

Visualizing Heterogeneous Data in Genomic Epidemiology

by

Anamaria Crisan

B. Comp., Queen's University, 2008

MSc., University of British Columbia, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL
STUDIES

(Computer Science)

The University of British Columbia
(Vancouver)

September 2019

© Anamaria Crisan, 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

Visualizing Heterogeneous Data in Genomic Epidemiology

submitted by **Anamaria Crisan** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Science**.

Examining Committee:

Tamara Munzner, Department of Computer Science
Supervisor

Jennifer L. Gardy, School of Population and Public Health
Supervisor

Raymond Ng, Department of Computer Science
Supervisory Committee Member

Bonnie Henry, School of Population and Public Health
Supervisory Committee Member

Anne Condon, Department of Computer Science
University Examiner

Leanne M. Currie, School of Nursing
University Examiner

David Aanensen, Wellcome Sanger Institute
External Examiner

Abstract

Technological innovations have allowed for a greater variety of data, most notably microbial genomic data, to be collected, integrated, analyzed, and visualized for epidemiological investigations. While analytic methods have evolved in light of this technological change, data visualizations systems have lagged behind.

I take a novel approach that integrates methods from information visualization, human computer interaction, machine learning, and statistics to address unmet data visualization needs in microbial genomic epidemiology (genEpi). This approach also enables me to generate study artifacts that can be used to address regulatory and organizational constraints arising in domains where the use of data is highly restricted. I first present a mixed methods approach to **understand** the needs, data, tasks, and constraints of public health stakeholders that are charged with interpreting the findings of these data. I demonstrate how this approach can be used to communicate new and heterogeneous types of data in a clinical report that is read by stakeholders in different roles. I next present a novel method for systematically reviewing data visualizations that I use to develop a Genomic Epidemiology Visualization Typology (GEViT), which enables others to **explore and characterize** the way the data could be visualized. Finally, I use these collective findings to inform the **design and implementation** of data visualization tools: Adjutant, the GEViT Gallery, minCombinR, and GEViTRec. Adjutant enables rapid and unsupervised topic clustering of PubMed article corpuses to aid systematic and literature reviews. The GEViT gallery is a browsable interface for exploring data visualizations specific to the microbial genEpi domain. minCombinR lowers

the burden to stakeholders for generating combinations of data visualizations for heterogeneous data. Finally, GEViTRec takes a novel approach to the automatic generation of data visualizations that can help stakeholders familiarize themselves with new data. All of these tools integrate with analytic methods.

This research makes novel contributions to the design and implementation of data visualization systems that impact microbial genomic epidemiological data collected for public health investigations. The challenges addressed here are not unique to this domain and my contributions are extensible to other domains grappling with heterogeneous, multidimensional, and restricted data.

Lay Summary

New technologies are enabling public health agencies to collect more data of many different types, which can be used to inform public health policy and practice. Yet, this new “big data” is challenging to analyze and to communicate to stakeholders that need to make decisions with data. In this dissertation research, I developed new approaches for understanding that relationships between data, how it used by stakeholders, and the ways that this data can be visualized. Data visualization forms an effective bridge between increasingly complex data and the methods that are used to analyze it. I have created new techniques and software systems to help stakeholders effectively analyze and visualize their data. My research makes important contributions toward building better analysis tools to help stakeholders in public health, and even beyond, work effectively with complex data.

Preface

Chapters of this dissertation have been previously published with different co-authors. I acknowledge the collaborative nature of this work by using “we” throughout the thesis, with the exception of the Introduction and Conclusion chapters where “I” is used. All research chapters are presented *as they were originally published*.

Chapters 1 and **8** have not been previously published and have been written by me with input from both Drs. Munzner and Gardy.

Chapter 2 has been previously published in the Beyond Time and Errors (BELIV) 2016 workshop proceedings [23], associated with the IEEE VIS week conference:

A. Crisan, J. L. Gardy, and T. Munzner. On regulatory and organizational constraints in visualization design and evaluation. Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization, 1:19, 2016.
doi:10.1145/2993901.2993911

I conducted the analysis for the case study and wrote the initial drafts of the publication. All authors contributed to the analysis and writing of the final publication.

Chapter 3 has been previously published in PeerJ [25]

A. Crisan, G. McKee, T. Munzner, and J. L. Gardy. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. PeerJ, 6:e4218, Jan. 2018.
doi:10.7717/peerj.4218

This research was borne out of a collaboration with the COMPASS-TB team at Public Health England (PHE) to redesign an existing clinical report for next generation sequencing of data. In addition to the report re-design collaboration with PHE, the larger goal of this study was also to collect stakeholder information that could be used to inform subsequent research aims. All authors jointly conceived of the study designs and contributed to the writing of the final publication. Geoffery McKee and I implemented the online surveys, conducted the analysis, developed the final re-designed clinical report, and wrote the initial drafts of the publication. Zipeng Liu, Kimberly Dextras-Romangnino, Dylan Dong, and George Hattab participated in the Design Sprint portion of this research and contributed prototype designs that were evaluated the Design Choice Questionnaire. The content of **Appendix A** was also published alongside this work and includes the online surveys deployed for this study and the notes I compiled to justify design the final clinical report. We have also produced a LaTeX template for the clinical report that is available online:

<https://github.com/amcrisan/TB-WGS-MicroReport>

All study resources and materials that could be publicly released were made available online ahead of publication:

<https://github.com/amcrisan/TBReportRedesign>

Chapter 4 has been previously published in Bioinformatics [26]:

A. Crisan, T. Munzner, and J. L. Gardy. Adjutant: an R-based Tool to Support Topic Discovery for Systematic and Literature Reviews. *Bioinformatics*, 35(6):10701072, 08 2018. doi:10.1093/bioinformatics/bty722

All authors contributed to the writing of the final publication. I developed and implemented the underlying programmatic logic and the Graphical User Interface (GUI) for Adjutant. The content of **Appendix B** was published alongside the initial publication and contains substantial additional analyses on Adjutant’s approach to rapid and unsupervised topic clustering.

Adjutant is available as an open source R package on GitHub:

<https://github.com/amcrisan/Adjutant>

Chapter 5 has been previously published in Bioinformatics [24]:

A. Crisan, J. L. Gardy, and T. Munzner. A Systematic Method for Surveying Data Visualizations and a Resulting Genomic Epidemiology Visualization Typology: GEViT. *Bioinformatics*, 35(10):16681676, 09 2018.
[doi:10.1093/bioinformatics/bty832](https://doi.org/10.1093/bioinformatics/bty832)

All authors jointly conceived of the study and contributed to the writing of the final publication. I developed the initial study ideas, created and implemented the systematic review and analysis methodology, conducted the preliminary analyses, and finally implemented and deployed the online gallery (<http://gevit.net>). The content of **Appendix C** was published alongside this work and includes additional information about the methodologies used in study as well as some supplementary figures. The version of the publication that appears here has a slight modification relative to the original publication. Following feedback from the research community, we changed the names of the combinations from “Composite”, “Small Multiples”, “Many Types Linked”, and “Many Types General” to “Spatially Aligned”, “Small Multiples”, “Visually Aligned”, and “Unaligned”, respectively.

All study resources and materials were made available online ahead of publication:

<https://github.com/amcrisan/GEViTAnalysisRelease>

Chapter 6 was submitted for publication [27]:

A. Crisan, S. Fisher, S. Kasica, J.L. Gardy, and T. Munzner (2019). minCombinR: Coordinating Chart Combinations with Minimal Specifications.
Submitted for Publication

Myself, Dr. Munzner, and Shannah Fisher conceived of the minCombinR’s architecture and contributed to the writing of the final publication. Myself and Shannah implemented minCombinR as an open source package and conducted tests into its capabilities. Stephen Kasica help with the comparison to other tools. **Appendix D** was submitted along this work.

All study resources and materials were made available online ahead of publication :

<https://github.com/amcrisan/minCombinR>

Chapter 7 was submitted for publication [28]:

A. Crisan, J.L. Gardy, and T. Munzner (2019). GEViTRec: Domain-Aware Visualization Recommendation for Data Reconnaissance and Harmonization.
Submitted for Publication

Myself and Dr. Munzner conceived of the original project objectives and contributed to the writing of the final publication. Dr. Munzner and I jointly conceived of the data reconnaissance and task wrangling conceptual framework presented in this work. I developed and implemented the algorithmic logic behind GEViTRec and wrote the initial drafts of the publication. The content of **Appendix E** was submitted alongside this work as a proof-of-concept for GEViTRec.

All study resources and materials were made available online ahead of publication:

<https://github.com/amcrisan/GEViTRec>

Presentation Style of Dissertation, Chapters, and Appendices

The manuscripts that comprise the aforementioned research chapters are written for different audiences and venues, as such they have different writing styles that are dependent upon the publication conventions of each community. Throughout, I also treat the term “data” as plural (for example, I write ‘these data’ and not ‘the data’).

Chapters 2, 6, and 7 are written primarily for an infovis research audience, have an informal tone, and have a looser manuscript structure that includes an introduction, related work (either at the beginning or the end of the manuscript), a theoretical description of an algorithm or toolkit, implementation details, results, discussion and conclusions.

Chapters 3 and 5 are written for a bioinformatics and genomic epidemiology audience, have a more formal tone, and follow a so-called “traditional laboratory style” manuscript, that is they have five strictly defined sections : introduction, materials and methods, results, discussion, and conclusions.

