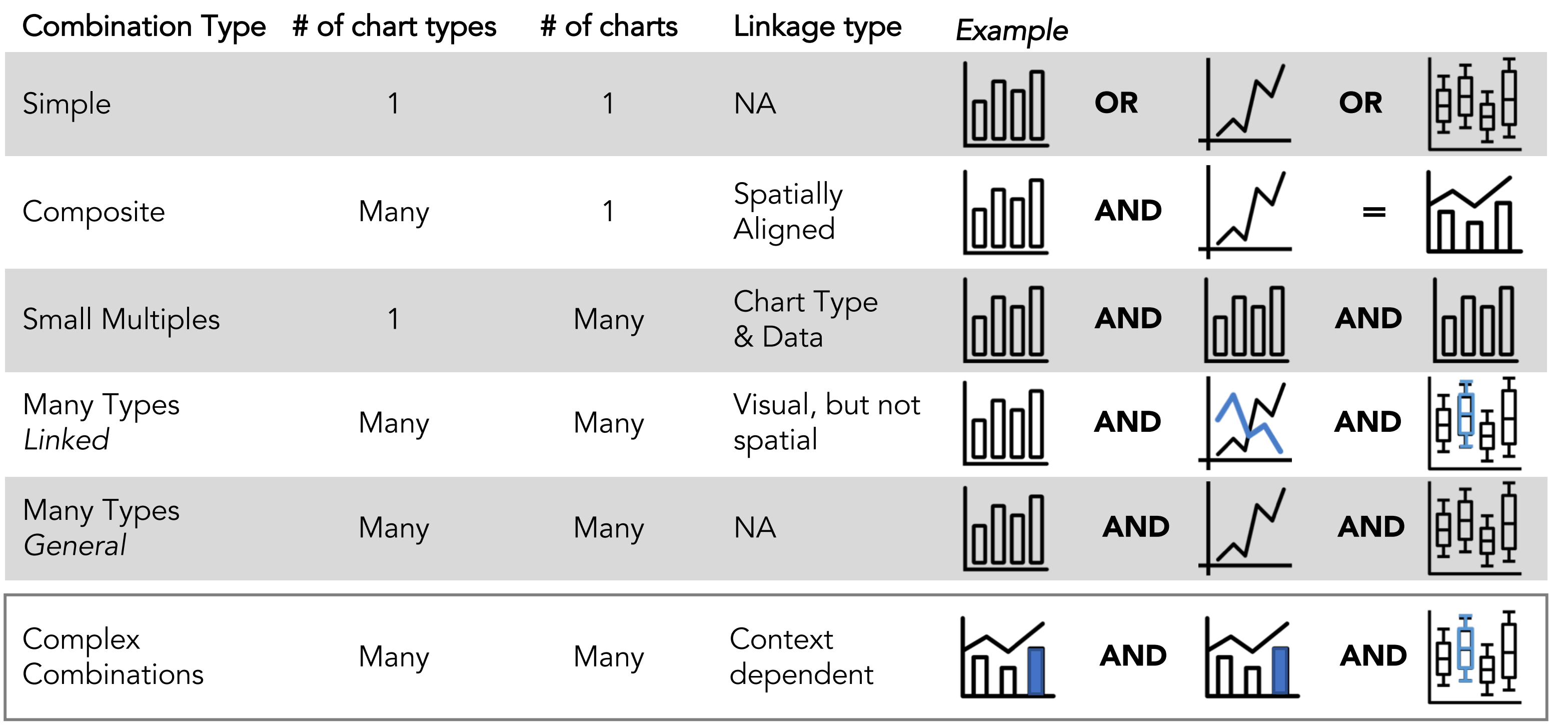


Figure 4 Chart Combinations in GEViT. The six combination types differ based on the number of chart types, the number of charts, and the approach to linking them together.



Combination Type	# of chart types	# of charts	Linkage type	*Example*
Simple	1	1	NA	OR OR
Composite	Many	1	Spatially Aligned	AND =
Small Multiples	1	Many	Chart Type & Data	AND AND
Many Types *Linked*	Many	Many	Visual, but not spatial	AND AND
Many Types *General*	Many	Many	NA	AND AND
Complex Combinations	Many	Many	Context dependent	AND AND

Figure 5 Chart Enhancements in GEViT. a) Our characterization of marks and their associated aesthetics properties is based on longstanding conventions in the visualization literature[15,19] with roots in Bertin's *Semiology of Graphics*[20]. Illustrative examples are shown for **b)** a tree and **c)** node-link chart types