Chapter 4 is also written for a bioinformatics audience, but as an “application note” – a format that briefly describes a software application.

Following the convention of the information visualization research literature I refer to the work conducted within these chapters as *projects*. A project can comprise one or more studies.

The appendices are all written in a relatively informal and conversational tone. They comprise additional figures, tables, details on the methods, and finally tutorials that were published online and also submitted as supplementary materials that accompanied their respective publications.

A final stylistic note is on the structure of this thesis document. As already indicated, chapters are presented in their published (or submitted) structure, which means that each chapter contains an abstract, its own introduction, and conclusions. This is different than other thesis presentation styles that have a single Introduction and Conclusion for the entire document; my chosen presentation style is in keeping with a manuscript thesis format. In this thesis document, I used the Introduction (Chapter 1) to present an overview of the individual research chapters, a summary of their contributions, and how all of these chapters are tied together. In the Conclusions (Chapter 8), I once again summarize the overall findings of these chapters and their contributions, but also comment on the post-publication reception of my research by the public health and visualization communities.

Ethics Approval

The studies described in this dissertation work conducted with the approval of the UBC Behavioral Research Ethics Board, certificate number H10-03336.

Table of Contents

Abstract iii

Lay Summary v

Preface vi

Table of Contents xi

List of Tables xix

List of Figures xx

Glossary xxiii

Acknowledgments xxiv

Dedication xxvii

1 Introduction 1

 1.1 Situating My Research in Prior Work 3

 1.2 Research Overview 7

 1.2.1 Chronology 10

 1.3 Summary of Research Projects and Contributions 11

 1.3.1 Chapter 2: Regulatory and Organizational Constraints
 in Visualization Design and Analysis 12

1.3.2	Chapter 3: Evidence-based design	14
1.3.3	Chapter 4: Adjutant	16
1.3.4	Chapter 5: GEViT	17
1.3.5	Chapter 6: minCombinR	19
1.3.6	Chapter 7: GEViTRec	21
1.3.7	Summary of Contributions	23

2 On Regulatory and Organizational Constraints

	in Visualization Design and Evaluation	24
2.1	Introduction	25
2.2	Defining External Constraints	26
2.2.1	Implications for Evaluation	27
2.2.2	Example: Hypothesis Generation Considered Harmful	28
2.2.3	Example: Agile Development Considered Harmful .	29
2.3	Prior Work	29
2.3.1	Visualization Methodologies	30
2.3.2	External Disciplines	31
2.4	Guidelines for Evaluating External Constraints	32
2.4.1	Defining Stakeholder Roles	33
2.4.2	Generation of Additional Artifacts	36
2.4.3	Methods	37
2.5	Case Study: Healthcare	39
2.5.1	Constraints in Healthcare	41
2.5.2	Lessons Learned in Developing a TB Decision Sup- port Tool	42

3 Evidence Based Design:

	Applying a Design Study Methodology to the Redesign of a Whole Genome Sequencing Clinical Report	49
3.1	Introduction	50
3.1.1	Human-Centered Design in the Clinical Laboratory . .	51
3.1.2	Collaboration Context COMPASS-TB	52

3.2	Materials and Methods	54
3.2.1	Overview of Design Study Methodology	54
3.2.2	Discovery Stage	56
3.2.3	Design Stage	57
3.2.4	Implementation Stage	59
3.3	Results	60
3.3.1	Experts Emphasized Prioritizing Information and Re- vealed Constraints	60
3.3.2	Experts Vary in Their Perception of Different Data Types	61
3.3.3	WGS Data is Vital, but Some Lack Confidence in its Interpretation	64
3.3.4	Respondent Consensus Suggests a Role for WGS in Diagnosis and Treatment Tasks	65
3.3.5	Prototyping Via a Design Sprint Produces a Range of Design Alternatives	66
3.3.6	The Design Choice Questionnaire Quantifies Partici- pant Preferences for Specific Design Elements	68
3.3.7	Qualitative Data Affords Additional Insights into Re- port Design	71
3.3.8	Developing a Final Report Template	74
3.4	Discussion	76
3.5	Conclusions	80

4 Adjutant:

An R-based Tool to Support Topic Discovery for Systematic and Literature Reviews		81
4.1	Introduction	82
4.2	Implementation Details	82
4.3	Usage Scenario	85
4.4	Conclusion	85

5	GEViT:	
	A Systematic Method for Surveying Data Visualizations and a Resulting Genomic Epidemiology Visualization Typology .	86
5.1	Introduction	87
5.2	Methods	90
5.2.1	Developing a Method for the Systematic Analysis of Data Visualizations	90
5.2.2	A Systematic Analysis of Data Visualizations from the Infectious Disease Genomic Epidemiology Re- search Literature	91
5.2.3	Visualization Analysis	94
5.3	Results	95
5.3.1	Literature Analysis	95
5.3.2	Visualization Analysis	99
5.4	Discussion	106
5.4.1	Implications of our Findings for Visualization Design	108
5.4.2	Implications of our Findings for the Genomic Epi- demiology Community	110
5.5	Conclusion	111
6	minCombinR:	
	Coordinating Chart Combinations with Minimal Specifications	112
6.1	Introduction	113
6.2	Domain Motivation and Design Decisions	115
6.2.1	GEViT Findings	116
6.2.2	Design Decisions	117
6.3	Related Work	118
6.3.1	Stand-Alone Applications	119
6.3.2	Charting Libraries and Packages	120
6.3.3	Domain-Specific Tools	121
6.4	Design of minCombinR	122
6.4.1	From Typology to Toolkit	122

6.4.2	Gradual Binding Architecture	124
6.4.3	Specification	125
6.4.4	Creation and Integration	129
6.4.5	Arrangement and Display	131
6.5	Implementation	132
6.5.1	User Functions and Specifications	133
6.5.2	Supported Data and Chart Types	137
6.5.3	Combination Control Flows	137
6.6	Results	139
6.6.1	Showcasing minCombinR on Different Datasets . .	140
6.6.2	Comparison to Existing Tools	140
6.7	Discussion and Future Work	144
6.8	Conclusion	145

7 GEViTRec:

Domain-Aware Visualization Recommendation for

	Data Reconnaissance and Harmonization	146
7.1	Introduction	147
7.2	Background	150
7.3	Data Reconnaissance and Task Wrangling	151
7.3.1	Operational Definitions	152
7.3.2	Conceptual Framework	153
7.4	Formalisms for Visualization Recommendation	156
7.4.1	Domain Prevalence Design Spaces	157
7.4.2	Data Model	158
7.4.3	Visualization Specification	159
7.5	General Algorithm	160
7.5.1	Mapping From Datatypes to Visual Encodings with a Design Space	160
7.5.2	Data Harmonization and Entity Graph Generation .	162
7.5.3	Ranking Paths Within the Entity Graph	164
7.5.4	Generating Specifications	166

7.5.5	Composing Views for Display	169
7.6	Implementation of GEViTRec	169
7.7	Results	171
7.8	Related Work	175
7.8.1	Rule-Based Approaches	176
7.8.2	Ontology-Based Approaches	177
7.8.3	Machine Learning	177
7.8.4	Stack Comparisons	178
7.9	Discussion and Future Work	179
7.9.1	Generalizability	179
7.9.2	Is Relevance Relevant?	180
7.10	Conclusion	181
8	Reflections and Conclusion	183
8.1	Reflections on Research Projects and Contributions	185
8.1.1	Regulatory and Organizational Constraints	185
8.1.2	Evidence Based Design	186
8.1.3	Adjutant	188
8.1.4	GEViT and the GEViT Gallery	188
8.1.5	minCombinR	191
8.1.6	GEViTREC	191
8.2	Reflecting on the Merits and Challenges of Interdisciplinary Research	193
8.3	Overall Limitations and Future Work	195
8.4	Concluding Remarks	198
	Bibliography	199
A	Evidence Based Design Supplemental Materials	218
A.1	Supplemental Figures	219
A.2	Supplemental Tables	220
A.3	Justification for Final Design Choices by Section	228
A.3.1	Analysis of Quantitative and Qualitative Results	228

A.3.2	ISO15189 Requirements	232
A.4	Task and Data Questionnaire Online Survey	238
A.5	Design Choice Questionnaire Online Survey	254
B	Adjutant Supplemental Materials	277
B.1	Adjutant Implementation Details	277
B.2	Adjutant in Action	279
B.2.1	t-SNE with Simulated Data	279
B.2.2	Investigating Adjutant with Real Data	300
B.2.3	Alternative Approaches	312
C	GEViT Supplemental Materials	316
C.1	Supplemental Methods for Visualization Analysis	316
C.2	Supplemental Figures	318
C.3	Supplemental Tables	319
D	minCombinR Supplemental Materials	336
D.1	Generating Simple Charts with minCombinr	336
D.1.1	Common Statistical Charts	338
D.1.2	Colour Charts	341
D.1.3	Relational Charts	342
D.1.4	Spatial Charts	343
D.1.5	Tree Charts	347
D.1.6	Genomic charts	347
D.1.7	Temporal Charts	351
D.1.8	Images	353
D.2	Generating Combinations of Charts with minCombinR	358
D.2.1	Unaligned	359
D.2.2	Small Multiples	360
D.2.3	Colour Aligned Combinations	366
D.2.4	Spatially Aligned Combinations	368
E	GEViTRec Supplemental Materials	375

E.1	Data Harmonization	377
E.2	Generate Specifications	377
E.3	Generated Views	378

List of Tables

Table 3.1	Evidence based design study participants	60
Table A.1	Task and data questionnaire study participants	220
Table A.2	Respondents anticipated future use of molecular/genomic data	221
Table A.3	Respondents' confidence to interpret laboratory data . . .	222
Table A.4	Respondents' confidence using genomic data	223
Table A.5	Identification barriers impacting respondents' workflows	224
Table A.6	Summary of design choice questionnaire results	225
Table C.1	External list of pathogens	320
Table C.2	Mapping of bigrams to <i>a priori</i> concepts	322
Table C.3	Master list of sampled articles	334
Table C.4	Final set of pathogens and pathogen clusters	334

List of Figures

Figure 1.1	Overview of research projects, questions, and contributions	9
Figure 1.2	Doctoral research timeline	10
Figure 2.1	Summary of our proposed additions to the Design Study Methodology	30
Figure 2.2	Power interest matrix of stakeholder roles	33
Figure 3.1	Initial COMPASS-TB report design	53
Figure 3.2	Human-centered design approach	55
Figure 3.3	Expert consensus for workflow tasks and data	63
Figure 3.4	Digital mockups of complete report prototypes	66
Figure 3.5	Isolated design components	67
Figure 3.6	Design choice questionnaire results	70
Figure 3.7	Original and revised reports	74
Figure 4.1	Adjutant user interface	83
Figure 5.1	GEViT method and application overview	89
Figure 5.2	Summary of literature analysis steps and document sam- pling	96
Figure 5.3	Summary of literature analysis results	97
Figure 5.4	Chart types in GEViT	102
Figure 5.5	Chart combinations in GEViT	103
Figure 5.6	Chart enhancements in GEViT	104

Figure 6.1	Architectural layers of minCombinR	123
Figure 6.2	User and derived partial specifications	124
Figure 6.3	Overall flow of specifications and control in minCombinR	126
Figure 6.4	Currently implemented chart types in minCombinR . .	132
Figure 6.5	Colour aligned combination of disparate static charts . .	135
Figure 6.6	Code, control flow, and resulting displays for the four combination types	136
Figure 6.7	minCombinR comparison to related work	143
Figure 7.1	Conceptual framework for data reconnaissance and task wrangling	153
Figure 7.2	Data reconnaissance and task wrangling phases over time	154
Figure 7.3	Comparing different approaches to human centered design	154
Figure 7.4	Data reconnaissance and task wrangling phase breakdown	155
Figure 7.5	Data harmonization and entity graph generation schematic algorithm overview	163
Figure 7.6	GEViTRec internal visual encoding templates	166
Figure 7.7	Mapping from datatypes to chart types	168
Figure 7.8	GEViTRec code and resulting visualization	172
Figure 7.9	GEViTRec results with Ebola outbreak data	173
Figure A.1	Survey responses with confidence intervals	219
Figure B.1	Simple example: two class simulated distributions	281
Figure B.2	Simple example: applying t-SNE	283
Figure B.3	Simple example: t-SNE results with varying parameters	284
Figure B.4	Simple example: t-SNE results with varying parameters II	286
Figure B.5	Simple example: hdbscan on dimensionally reduced data	287
Figure B.6	Simple example: clusters resolved by Adjutant	289
Figure B.7	Complex example: multiclass simulated distributions .	292
Figure B.8	Complex example: t-SNE results with varying parameters II	293
Figure B.9	Complex example: hdbscan on dimensionally reduced data	295

Figure B.10 Complex example: clusters resolved by Adjutant	297
Figure B.11 Complex example with noise	298
Figure B.12 Complex example with noise: t-SNE results with varying parameters	298
Figure B.13 Complex example with noise: hdbscan on dimensionally reduced data	299
Figure B.14 Complex example with noise: clusters resolved by Adjutant	299
Figure B.15 Real data: clusters identified by Adjutant	303
Figure B.16 Real data: distribution of terms across clusters	306
Figure B.17 Real data: distribution of terms across clusters II	307
Figure B.18 Real data: distribution of terms across clusters III	310
Figure B.19 Real data: distribution of terms across clusters IV	311
Figure B.20 Real data: LDA clusters and topics	314
Figure B.21 Real data: LDA gamma distribution	315
Figure C.1 Literature mining methods	318
Figure C.2 Qualitative and quantitative visualization analysis methods	318
Figure C.3 <i>A priori</i> concepts distributed among pathogens and the number of bigrams assigned to each concept	319
Figure C.4 Distribution of chart types across articles and the co- occurrence of chart types with figures	319
Figure E.1 GEViTRec entity graph	378
Figure E.2 GEViTRec generated view #1	379
Figure E.3 GEViTRec generated view #2	380
Figure E.4 GEViTRec generated view #3	381
Figure E.5 GEViTRec generated view #4	382
Figure E.6 GEViTRec generated view #5	383

Glossary

BCCDC British Columbia Centre for Disease Control

DSM Design Study Methodology

genEpi Genomic Epidemiology

GEViT Genomic Epidemiology Visualization Typology

GUI Graphical User Interface

HCI Human Computer Interaction

infovis Information Visualization

PHE Public Health England

TB Tuberculosis

UBC University of British Columbia

WGS Whole Genome Sequencing

Acknowledgments

Foremost I would like to thank my supervisors, Dr. Tamara Munzner and Dr. Jennnifer Gardy, for their support and guidance over these past four years. They gave me the opportunity to pursue a research trajectory I deeply cared about and gave me the freedom to explore new methodological approaches that satisfied my intellectual curiosities.

I would like to thank the members of my supervisory committee, Dr. Bonnie Henry and Dr. Raymond Ng. Given the interdisciplinary nature of my research, it was important for me to have voices from both my chosen disciplines, epidemiology and computer science, appraise my work. Both Drs. Henry and Ng brought important perspectives to my research and I am grateful for their feedback.

The members of the University of British Columbia (UBC) Information Visualization (infovis) groups, both past and present, have also provided a valuable perspective to my research and were often among the first to appraise its strengths and weakness. I would like to thank Michelle Borkin, Matthew Brehmer, Kimberly Dextras-Romangnino, Dylan Dong, Madison Elliott, Shannah Fisher, George Hattab, Stephen Kasica, Zipeng Liu, and Michael Oppermann.

I have also been fortunate to have collaborators at the British Columbia Centre for Disease Control (BCCDC), that were my colleagues prior to and throughout my doctoral studies: Dr. Robert Belshaw, Ms. Catharine Cham-

bers, Dr. Victoria Cook, Mr. Michael Coss, Dr. Jennifer Guthrie, Dr. James Johnston, Dr. William Hsiao, Dr. Geoffery McKee, Dr. Michael Otterstatter, Dr. Natalie Prystajecky, and Dr. David Roth. The members of the Canadian Bioinformatics Workshop on Microbial Genomic Epidemiology and the Biological Data Visualization community have also played an important role in helping me refine and disseminate my research and I am grateful for their collegiality. Lastly, among my colleagues I wish to thank the Bedford Lab at the Fred Hutch Cancer Research Centre, who let me occupy a space in their lab during my many visits to Seattle.

I have been fortunate in my professional career to develop lasting friendships that have helped me make the difficult decisions in my research career. Dr. Christine Buerki has been a supportive mentor for nearly a decade. She helped me develop into a better researcher and helped me to get started along the path toward my doctoral research. I am immensely grateful to her. Dr. Ruth Miller (and her dad Bill) gave me the push I needed to apply to a doctoral program and I am grateful for their confidence in me when I myself was unsure. Finally, Ms. Marguerite du Plessis, to whom I am grateful for her friendship, encouragement, and willingness to listen to my rants.

I would also like to acknowledge my funding sources: Vanier Canada Graduate Scholarships, UBC Four Year Fellowships, the Li Tze Fong Memorial Award (UBC Affiliated Fellowships), and the UBC Public Scholars Program. I have also been fortunate to have much of my conference travel sponsored through awards from several organizations: the American Microbiology Society, the Canadian Bioinformatics Workshop, the Canadian Institutes of Health Research Institute for Population and Public Health, the Department of Biomedical Informatics at Harvard Medical School, the International Society for Computational Biology, the IEEE VIS Doctoral Colloquium, the University of Manitoba, and the Wellcome Genome Campus Scientific Conferences.

Most important of all, I would like to thank my family. I arrived with my Mom, Dad, and younger brother in Canada in July of 1990 and I defended this dissertation 25 years to the day that I became a Canadian citizen. When we arrived in Canada, my parents hoped to open up as many opportunities to me as possible and along the way they encouraged me explore my interests and curiosities. Because of this support, I grew up into the scientist I had always hoped to become. From a distance, my Romanian grandparents, cousins, aunts, and uncles sent their love and encouragement as well. Over time I have been fortunate to extend this family, and so would also like to thank the Brehmer clan for welcoming me and cheering me on.

The support of my family was important to taking on this doctoral research. Deciding to return to graduate school from industry was tough decision and one that I was uncertain about in the beginning. You don't quite know until the end if it will all pay off, or if you have forgone opportunities by choosing academia rather than remain in the fray of industry. When I won the Vanier Scholarship I called my mom and dad to tell them the news and there was a pause on the phone before my dad said "we made it". In those words, my concerns about lost opportunity costs dissolved and I was able to pursue a research program that I was fully invested in and proud of.

Lastly, but most importantly, an especially grateful note of thanks is to my husband Dr. Matthew Brehmer, who has been my source of inspiration, sanity, terrible puns, food, understanding, and love. It was been a joy to be on this journey together.