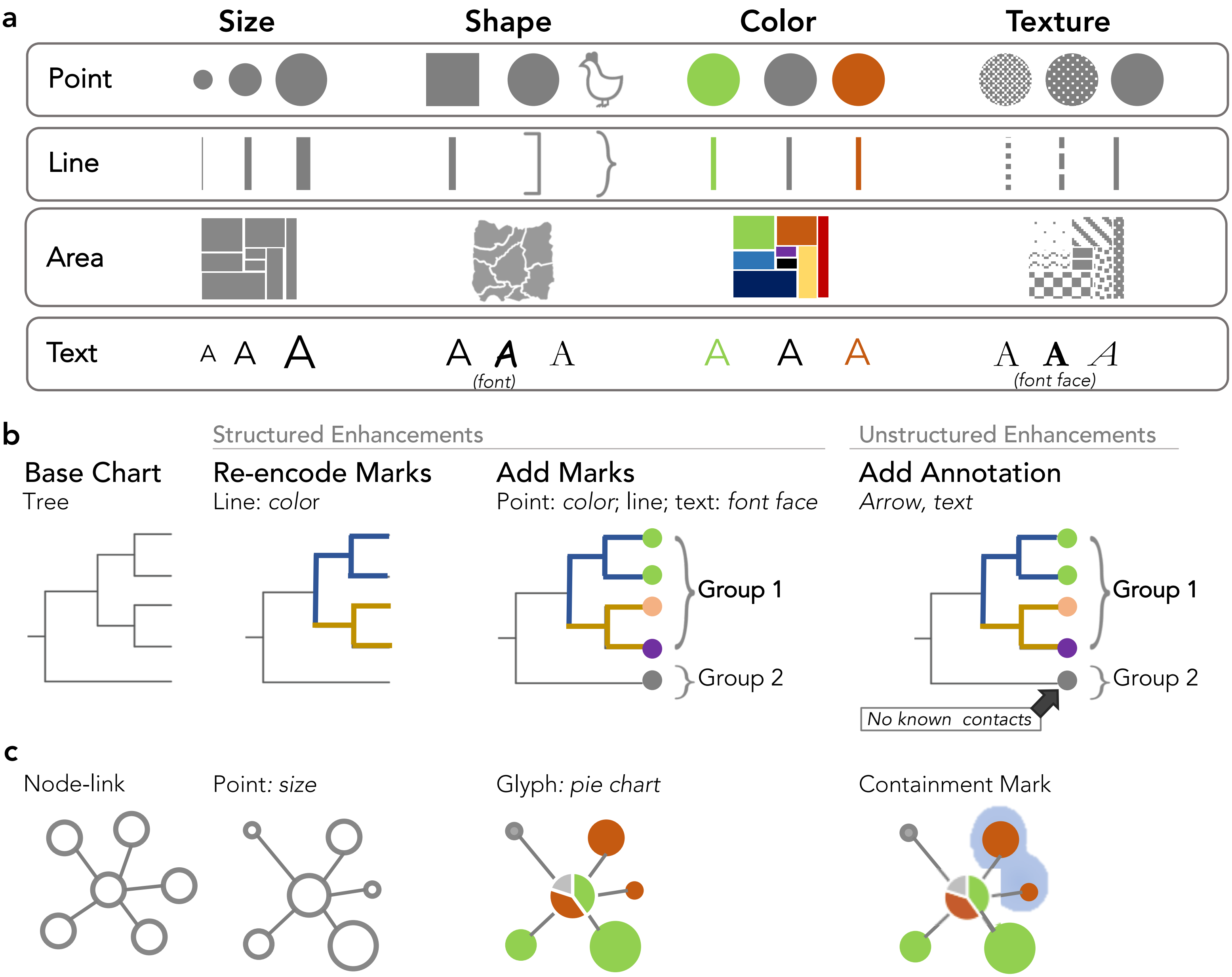
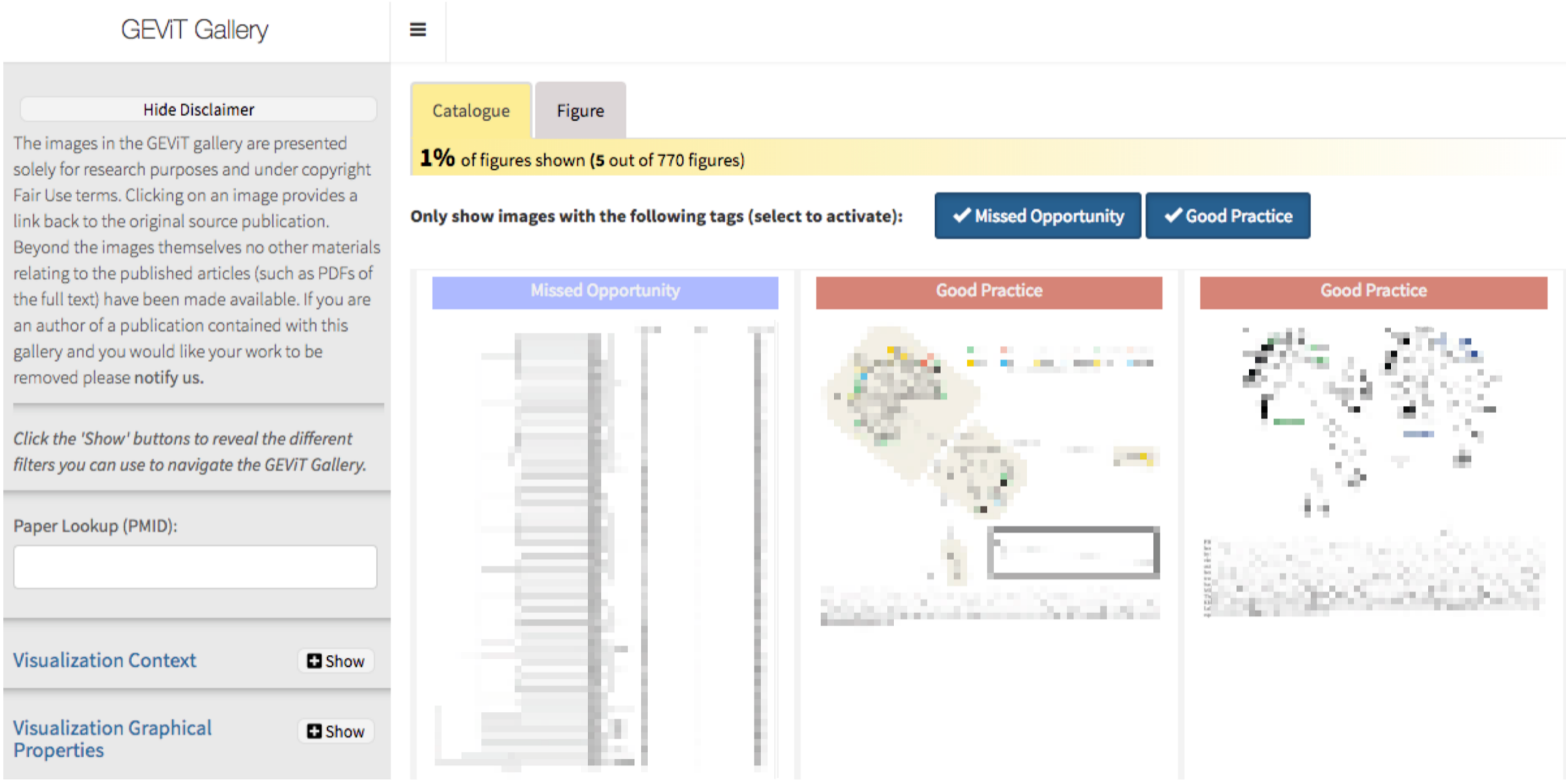


Figure 6. GEViT Gallery. A screen shot of the resulting GEViT gallery, available online at: http://gevit.net. Images in the GEViT gallery are intentionally blurred for this publication. The GEViT gallery provides links back to the original source publication and presents the images under fair use copyright terms.



The images in the GEViT gallery are presented solely for research purposes and under copyright Fair Use terms. Clicking on an image provides a link back to the original source publication. Beyond the images themselves no other materials relating to the published articles (such as PDFs of the full text) have been made available. If you are an author of a publication contained with this gallery and you would like your work to be removed please **notify us.**

Click the 'Show' buttons to reveal the different filters you can use to navigate the GEViT Gallery.

Paper Lookup (PMID):

1% of figures shown (**5** out of 770 figures)

Only show images with the following tags (select to activate):

Online Methods for

A method for systematically surveying data visualizations in infectious disease genomic epidemiology

Anamaria Crisan, Jennifer Gardy, and Tamara Munzner

As with the presentation of the results, the methods are split up into the literature mining and visualization analysis phases. A detailed step-by-step overview of our methods are also shown in supplemental Figures S2 and S3. Our analysis notebooks, data, and associated documents are available online at: `https://github.com/amcrisan/GEViTAnalysisRelease`

Importantly, we use, analyze, and present figures from research articles under "Fair Use Terms", which allows us to use copyrighted materials for research purposes. We make provisions to link back to the original work from which figures are extracted, and do not make any other materials available beyond the figures and article metadata data obtained from PubMed.

LITERATURE ANALYSIS

Aspects of our literature analysis have, with some modification, been turned into an R package called Adjutant, which is available at `https://github.com/amcrisan/adjutant`. A pre-print for Adjutant is available online at `https://www.biorxiv.org/content/early/2018/03/27/290031` and describes the methodology we have used. We do not repeat that methodology in detail here, but we do describe it, and indicate where there are discrepancies between Adjutant's final implementation, and this analysis.

Search Terms. We searched for articles related to infectious disease genomic epidemiology that were published within the past ten years. We used two queries*,* 1) *`(genome AND (outbreak OR pandemic OR epidemic)) OR "genomic epidemiology"`* and 2) *`(genomic epidemiology OR molecular epidemiology) AND (bacteri* OR vir* OR pathogen) AND Genome`* combined their results and retaining only unique records for further analysis.