Dedication

For my grandmothers, Ana and Maria

Chapter 1

Introduction

We are facing many new challenges, and these cannot be understood by using the visual metaphors we've been using for centuries — Manuel Lima

Data visualization has been a component of public health research and practice since John Snow created his infamous 1854 cholera map. The dominant narrative of Snow's story is that by plotting the cases of infected individuals on a map, he formulated a hypothesis that the Broad street pump was the source of the outbreak. He verified this hypothesis by removing the pump handle and ending the outbreak. The factually accurate narrative is more complex and illustrates the dynamic interplay between data, statistical analyses, visualization, and actionable insights. Snow initially undertook a considerable statistical effort to implicate water, and not commonly held belief of bad air ("miasmata"), as the vector transmitting the cholera contagion. While he suspected the Broad Street pump as the source of the outbreak he still developed the cholera case map to confirm his hypothesis, which was further verified by removing the pump handle. It was through the combined expository power of the statistical procedures and data visualization that Snow produced a potent analytic repertoire that today still inspires analysts in public health and beyond. For his investigation and resolution of the cholera outbreak, Snow is credited as the founder of the discipline of epidemiology.

Since Snow’s seminal research epidemiology has significantly evolved, introducing new, heterogeneous, and multidimensional sources of data, together with increasingly complex analytic procedures [42]. Within public health, an influx of new whole genome data has changed the resolution that stakeholders, which includes clinicians, nurses, policy makers, epidemiologists, analysts, and researchers, can investigate disease outbreaks. However, genomic data also introduced new complexities; it was difficult to integrate genomic data with currently existing data sources, including tabular data from electronic health records, contact network data, and spatial data [23, 25]. While new statistical procedures have emerged to respond to these new and complex datasets, data visualisation techniques have not evolved at a same pace and there remains a considerable need to better integrate statistical and visual analysis methods [18]. Thus, that interplay between analysis and visualization that Snow elegantly demonstrated is at risk of being disrupted, with the consequence of introducing literal blind spots into modern epidemiological investigations.

I became interested in the interplay between statistics and data visualization while analyzing some of these aforementioned complex datasets. Like Snow, I relied on both statistical and visual approaches to formulate a more complete understanding of these data and to prioritize more viable and actionable insights over other findings. As both my data and analysis procedures would continue to grow in complexity, I began to encounter limitations with existing data visualization tools. I discovered I was not alone. Through this doctoral research, I sought to establish new approaches for visualizing data stemming from microbial Genomic Epidemiology (genEpi) investigations. However, knowing that genEpi is one facet of the growing discipline of Data Science, I also sought to produce technical and methodological contributions that could generalize beyond the specification application context I present in this thesis.

1.1 Situating My Research in Prior Work

Each chapter of this dissertation contains a detailed introduction with a review of the literature that is relevant to its subject matter. This introduction provides a summary of prior research that emphasizes studies and findings that were influential to my overall dissertation. I also present gaps in the prior work that my research sought to address.

Data visualizations developed for public health have long been drawn by hand and thus could be as expressive as the imagination of its creator. Technological innovations have since allowed public health stakeholders to use computers to generate data visualizations, which enabled stakeholders to incorporate wider variety of data types, including microbial genomic data. If stakeholders are sufficiently technically savvy, or have support from technical personnel, they can expressively create data visualizations libraries from within R, Python, or Java Script (using packages such as ggplot [124], matplotlib [56], D3 [9], vega-lite [100]) and link the generation of these visualization to analysis procedures. However, few public health stakeholders have such resources and so rely primarily on systems developed by others. Even in instances when stakeholders are well resourced, they may lack the time to create bespoke custom solutions and so still rely on a rich ecosystem of tools to expressively generate data visualizations.

Data in modern genomic epidemiology investigations are drawn from heterogeneous sources that must be integrated, transformed, and analyzed, and visualized together. This heterogeneity of data adds a level of complexity to the design and implementation of both analytic and visualization tools. However, existing data visualization systems are still limited in the types of data they support, the range of visualizations they can produce, and their ability to connect to different analytic methods. Overall, it can still be complex to generate expressive data visualizations from these complex data. I have experienced this limitation in my own prior research, and it was a motivating factor in undertaking this research. To illustrate this point, I will describe

some of these existing tools and the way I have observed them being used by public health stakeholders that must collect, analyze, and interpret the results from heterogeneous genomic epidemiology data.

I will begin by discussing tools available for visualizing the most ubiquitous data in genomic epidemiology: phylogenetic data, which shows the evolutionary relationships of specimens isolated from infected individuals over time. There exist a number of systems to view these data and among the most widely used are Tree Viewer [55], ggtree [135], and ape [87]. These systems that can visualize phylogenetic tree data together with associated data that contains additional contextual information (i.e. geographic regions, year specimen was acquired, etc.). The latter two, ggtree and ape, are R packages that can integrate with a variety of analytic procedures and other visualization libraries in R, while Tree Viewer, although widely used, is much more limited. Still, these systems primarily produce a visualization of one or more phylogenetic trees and to visualize other types of data, for example genomic, network, or spatial (geographic) data, stakeholders must turn to other tools. It then becomes necessary to integrate the visualization results of multiple tools, a procedure that is called ‘post-processing’, in order to arrive a final data visualization. For example, a stakeholder may need to visualize network data using Cytoscape [105], geographic data using ArcGIS, genomic data via the Integrated Genomics Viewer (IGV) [97] or Island Viewer [30], and tabular data (also referred to as a linelist to public health stakeholders) via Excel or Tableau. Statistical analysis may be conducted in SAS, R, or some other tool. Finally, Adobe Illustrator, PowerPoint, or InkScape, may be used to integrate all of these visual results together. This is just one example of a combination of systems among many potential options. In recent years, data visualization systems have been developed to enable a stakeholder to more easily integrate and visualize heterogeneous genomic epidemiological data. These systems include GenGIS [88], Microreact [4], and Nextstrain [48], which support the visualization of genomic, geographic, temporal (Microreact & Nextstrain), some network (Nextstrain), and some

genomic (Nextstrain) data. These tools represent the current state of the art, but still have considerable limitations toward the types of data, visualizations, and analyses they support.

Although these systems exist they are not widely used outside of a research context. An excellent review by Carroll [18] indicates a number of reasons as to why this may be the case, such as poor fit for different stakeholders (like nurses and clinicians), inability to integrate with existing clinical, treatment, and surveillance workflows, and constraints that limit how data can be used and whether a system can be installed on a stakeholder’s workstation. Digging more deeply into the concerns that Carroll raises, it is possible to also see how constraints stemming from data access and use as well as stakeholder familiarity with new and emerging data types all play critical roles that impact the design and implementation of data visualization systems. Unfortunately, the existing literature in microbial genomic epidemiology, or bioinformatics more generally, offers very little guidance toward understanding those needs, data, and tasks (the procedures stakeholders perform with data) as well as the constraints on these data. For a more principled approach to the design and implementation of data visualization systems, it becomes necessary to use the methods from another discipline.

A large body of literature in the Information Visualization (infovis) research community is precisely dedicated to understanding the needs, data, and tasks of users as well as exploring and characterizing ways to visualize these data. Using techniques from both quantitative and qualitative user research [57], the infovis literature advocates a so-called design study methodology to elicit needs, constraints, data, and tasks that through an iterative process are then married to visual encodings (a more technical and precise term for data visualization). There are a number of design study methodologies, but the most widely used one is reported by Sedlmair *et al.* [103] and is referred to as the DSM. The iterative cycles of ‘the DSM’ are intended to encourage both developers and stakeholders to consider multiple different visual alternatives before arriving at a final visual encoding that is suitable

for the needs, data, and tasks of stakeholders. For example, a developer may initially consider a phylogenetic tree with branches coloured to represent different geographic origins of specimen isolates. However after consulting with stakeholders, and through iterative refinements, she might instead end up at a visual encoding of a tree and map with a co-ordinated colour scheme. In other words, in order to identify an optimal solution the DSM is intended to help developers and stakeholders navigate a visualization design space, which is a conceptualization of all possible ways data could be visualized. The DSM approach is in contrast to an *ad hoc* approach to the visualization of data, which does not consider multiple alternatives or have an awareness of a visualization design space. The infovis research literature provides concrete examples of such visualization design spaces, for example Set Vis [1], Tree Vis [101], or BioVis Explorer [58]. More generally, infovis research also continues to explore the potential types of visual encodings that could exist for different data types, thus constantly expanding the size of a visualization design space.

The prior research stemming from microbial genEpi and infovis literature provide different approaches for visualizing data. Absent is a more systematic and unified methodology that integrates methods and practices from the research disciplines underpinning both communities. Such an integration would benefit both disciplines and support the development of better data visualization tools capable of linking to analytic procedures. For stakeholders representing public health, bioinformatics, and microbial genEpi, an awareness of the infovis research methodologies would allow them to develop useful data visualization tools and to meaningfully assess their efficacy. Importantly, such an awareness would also motivate bioinformatics developers to more carefully consider the evolving data landscape that arises from technological changes. Likewise, the landscape of data, diverse stakeholders, and unavoidable constraints, can result in novel approaches to design studies methodologies. In particular, design study methodologies could benefit from more robust and systematic evidence based approaches that are the norm in

public health investigations.

This dissertation presents a unified methodology that borrows techniques from a number of disciplines to address presently unmet needs toward the visualization and analysis of microbial genomic epidemiological data. The results of applying this methodology are novel contributions of data and visualization findings, algorithms, and tools that are specific to microbial genEpi, but that can also extend beyond this domain.

1.2 Research Overview

My doctoral research undertook an interdisciplinary approach that integrated techniques from public health, human computer interaction, machine learning, and information visualization across several research projects. These projects sought to **understand** stakeholders needs, data and tasks, **explore and characterize** existing data visualization strategies, in order to define the problems in how and why stakeholders in genEpi visualized data. Next, my research sought to **design and implement** new data visualization systems and algorithms as solutions to previously identified visualization problems.

These research projects and their contributions are summarized in detail in the subsequent Section 1.3 and here I present a brief overview of the different research chapters comprising this thesis (Figure 1.1)

In Chapters 2 and 3, I present two projects I undertook to **understand** stakeholders needs, data, and tasks. In Chapter 2, I present a data visualization case study in public health to demonstrate how regulatory and organization constraints impact data access and consequently visualization design and evaluative approaches. In Chapter 3, I gathered both quantitative and qualitative evidence to understand stakeholders' available data, their ability to interpret these data, and how these data are used for different tasks. Both projects were carried out collaboratively with teams at the British Columbia Centre for Disease Control (BCCDC) and Public Health England (PHE) and focused on genomic epidemiology applied to Tuberculosis (TB).

In Chapter 5, I present a project I undertook to **explore and characterize** data visualization strategies already in use by stakeholders. I present a novel method for systematically reviewing data visualizations to derive a Genomic Epidemiology Visualization Typology (GEViT) that can characterize a so-called *domain prevalence visualization design space*. The work presented in Chapter 5 complements the research in Chapters 2 and 3 by combining a holistic community perspective with that of a smaller and concentrated group of stakeholders. Importantly, the research in Chapter 5 links specific types of data, which are explored in detail in Chapter 3, to visual representations already in use by the genEpi community.

Finally, in Chapters 4, 5, 6, and 7, I present the **design and implementation** of data visualization solutions that build off of the findings of the prior research projects. In Chapter 4, I present Adjutant, a novel method with an accompanying system for rapid and unsupervised topic clustering of text data. In Chapter 5, I present a data visualization gallery intended to help stakeholders explore alternative visual designs for different data types and creation contexts; for example, browsing visualizations that show genomic data visualized in an outbreak. Both Adjutant and the gallery are products of the systematic review method presented in Chapter 5. In Chapter 6, I present minCombinR, a toolkit that minimizes the amount of code stakeholders need to write in order to produce different types of charts and combinations of charts that I catalogued in Chapter 5. Finally, in Chapter 7, I present GEViTRec, an algorithm for automatically generating data visualizations given only input data. The GEViTRec algorithm uses the collective evidence from previous chapters to link data to visual representations, identify relevant visualizations, and finally to render visualizations to the display for the user. Adjutant, minCombinR, and GEViTRec are all implemented in R and are able to integrate with data analyses procedures within the larger R ecosystem.

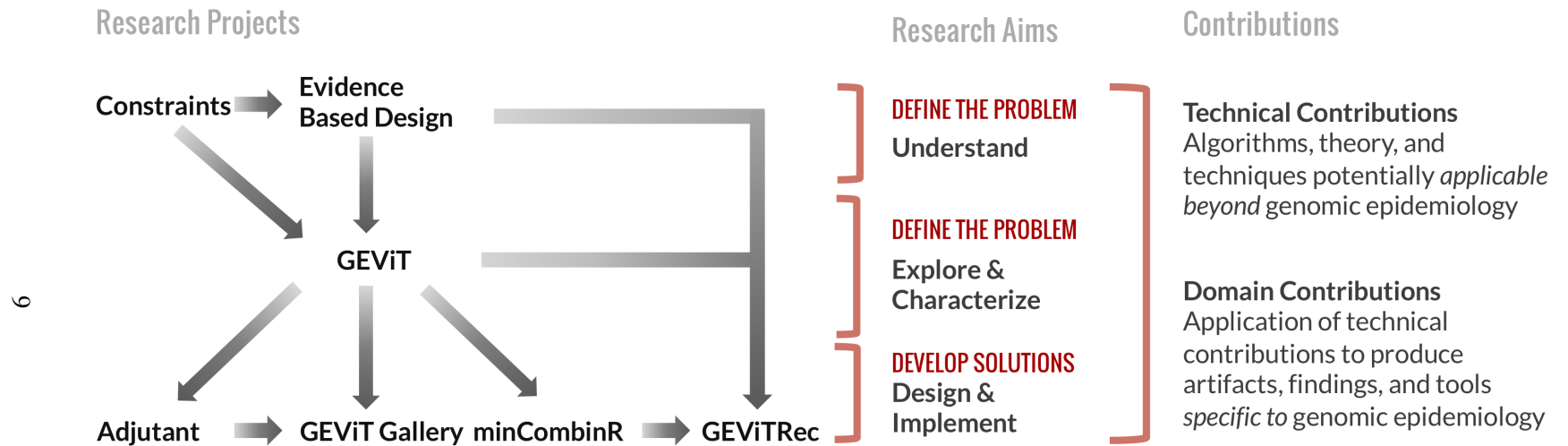


Figure 1.1: Overview of research projects, questions, and contributions. Research projects are structured to address specific research aim. My first research aim concerned defining the problems of how and why genEpi stakeholders visualized data and where there existed limitations. The subsequent research aim focused on developing solutions for the problems identified. These research projects are inter-related whereby prior project results establish the trajectories of future project directions. Finally, these research projects collectively contribute both technical and domain specific contributions.

1.2.1 Chronology

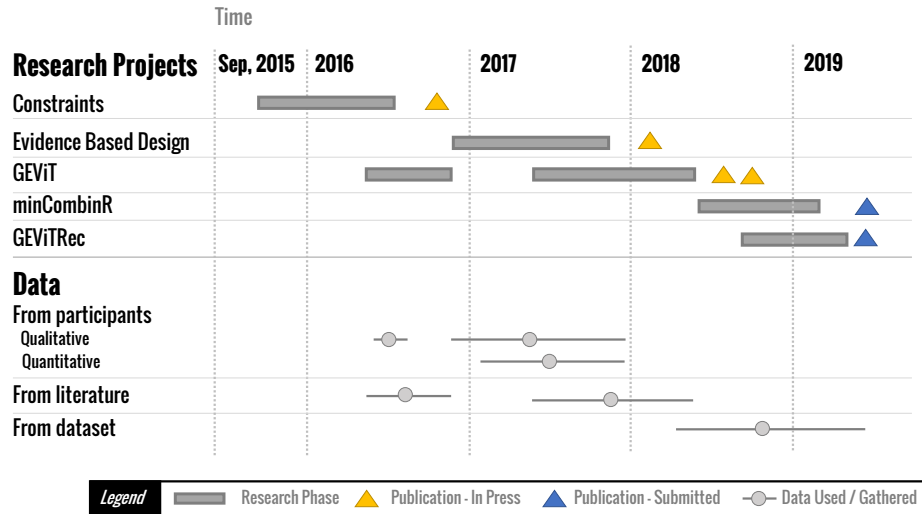


Figure 1.2: Doctoral Research Timeline. This timeline contains the different research projects presented in the chapters of the dissertation, along with the duration of time spent on each project and the publication status. I have also included the different sources of data (participants, literature, and datasets) that I gave generated and used in this dissertation.

The overarching trajectory of my research is to understand, explore and characterize, and finally design and implement data visualization tools; this trajectory was established at the outset of my doctoral research. Consequently, the chronological order that these research projects were undertaken (Figure 1.2) is important because the results of prior projects influenced the research approaches undertaken in the subsequent projects.

I conducted an initial pilot project with stakeholders at the BCCDC to develop and deploy an analytic data visualization to support management and control of tuberculosis. The initial project allowed me to identify data, constraints, missing stakeholders, and available infrastructure for data visualizations tools. These results are summarized in Chapter 2, entitled **Regulatory and Organizational Constraints in Visualization Design and Analysis**. I built upon these initial results through a collaborative project with PHE in

Chapter 3, **Evidence-based Design**. Next, I connect data to visualizations through the development of a **GEViT** (Chapter 5), and concurrently developed **Adjutant** (Chapter 4) and the **GEViT Gallery**. These aforementioned chapters have all been published and also presented in venues that target a *bioinformatics, genomic epidemiology, and public health* audience.

The **minCombinR** toolkit (Chapter 6) and **GEViTRec** algorithm (Chapter 7) were concurrently developed. In fact, GEViTRec relies on minCombinR to create data visualizations that are viewed by stakeholders. These chapters target a *information and biological visualization* audience.

1.3 Summary of Research Projects and Contributions

I have produced both **technical** and **domain specific** contributions stemming from the research projects presented in Chapter 2 to 7, inclusive. Technical contributions produce a methodology, technique, or algorithm that is intended to be generally applicable. Domain specific contributions are artifacts that result from the application of my technical contributions to a specific application domain, in this case genEpi. For example, in Chapter 5 I developed a method for systematically reviewing data visualizations, which is a technical contribution, and produce both a typology and online gallery, both domain specific contributions, that demonstrate the results of my method when applied to a domain specific research question.

In this section I present a summary of these research projects and their contributions. I also provide additional contextual information for how all of these research projects and their contributions are linked together.