Data Preparation. The document corpus included only PubMed IDs, year of publication, authors, article titles, article abstract, and associated Medical Subject Heading (MeSH) terms (if there were any). Titles and abstracts were decomposed into single terms, stemmed, and filtered as described in the Adjutant paper. We calculated the term frequency inverse document frequency (td-idf) metric each term, created a sparse Document Term Matrix (DTM) for further analysis. A separate dataset of bigram terms was also prepared but used only for purposes of linking articles to *a priori* concepts (see Main text).

Unsupervised Clustering. We used the t-SNE and hdbscan algorithms to perform an unsupervised clustering using the DTM. While numerous sources advise against clustering on t-SNE results we found that on large document corpuses this approach worked well as we verified with the validity checks described below. We used the Barnes-Hut implementation of t-SNE[21], which allows for some acceleration at the cost of accuracy, with the perplexity parameter set to 100 and otherwise default parameters of the R package implementation[22]. We then used hdbscan[23] on the t-SNE co-ordinate to derive the topic clusters. Clusters are sensitive to the minimum number of cluster points (minPts) parameter supplied to the hdbscan, and so we tried different minPts values (50, 75, 100, 125, 150, 250, 500, 1000), observing how the cluster compositions changed. We observed that some articles never held membership in any cluster irrespective of the parameter settings and labelled those as “never clustered”, in contrast to articles that were simply not clustered with our specific final parameter settings that are labeled as “currently unclustered”. The final set of clusters are a blend of separate parameters (75 and 150). The topic of each cluster is assigned by using the top two most frequent terms within each cluster. Upon observing the cluster results, we validated our clusters using an external list of human pathogens and assessed the correspondence between pathogen terms and cluster topics.

Linking To *A Priori* Concepts. We used the dataset of bigrams and filtered out those that occurred in fewer than 10 articles within a cluster or fewer than 10% of bigrams across bigrams in the corpus. The remaining bigrams were mapped to a set of *a priori* defined concepts, except for bigrams excluded because they were common writing colloquialisms or could not be clearly mapped. This mapping was conducted through iterative internal discussions, in a similar spirit to the visualization analysis described below. We deemed this result acceptable for our analysis needs and did not attempt to further validate it.

Document Sampling. We sampled one document for each *a priori* concept within each topic cluster. Each sampled article was examined and either considered acceptable for further analysis or rejected. Reasons for rejection included: article did not contain any figures (main reason); full text article not accessible; article not in English; article was mainly about a technique (i.e. laboratory technique or bioinformatics method); article did not include humans (animals only, which we considered out of scope); article was a systematic review (figures were mainly illustrations and not data visualizations). For each rejected article, we resampled two additional articles and chose only one article (assuming both were not rejected) for further analysis. Based upon the analysis of the first round of sampling, the second round only sampled articles from 2011 onwards to increase the chance of sampling articles containing figures, and also attempted to sample underrepresented *a priori* concepts from the first round. Table S3 contains a list of all the articles, which round they were sampled in, whether they were included or rejected, and the reason for rejection.

Figure and Table Extraction. To properly capture the figures and their captions, we manually extracted them from PDFs of the sampled articles. Images were only excluded if they were CONSORT diagrams, flow diagrams (excepted only if a data visualization was overlain) or were illustrations. We also included a small number of "missed opportunity" tables, which were stand-alone tables that we felt could have been visualized. This determination was subjective but included tables that were matrices of numbers or large tables of patient metadata where each row consisted of a patient (but demographic tables and statistical summaries were *not* considered missed opportunity tables).

VISUALIZATION ANALYSIS

Figure Analysis. We analyzed whole figures; we did not break them up into individual parts because we wanted to understand the potential interplay between subfigures. For example, if a paper contains three figures (Fig. 1, Fig.2, and Fig. 3) each figure was analyzed separately, whereas if the third figure contains two parts (i.e. Fig. 3A, Fig 3B) those two parts were analyzed *together*.

We generated a descriptive mechanism using qualitative open and axial coding techniques that are routinely used within human-computer interaction (HCI) research[24], which grew out of the Grounded Theory Method developed in the social science fields of sociology, psychology, and anthropology[25]. As we assume that many readers are quantitative researchers, we will briefly describe these techniques in more detail. Grounded Theory refers to a general set of methods used by qualitative researchers to inductively analyze and construct a theory about some phenomenon that is "grounded" in data[24]. In general terms, the idea of Grounded Theory is similar in spirt to unsupervised analysis methods that are applied in quantitative research[26] since both approaches rely on emergent pattern matching that is found within the data rather than applying a specific hypothesis or theory; in qualitative methods the human resolves the relevant patterns, in quantitative methods generally the algorithm does. Curating and labelling data is also standard practice for developing image-based machine learning training datasets and these approaches likely use qualitative techniques without referring to them. We have also found that qualitative research approaches are useful when trying to explore some data without any pre-conceived notions of what the outcomes should be.

The core foundation of Grounded Theory Methods (GTM) rests upon different approaches for assigning descriptive codes to data, typically chunks of text, that become the basis for further analysis[25]. Two widely used approaches are open and axial coding, the latter allowing a researcher to develop hierarchical relationships between codes. Codes are subjectively assigned to data and refined over multiple rounds of data interrogation until a final set of descriptive codes are agreed upon. Notions of validity and generalizability within qualitative research are different than within quantitative research, but there is a notion of at least internal validity for qualitative research and some agreed upon conventions to assess the robustness of the work (see Maxwell[27], Chapter 6), which we have applied in our own research.

We note that the application of GTM is different between the social sciences and HCI, with one large difference being that HCI and information visualization (infovis) researchers frequently apply GTM to text[28], video, and image data[29] whereas social scientists tend to primarily use interview text (although some examples of image analysis with social sciences exist[30]). Our application of GTM, and especially open and axial coding, is drawn from the HCI infovis research traditions, and we also build upon established terminology and ideas from Munzner's Visualization Analysis and Design[15]. We ourselves are primarily quantitative researchers and thus further apply a specific interrogative lens to the way we use GTM. There exists a fascinating and broader discussion about mixed methods approaches to augment the best properties of both qualitative and quantitative research methods[31] , which is beyond the application of this work but that the reader should be aware of.

REFERENCES

21. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15,** 3221–3245 (2014).
22. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. (2015). url: https://github.com/jkrijthe/Rtsne
23. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. *Adv. Knowl. Discov. Data Min.* 160–172 (2013). doi:10.1007/978-3-642-37456-2_14
24. Jacko, J. A. *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition*. (CRC Press, Inc., 2012).
25. Charmaz, K. *Constructing grounded theory: a practical guide through qualitative analysis*. (Sage, 2006).
26. Muller, M., Guha, S., Baumer, E. P. S., Mimno, D. & Shami, N. S. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. *Proc. Gr.* 0–6 (2016). doi:10.1145/2957276.2957280

27. Maxwell, J. A. *Qualitative Research Design: An Interactive Approach*. *Applied social research methods series* **41,** (2013).

28. Furniss, D., Blandford, A. & Curzon, P. Confessions from a grounded theory PhD: Experiences and lesson learnt. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*(2011).

29. Sedlmair, M., Munzner, T. & Tory, M. Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices. *IEEE Trans. Vis. Comput. Graph.* **19,** 2634–2643 (2013).

30. Liebenberg, L., Didkowsky, N. & Ungar, M. Analysing image-based data using grounded theory: the Negotiating Resilience Project. *Vis. Stud.* **27,** 59–74 (2012).

31. Creswell, J. W. & Piano, V. L. Designing and Conducting Mixed Methods Research. *Aust. N. Z. J. Public Health* **31,** 388–388 (2007).

Supplemental Material for

A method for systematically surveying data visualizations in infectious disease genomic epidemiology

Anamaria Crisan, Jennifer Gardy, and Tamara Munzner

Contents

A reminder that analysis notebooks are also available at:
`https://github.com/amcrisan/GEViTAnalysisRelease`

Supplemental Figures

Figure S1 Overview of our approach to construct a visualization design space. This approach is split into two distinct, but connected phases, consisting of a literature analysis and followed by a visualization analysis phase that itself consists of a qualitative and quantitative analysis component. We overlay these phases as concrete steps in resolving our primary research objective, which is stated below.