1.3.1 Chapter 2: Regulatory and Organizational Constraints in Visualization Design and Analysis

Information visualization research exists within a larger data ecosystem that governs *what* data can be used and also *how* it should be visually represented for analysis. In Chapter 2, I have defined regulatory and organizational constraints and argue that such constraints are not adequately accounted for in existing infovis design and evaluative methodologies [23]. I demonstrate how these constraints implicate visualization design and analysis through a presentation of a case study that documented the results of a pilot project intended to build a visual analytics tool for the tuberculosis team at the BCCDC.

To ameliorate the visualization design and evaluation difficulties that are introduced by regulatory and organization constraints I have modified a widely used Design Study Methodology (DSM) [103]. First, I introduced the idea of a “power-interest” matrix that allows researchers to generate a more fine-grained classification of stakeholders than the existing DSM enables. Importantly, by classifying stakeholders according to both their power over a project and their interest in its outcomes, researchers are able to prioritize those stakeholders that regulate data access, researcher processes, and whether software can be installed on stakeholder work stations. Next, I borrowed from the statistical analysis literature to demonstrate how synthetic datasets can be generated and used in infovis research processes. I also show how initial tool prototypes developed using this synthetic data can be used to get buy-in from stakeholders and support data access. Importantly, the majority of infovis research literature around design studies methodologies emphasis the use of real data through the design and evaluation process. My research concretely demonstrates both the value and necessity of generating representative synthetic datasets for analysis.

Through a case study carried out in collaboration with stakeholders at the BCCDC interested in TB prevention and control, I produce an example of a

“power-interest” matrix and a synthetic dataset used for prototyping a data visualization tool. I also demonstrate how these and other artifacts generated from the research processes were important for communicating with stakeholders and getting buy-in from gatekeepers. I conclude this chapter by summarizing a set of general recommendations to tackle regulatory and organizational constraints through the data visualization design and evaluation process.

This initial investigation into the constraints that govern data usage and access would prove to be influential in the research projects undertaken in subsequent chapters. The “power-interest” matrix I developed was important for understanding stakeholder availability, data knowledge, and priorities. These findings would motivate the projects I present in Chapters 3 and 5 and also influenced my research methodologies for these projects. Moreover, an awareness of regulatory and organizational constraints influenced the *kinds* of tools I should develop since many stakeholders encountered similar constraints when they carried out their own research. The minCombinR toolkit and GEViTRec algorithm presented in Chapters 6 and 7, respectively, are both influenced by both data accessibility and stakeholders research needs in light of regulatory and organizational constraints.

Contributions for Chapter 2

Technical Contributions:

- Formal definitions of regulatory and organizational constraints and an assessment for their impact in visualization design and evaluation
- Incorporation of project management techniques with a DSM
- Strategies for generating synthetic datasets in visualization design and evaluation

Domain Specific Contributions:

- Case study documenting a data visualization pilot project for genEpi, including a tool prototype and initial evaluation
- Stakeholder classification for public health roles

1.3.2 Chapter 3: Evidence-based design

Advances in technology have enabled public health stakeholders to collect greater quantities of more variable data types to manage and control disease outbreaks. In Chapter 3, I sought to gather evidence of the types of data used in different public health diagnostic, treatment, and surveillance tasks, as well as to understand how confident stakeholders are interpreting these data for their tasks [25]. My co-authors and I partnered with the COMPASS-TB team from PHE to re-design a TB Whole Genome Sequencing (WGS) clinical report. Through the process of the clinical report re-design, I also sought to gather data that would be instrumental for informing subsequent data visualization tool development.

The project reported in this chapter employed a multi-phased mixed methods research approach integrated with the modified DSM (from Chapter 2). I also used the stakeholder classifications from Chapter 2 to identify study participants. The project's first phase used an exploratory sequential model study design [22] to conduct semi-structured interviews with a selective number of stakeholders in order to established workflows, specific tasks, and finally the data used to complete those tasks. These qualitative findings were then transformed into an online questionnaire that established the extent of consensus (if any) among stakeholders that some data were used to complete a specific task. I also assessed stakeholder ability to interpret these data in order to perform diagnostic, treatment, or surveillance tasks. I used the qualitative and quantitative results to inform the second project phase of generating paper prototypes of alternative clinical report designs. I collaborated with the information visualization research group at University of British Columbia (UBC) to come up with report prototypes, which I later converted

into another online questionnaire to assess stakeholders' preference for specific representations of WGS data. To assess preference, study participants were asked to select or rank preferred data representations and to provide textual comments to justify their preferences. While we predominantly used these findings to generate a final re-design of the WGS clinical report, our results identified important and general principles for communicating complex data to stakeholders. These principles are summarized as three experimental and five design guidelines. All study materials are available in Appendix A. The re-designed clinical report was publically deployed and also integrated into the analysis pipelines of PHE and others.

A revealing insight from this research project was that there existed very little stakeholder consensus toward the types of data and even visual representations for genEpi surveillance tasks. These are tasks that monitor populations for the emergence and progression of disease outbreaks. Stakeholders primarily used case counts (number of infected individuals) to make assessments for surveillance tasks and while there existed considerable enthusiasm to use new WGS technologies it was still not very clear how to use these new data. By comparison the role of different data types, even genomic data, was much more clearly defined for diagnostic and treatment tasks. These findings laid out a clear path for future work to support surveillance tasks with multiple heterogeneous types of data.

Contributions for Chapter 3

Technical Contributions:

- A mixed methods research approach for information and visualization design and evaluation
- Experimental and design guidelines for communicating new and complex data to stakeholders

Domain Specific Contributions:

- Links between data and tasks for tuberculosis applications for genEpi
- Emphasizing areas of greatest need in genEpi
- A realized and deployed clinical report

1.3.3 Chapter 4: Adjutant

Chapter 4 presents Adjutant, a system for rapid and unsupervised topic clustering of PubMed articles [26]. Adjutant is used within the GEViT study (Chapter 5), but these two chapters are presented separately because they represent two different publications.

There exist many systems for topic classification of text, but I found that these systems placed a significant burden on the user to provide *a priori* labels for text documents in order to perform an accurate classification. Document labelling is time consuming and so I developed a rapid and unsupervised method for automatically clustering documents. In Appendix B, I include considerable materials that assessed Adjutant’s clustering accuracy and comparison to the widely used Latent Dirichlet Allocation (LDA). Adjutant is available as an R package that is distributed via GitHub. In addition to a rapid and unsupervised clustering method, I also developed an Adjutant Graphical User Interface (GUI) that allows stakeholders to initiate queries and explore the resulting clusters for themselves.

While Adjutant was developed to support the research presented in the subsequent Chapter 5, this system is capable of performing analysis on any PubMed queries.

Contributions for Chapter 4

Technical Contributions:

- A system for rapid and unsupervised topic clustering

1.3.4 Chapter 5: GEViT

Chapters 2 and 3 demonstrate the challenges of integrating new types of data into healthcare settings, despite enthusiasm from stakeholders to include these data in their surveillance tasks. These challenges also introduce complexity into the data visualization design process : *how can we build data visualization tools when stakeholders are also trying to understand these data and their uses?* My insight was that research communities lead the exploration of new data and its uses, including the visualization of these data. I wanted to leverage this collective community wisdom to speed up the data discovery and exploration processes through data visualization.

The research presented in Chapter 5 sought to identify and classify community strategies for data visualization by creating a method for systematically reviewing a data visualization corpus and generating a typology to describe and enumerate these strategies [24]. I demonstrate this method in action through its application to the genEpi research literature and the development of a **Genomic Epidemiology Visualization Typology (GEViT)**.

First I present a method to systematically survey data visualizations that can describe *why* a visualization was created, *how* it was created, and finally *how many* examples there are of specific data visualization strategies. This systematic review method consists of an initial literature analysis phase that uses Adjutant (Chapter 4) to conduct an unsupervised topic clustering of research articles. Topic clusters are used as strata within a random sampling procedure in order to harvest figures (data visualization) within these articles. The topic clusters provide a sense of *why* a visualization was created and my goal was to obtain a broad sample of visualization strategies across different creation contexts. The literature analysis phase produces a dataset of figures that are supplied to the qualitative and quantitative procedures of a visual analysis phase in order to describe *how* a visualization was created and *how many* examples there are of different visualization strategies. The qualitative stage of the visualization analysis phases uses iterative axial coding

techniques [20] to derive visualization annotations across three descriptive axes, chart types, chart combinations, and finally enhancements. Together, these descriptive axes form the taxonomy of a visualization typology that can describe and enumerate common visualization strategies. I refer to the final dataset of annotated and enumerated data visualization strategies as referred to as a *domain prevalence visualization design space*.

I applied this method to approximately 18,000 research articles from the genEpi research literature and I developed a visualization typology called “GEViT” through an analysis of 842 figures (data visualizations) derived from 221 articles sampled from 36 topics clusters. Enumerating the visualization strategies also summarized the diversity of current common genEpi visualization practices. What I found was that stakeholders used only a small set of visualization strategies, often just showing a single phylogenetic tree or a tree with an accompanying table. While there were examples of much more sophisticated visualization designs that emerged from my dataset, these were the exceptions as the majority were relatively simple and consisted of poor visualization design choices that left much of the data encoded as text in the visualization. Interestingly, while the strategies taken by individual research articles tended to be limited, the combinatorial space of visual designs revealed by GEViT showed that there existed many possibilities that the majority of stakeholders did not explore. Thus, the whole visualization design space was useful, whereas individual examples from a specific paper had variable and sometimes limited visual expression. In Appendix C, I present additional methodological details and figures that support the analysis presented in this chapter.