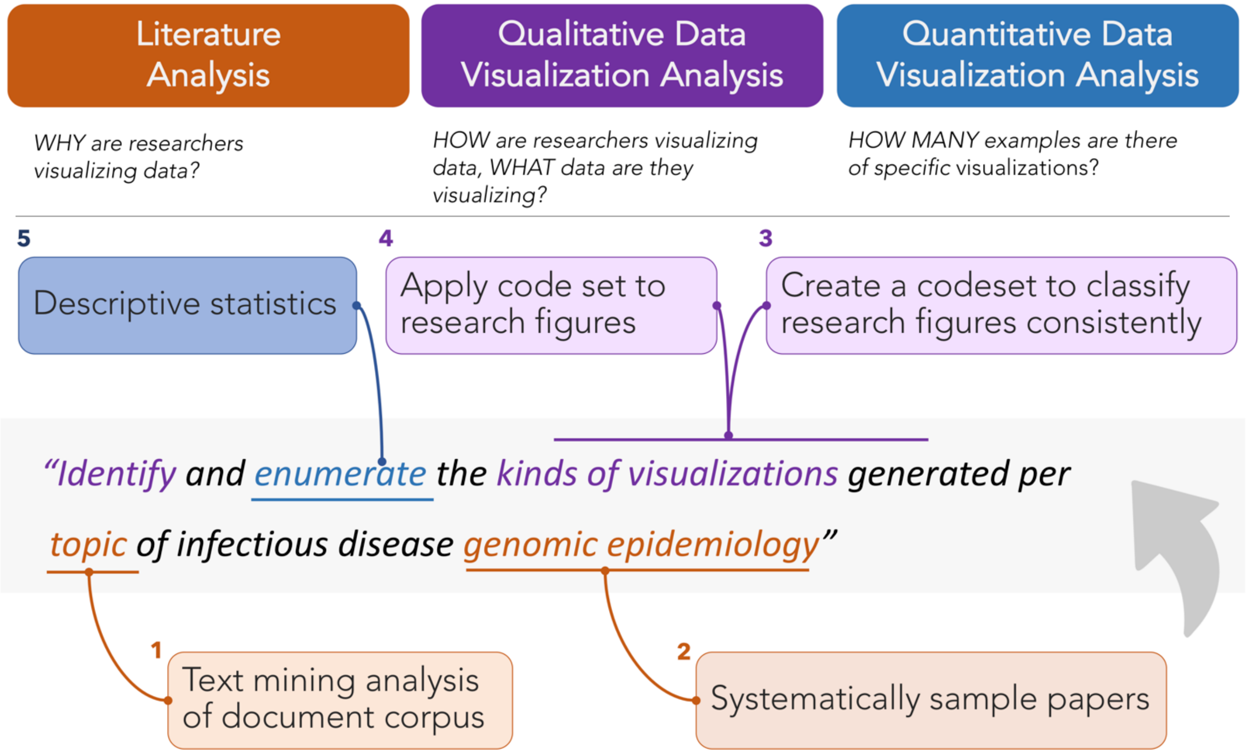
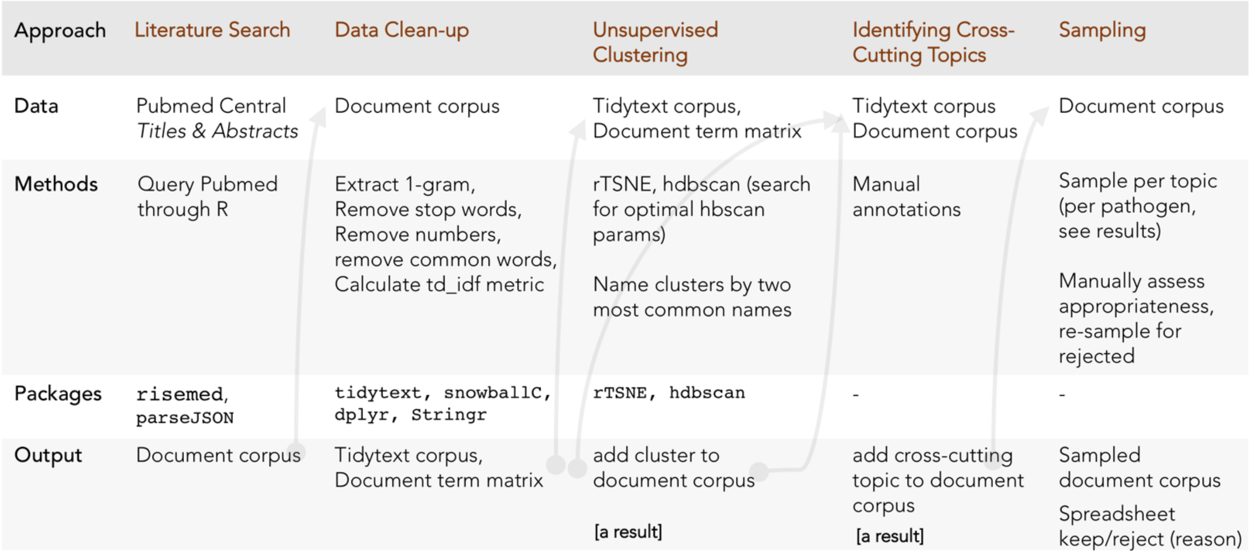
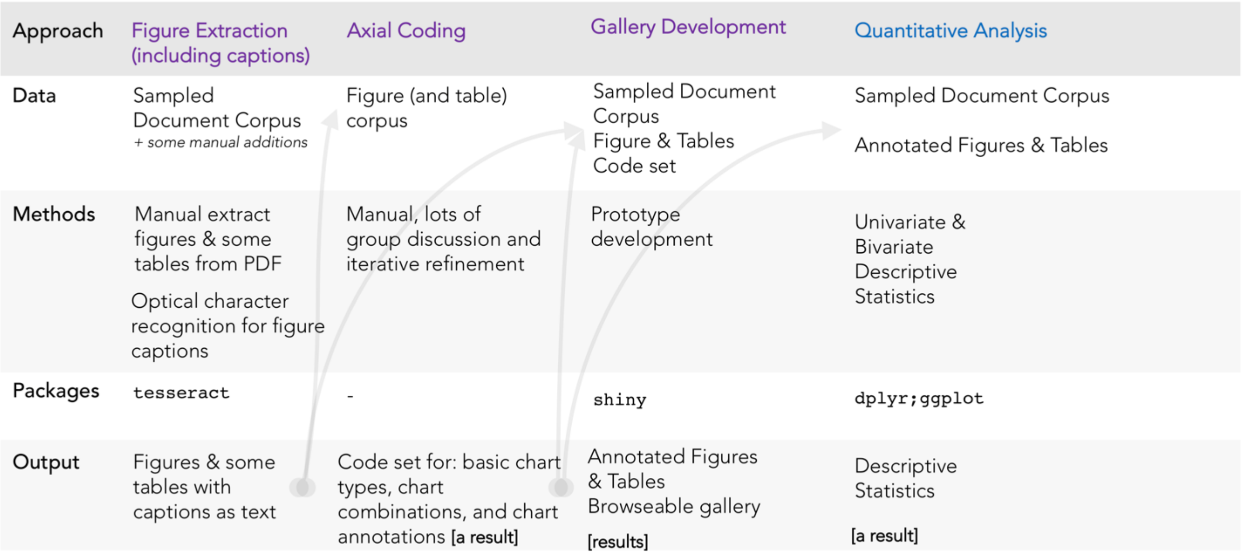