The GEViT project is an important component of my research that linked stakeholders’ data, tasks, needs, and existing data visualization strategies. The findings from this research directly influenced the development of three data visualization tools presented in this thesis. Within this chapter I present the GEViT gallery (<http://gevit.net>), which I included within the GEViT publication. The GEViT gallery allows stakeholders to explore

visualization designs across different creation contexts, or to browse visual alternatives via the typology terminology. I have also annotated examples of “good” and “missed opportunity” visualization practices to help guide stakeholders toward better practices.

Contributions for Chapter 5

Technical Contributions:

- A method for systematically reviewing data visualizations
- A method for generating a domain prevalence visualization design space

Domain Specific Contributions:

- GEViT, which classified genEpi data visualization strategies according to chart types, combinations, and enhancements
- A dataset of annotated data visualizations
- The GEViT gallery tool

1.3.5 Chapter 6: minCombinR

An analysis of the domain prevalence visualization design space generated by the GEViT study revealed that there did not exist a single data visualization tool that can visualize multiple data types. Furthermore, I found that there existed few tools that allowed stakeholders to integrate various statistical and phylogenetic analyses with the visualization of their heterogeneous datasets. Instead, stakeholders would need to programmatically generate data visualizations that accompanied their analysis, using R, Python, or JavaScript, a processes that was labour intensive and complex for the chart types, combinations, and enhancements revealed by GEViT.

Chapter 6 presents minCombinR, a toolkit that supports a minimal specification syntax for generating a variety of chart types and their combinations in R. Currently, stakeholders must consider both what visualizations they wish to generate from some data and to programmatically specify how those visualization should be rendered by the R graphics device. The minCombinR toolkit uses a declarative framework that allows stakeholders to simply describe the chart types and combinations that they would like to generate without having to specify *how* the resulting visualizations should be generated. A stakeholder is warned when certain visualizations are not possible to generate and is prompted through a set of steps to help them generate a viable data visualization specification. The result is that the minCombinR toolkit allows stakeholders to create visualization with as little as three lines of code. The minCombinR toolkit is developed as an R package and distributed via GitHub. It is capable of integrating with analytic methods in the R ecosystem.

In Appendix D, I present the applicability of minCombinR using various different chart types and their combinations using datasets specific and agnostic to genEpi.

Contributions for Chapter 6

Technical Contributions:

- A simplified syntax and declarative framework for rendering various chart types and chart combinations
- Chart harmonization through gradual binding

Domain Specific Contributions:

- minCombinR: a tool to support stakeholder's ability to more easily and reproducibly visualize their genEpi data

1.3.6 Chapter 7: GEViTRec

Making it easier for stakeholders to generate data visualizations that are relevant to them is one component helping stakeholders to explore different visualization alternatives in a visualization design space. While the GEViT gallery tried to help stakeholders explore different visual alternatives, I found that stakeholders still needed support to connect data to possible visualizations. In particular, stakeholders were still overwhelmed by the volume, novelty, and heterogeneity of new data and were not sure what data they had available. Together with Dr. Munzner, I developed a framework for data reconnaissance and task wrangling that describes processes stakeholders undertake to understand complex and emerging data landscapes. Chapter 7 presents the GEViTRec algorithm that supports stakeholders through data reconnaissance and task wrangling processes by automatically generating data visualizations informed by the GEViT visualization design space.

The GEViTRec algorithm is the culmination of my doctoral research and brings together the threads to environmental constraints and stakeholder knowledge (Chapters 2 and 3), with existing data visualization strategies (Chapter 5), and attempts to lower the burden for exploring and generating data visualization alternatives (Chapter 6).

In the data reconnaissance and task wrangling framework, we describe four repeated phases stakeholders undertake to explore data landscapes: acquire, view, assess, pursue. GEViTRec is designed to support data reconnaissance and task wrangling by helping stakeholders generate a quick, low effort, view of their data. While there exist data visualisation recommendation systems, for example Tableau’s ShowMe [69], the Voyager Systems [128, 129], and Draco [76], these systems all require that the user provide some initial specifications for the visual encoding before the systems recommend alternative data visualizations. Building on the collective community knowledge summarized in GEViT, the GEViTRec algorithm bypasses the need for any user input beyond the datasets she wishes to visualize. Moreover, currently exist-

ing visualization recommendation systems present just a single chart type, whereas GEViTRec builds chart combinations to reveal multiple aspects of these data through shared data linkages.

The underlying GEViTRec recommendation algorithm is intended to generalize to other applications, even though here I only demonstrate its applicability to genEpi data. This algorithm is designed to integrate different data types and extends existing systems by supporting both tabular and non-tabular data. To automatically generate visualization specifications, the algorithm uses the previously described domain prevalence visualization design space to rank visualizations according to their relevance to the stakeholder. The use of such a design space also injects an awareness of domain-specific conventions into the algorithmic recommendation procedure.

Contributions for Chapter 7

Technical Contributions:

- An algorithm for automated domain-aware visualization recommendation
- Recommendation beyond tabular data and singular charts with minimal user input
- Relevance as novel and computable metric for ranking visual encodings

Domain Specific Contributions:

- GEViTRec: automatic visualization recommendation for genEpi

1.3.7 Summary of Contributions

Taken together with the methods and contributions of these research chapters demonstrate how the different sources of knowledge can contribute to a data visualization design and evaluation processes. Collectively, this knowledge is represented in different forms, as text from interviews or research articles, as binary choices of preference, as ranks, as discrete numerical counts, through typologies, and finally as visualization design spaces. Through my research projects I have drawn upon these different sources of knowledge to **understand** the present needs and limitations of stakeholders, **explore and characterize** visualization strategies, and to use this knowledge to **design and implement** tools that address the unmet needs of stakeholders. These contributions extend current data visualization practice in genomic epidemiology and work toward the improved integration of data visualization with analysis. The research approach that I have taken and its resulting technical and domain specific contributions are novel in the way that methods and techniques from multiple different disciplines are used to create and integrate new sources of knowledge. Although the work that I present here is limited to an application context within the genomic epidemiology domain, this research approach, and especially its technical contributions, can serve to inform visualization research more broadly.

Chapter 2

On Regulatory and Organizational Constraints in Visualization Design and Evaluation

Privacy is not something that I'm merely entitled to, its an absolute prerequisite. — Marlon Brando

¹ Problem-based visualization research provides explicit guidance toward identifying and designing for the needs of users, but absent is more concrete guidance toward factors external to a user's needs that also have implications for visualization design and evaluation. This lack of more explicit guidance can leave visualization researchers and practitioners vulnerable to unforeseen constraints beyond the user's needs that can affect the validity of evaluations, or even lead to the premature termination of a project. Here we explore two types of external constraints in depth, regulatory and organizational constraints, and describe how these constraints impact visualization design and evaluation. By borrowing from techniques in software development,

¹This chapter has been previously published [23]:

A. Crisan, J. L. Gardy, and T. Munzner. On regulatory and organizational constraints in visualization design and evaluation. Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization, 1:19, 2016. doi:10.1145/2993901.2993911

project management, and visualization research we recommend strategies for identifying, mitigating, *and* evaluating these external constraints through a design study methodology. Finally, we present an application of those recommendations in a healthcare case study. We argue that by explicitly incorporating external constraints into visualization design and evaluation, researchers and practitioners can improve the utility and validity of their visualization solution and improve the likelihood of successful collaborations with industries where external constraints are more present.

2.1 Introduction

Simon’s parable of *The Ant on the Beach* asks readers to consider the trajectory of an ant as it walks along a beach: “Viewed as a geometric figure, the ant’s path is irregular, complex, hard to describe. But its complexity is really a complexity in the surface of the beach, not a complexity in the ant” [110]. The parable highlights the importance of describing both the agent of action and the broader environment that acts upon that agent [118]. In problem-based visualization research and other user-centred methodologies, that agent is the user. While a focus on the user does not exclude consideration of her broader environment, little of the visualization research literature has been dedicated to precisely understanding how factors external to a user’s needs affect design and evaluation [62]. External factors can constrain the scope of the design space because, irrespective of user preferences, some solutions can never be implemented in their contextual environments. If researchers are unaware of these external factors from the project outset, they may develop and evaluate a visualization solution that cannot be used. For example, the authors of WeaVER, a tool that visualizes ensemble weather data, identified obstacles to data access, barriers of installing their visualization tool on locked-down workstations, and difficulty obtaining raw data as factors affecting their ability evaluate the tool’s design [93]. The discussion of external factors is not absent from the visualization research, but there

does not exist more explicit guidelines toward incorporating factors from a user’s contextual environment into visualization design and evaluation.

In this chapter I propose that these external factors should be modelled as **constraints** [118] that must be incorporated into visual and interaction design choices so as to yield relevant evaluations. I suggest strategies that visualization researchers can use to identify these constraints and provide recommendations for how constraints can be evaluated throughout a project’s life cycle. I also demonstrate how these suggested strategies can be practically applied by presenting a case study in a healthcare environment, where external constraints can present many challenges for visualization researchers. Finally, I layout how these constraints impact data visualization design and analysis and moreover underlay the motivations for research I present the subsequent chapters.

2.2 Defining External Constraints

We have defined **external constraints** as any factor affecting visualization design and evaluation that is separate from the user’s problem or needs and that are drawn from the user’s contextual environment. In this section, we further separate these external constraints into two broad categories – regulatory and organizational constraints. In the context of this paper, we limit the definition of regulatory and organizational constraints to data access and the use of data for research purposes, because data is central to visualization research.