Literature Analysis
Qualitative Data Visualization Analysis
Quantitative Data Visualization Analysis
WHY are researchers visualizing data?
HOW are researchers visualizing data, WHAT data are they visualizing?
HOW MANY examples are there of specific visualizations?
5
Descriptive statistics
4
Apply code set to research figures
3
Create a codeset to classify research figures consistently
"Identify and enumerate the kinds of visualizations generated per topic of infectious disease genomic epidemiology"
1
Text mining analysis of document corpus
2
Systematically sample papers

Figure S2 Literature Mining Methods.



Approach	Literature Search	Data Clean-up	Unsupervised Clustering	Identifying Cross-Cutting Topics	Sampling
Data	Pubmed Central *Titles & Abstracts*	Document corpus	Tidytext corpus, Document term matrix	Tidytext corpus Document corpus	Document corpus
Methods	Query Pubmed through R	Extract 1-gram, Remove stop words, Remove numbers, remove common words, Calculate td_idf metric	rTSNE, hdbscan (search for optimal hbscan params) Name clusters by two most common names	Manual annotations	Sample per topic (per pathogen, see results) Manually assess appropriateness, re-sample for rejected
Packages	`risemed`, `parseJSON`	`tidytext, snowballC, dplyr, Stringr`	`rTSNE, hdbscan`	-	-
Output	Document corpus	Tidytext corpus, Document term matrix	add cluster to document corpus [a result]	add cross-cutting topic to document corpus [a result]	Sampled document corpus Spreadsheet keep/reject (reason)

Figure S3 Qualitative and Quantitative Visualization Analysis Methods.



Approach	Figure Extraction (including captions)	Axial Coding	Gallery Development	Quantitative Analysis
Data	Sampled Document Corpus *+ some manual additions*	Figure (and table) corpus	Sampled Document Corpus Figure & Tables Code set	Sampled Document Corpus Annotated Figures & Tables
Methods	Manual extract figures & some tables from PDF Optical character recognition for figure captions	Manual, lots of group discussion and iterative refinement	Prototype development	Univariate & Bivariate Descriptive Statistics
Packages	`tesseract`	-	`shiny`	`dplyr;ggplot`
Output	Figures & some tables with captions as text	Code set for: basic chart types, chart combinations, and chart annotations [a result]	Annotated Figures & Tables Browseable gallery [results]	Descriptive Statistics [a result]

Figure S4 *A priori* concepts distributed among pathogens (a) and the number to bigram assigned to each concept (b).

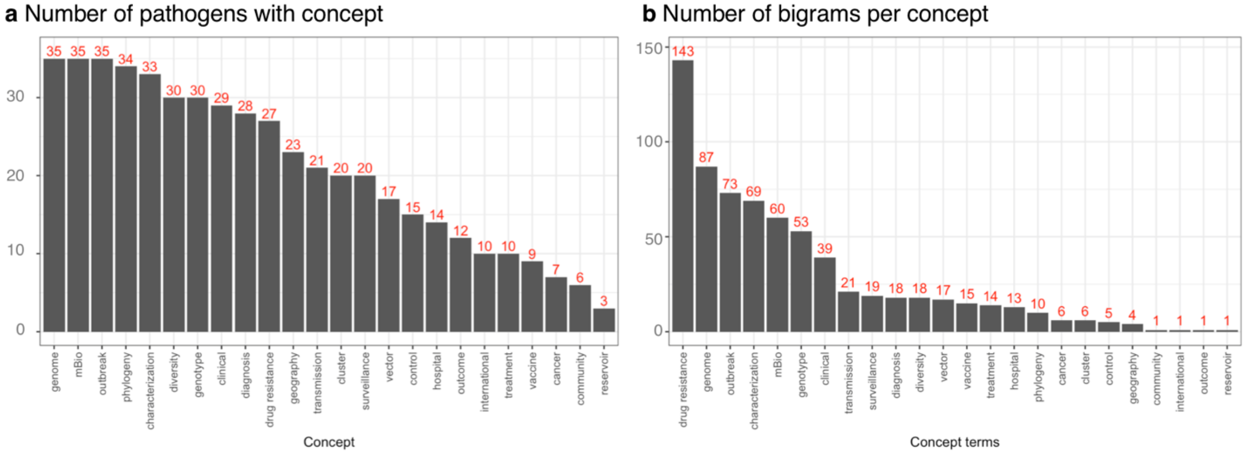
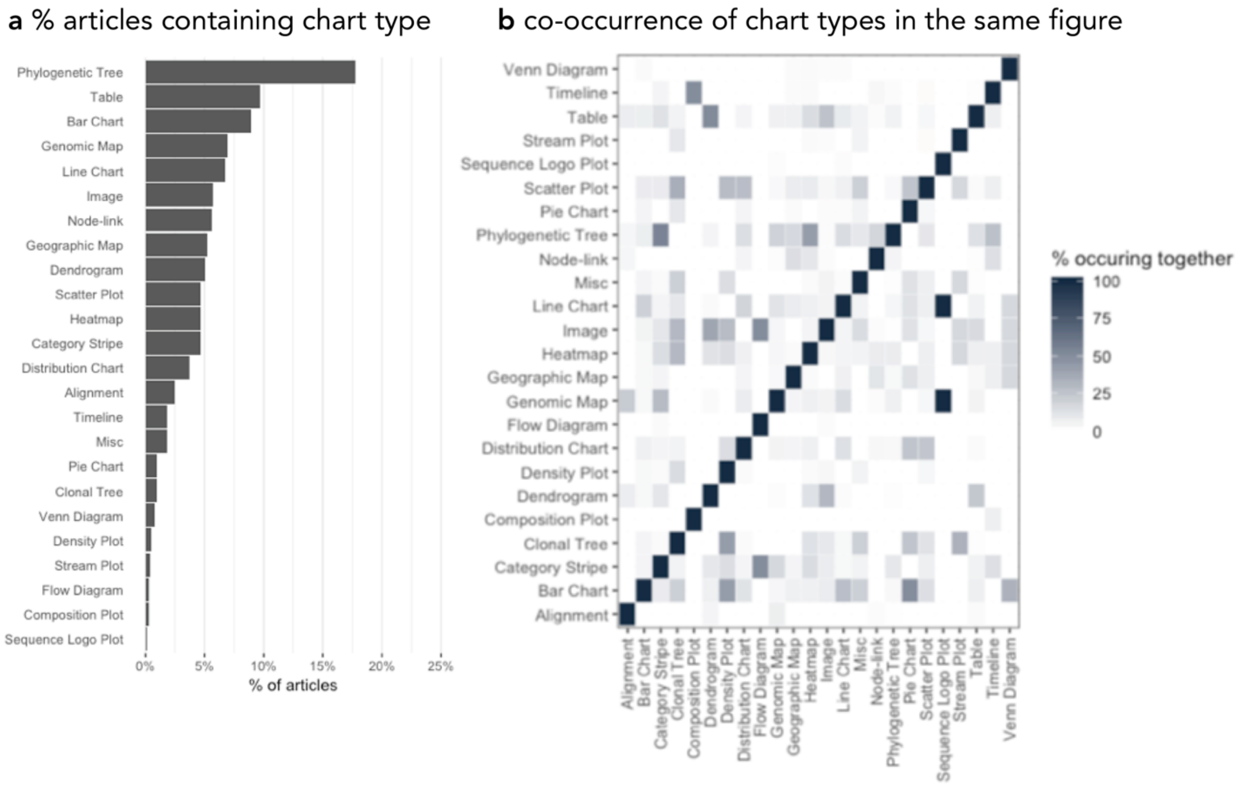


Figure S5 Distribution of chart types of chart type across articles (a) and the co-occurrence of chart types with figures (b)



Supplemental Table Captions

Table S1 External list of pathogens. A list of human pathogens and their associated disease taken from Wikipedia (`https://en.wikipedia.org/wiki/List_of_infectious_diseases`) and used to validate the topic clustering by assessing whether the pathogen strings occur in clusters with the same name. Both the disease and the source of the disease were checked for a match within each document.

Table S2 Mapping of bigrams to concepts.

Table S3 Master list of sampled articles.