Regulatory constraints refer to legal requirements governing the collection, storage, and use of data. In contrast, **organizational** constraints are policies and practices that are not necessarily encoded in law and that can vary across different institutions and across communities. Examples of organizational constraints can include policies around the protection of trade secrets, protectionist tendencies toward data, availability of financial resources, or

institutional support for visualization projects [18, 102]. Importantly, organizational constraints encompass both the *interpretation* and the *enforcement* of regulatory constraints. Differences of interpretation mean that different institutions can have different data access and use policies, some being more restrictive than others, while still conforming to the law. Although these constraints are real, they should not discourage visualization researchers from collaborating with industries where regulatory and organizational constraints are present. By being aware of these constraints throughout the project's life cycle and explicitly incorporating them into visualization evaluation, researchers can enjoy fruitful collaborations, even within highly regulated industries.

2.2.1 Implications for Evaluation

Regulatory and organizational constraints have implications for design choices, often by restricting functionality and research processes (Section 2.2.3 and 2.2.2).

As result, these external constraints provide additional parameters that need to be considered during evaluation or can define how evaluation should take place.

For example, an additional parameter that needs to be evaluated is whether the visualization solution can be accessed by users, either by being installed on their work station or through web access, or whether IT constraints prevent local installations or uploading data to a web-based interactive platform. Such considerations can be missed when evaluating solely user's needs, as users themselves may not be fully aware of these constraints, or users may be inappropriately using their personal laptops for sensitive data and may not communicate they may be in violation of regulatory or organizational constraints.

There are different consequences for failing to account for these constraints.

Failure to account for organizational constraints typically affects the validity of evaluations, whereas failure to account for regulatory constraints may have legal repercussions for a researcher and also the user. For example, ignored organizational constraints may result in project delays or termination, or a lack of adoption of the proposed solution. However, researchers who fail to account for regulatory constraints are in violation of the law and could be subject to more severe penalties that involve the legal and judicial systems.

It is thus necessary to evaluate that a project is in compliance with these external constraints throughout the project's life cycle.

2.2.2 Example: Hypothesis Generation Considered Harmful

One of the common arguments for the use of visualization is to facilitate new insights [81]; that is, to generate new, testable hypotheses from data.

However, in some highly regulated industries such as healthcare, finance, or the government, the ethics of exploring or mining data to generate new hypotheses is often controversial and is sometimes considered inappropriate or even illegal – especially for data pertaining to individual people [96].

Both regulatory and organizational constraints influence exploratory analysis and hypothesis generation.

For example, organizations that routinely mine their users' data may have internal policies limiting who can mine this data, at what level of resolution (individual-level or aggregate), what can be reported and to whom, and what data may be unacceptable to use (for example, data from minors).

In highly regulated industries, legal boundaries also affect hypothesis-generating research. For example, personal data in Europe is subject to the recently adopted General Data Protection Regulation (EU 2016/679), which provides a framework governing multiple aspects of data use, including notice of collection, specified-purpose usage, consent, security, disclosure,

access, and accountability. Failing to adhere to the regulations can cost organizations fines of up to €1,000,000.

Visualization researchers who are new to highly-regulated environments might want to launch a visualization collaboration to specifically support hypothesis generation, yet might not be aware of the organizational or regulatory constraints that apply to their data that may preclude a successful outcome. Even researchers who have successful past collaborations with industrial partners with strict organizational constraints about the necessity of keeping proprietary data from leaking to the outside world may not realize the restrictions entailed by these kinds of regulatory constraints for any unauthorized data use whatsoever, even internally.

2.2.3 Example: Agile Development Considered Harmful

Many visualization researchers advocate agile and iterative methods for visualization design and evaluation, but these approaches are often at odds with the rigid information technology infrastructure typically in place in institutions like hospitals, banks, or government agencies [36]. Moreover, concerns about the dangers of uncontrolled data exploration are frequently so central that they even extend to the realm of software development methods for tools to manipulate that data. Many organizations in highly-regulated industries remain firm in their use of waterfall software development models, despite their known problems and inefficiencies, rather than adopting more agile options [63] [38].

2.3 Prior Work

The central prior work appears both within the visualization literature and in other domains. In this section, we appraise the extent to which prior work has equipped visualization researchers to identify, incorporate, and evaluate external constraints.

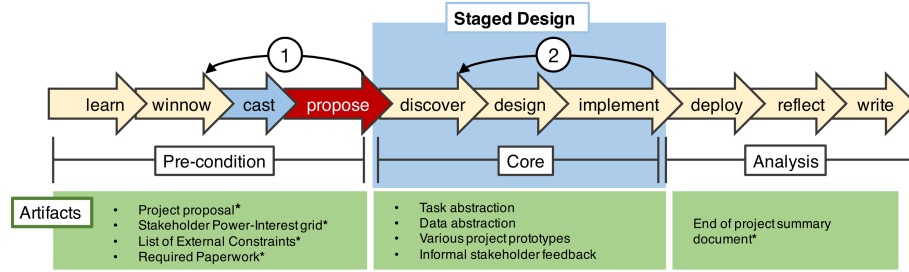


Figure 2.1: Summary of our proposed additions to the Design Study Methodology [103]: changes to the cast stage, a new propose stage, the generation of the starred artifacts, and identifying two of the many possible checkback cycles as required rather than optional.

2.3.1 Visualization Methodologies

The visualization research literature sets forth a number of models and methods to approach problem-driven design and evaluation projects [74]. These include contributions from our own group – the Nested Model (NM) for Design and Validation [78], the follow-on Nested Blocks and Guidelines Model (NBGM) [73, 74], and a Design Study Methodology (DSM) [103] – and others, including Multi-dimensional In-depth Long-term Case studies (MILCs) [108] and the Human-Centered Design Cycle [66]. A central tenet of problem-driven research has been an emphasis on the needs of the target users and evaluating visualization design choices with respect to those needs. The “domain problem” of the NM or the “domain situation” of the NBGM, and also more recent work by Winters [126] to further characterize domain situations via the NBGM through a new conceptual framework, *could* be interpreted to include external constraints, but guidance is primarily offered toward identifying and evaluating user needs. Similarly, the DSM and MILC approaches acknowledge the importance of considering the broader context in which visualization tools are deployed, but we argue they do not sufficiently address external constraints. A small number of design studies and commentaries of design and evaluation methodologies have considered external constraints within the context of visualization

research. A study of large automotive companies warned of obstacles that are separate of “technical challenges but [include] political or organizational requirements” [102]. The authors suggested conducting pre-design studies to understand these factors in order to identify a feasible project path – a sentiment that was shared in a position paper on pre-design empiricism [12]. Both Brehmer [12] and Sedlmair [102] advocate for a variety of evaluation techniques at different design stages, with the thrust of their discussion focusing on a common agile motto “test early, test often”. Another study by Lam [62] uses a *scenario based* approach to evaluating visualization solutions that includes understanding environment and workplace practices, which they and others note is understudied in visualization research. Aside from identifying these evaluation scenarios through a literature review of visualization research, Lam *et. al* [62] do not provide more detailed guidance towards the the types of external constraints or how they may be identified and evaluated. The lack of explicit guidance toward evaluating visualization design with respect to external constraints means that individual researchers must devise strategies on an *ad hoc* basis, which some researchers may be more successful at than others.

2.3.2 External Disciplines

The design and evaluation of a system in the context of regulatory and organizational constraints is not unique to the domain of visualization research or practice. Some of the techniques used in visualization design studies are drawn from the larger set used in agile software development and related project management practices. For many visualization research projects, applying the complete set of agile methodologies and practices may be inappropriate – they do not capture some of the unique nuances of the visualization discipline and the agile framework can be too comprehensive and prescriptive for smaller, informal projects. However, for large, formal collaborations in industries where the external constraints are much more pronounced, certain agile techniques from the software development liter-

ature can be useful. In Section 2.4.3, we discuss specific techniques from the broader domain of agile software development that may be applicable toward design and evaluation of external constraints for highly-regulated environments.

Cognitive Work Analysis (CWA) is another broad framework frequently deployed in developing technologies for the workplace, especially where regulations or safety are paramount considerations [118]. Its roots lie in systems thinking and ecological psychology, and it takes the most holistic view of a user and their contextual environment. A subset of CWA methods are frequently harnessed for visualization design and evaluation, particularly for task analysis. Importantly, CWA advocates undertaking a “work domain analysis” to understand a user’s context because “it imposes constraints on the actions of the actors” [118]. Collectively, the agile and CWA literatures offer a number of strategies for identifying and mitigating external constraints, but these strategies will be most useful only when appropriately contextualized for the visualization research domain.

2.4 Guidelines for Evaluating External Constraints

We argue that the best way to mitigate external constraints is to proactively seek to identify them as early as possible, and to follow up by assessing whether they have been met as part of *formative evaluation* efforts throughout the project’s life cycle. We use the Design Study Methodology (DSM) [103] as a scaffold to provide specific recommendations to visualization researchers. We propose additional stakeholder roles within the cast stage and explicit communication strategies with them. We advocate the creation of several artifacts at many points, including at a new stage where a formal proposal is generated as part of a formative evaluation to assess project feasibility. These artifacts serve as checkbacks to specific previous stages, in contrast to the original DSM that simply encourages researchers to return to any prior stage of the framework as needs are noticed. We also

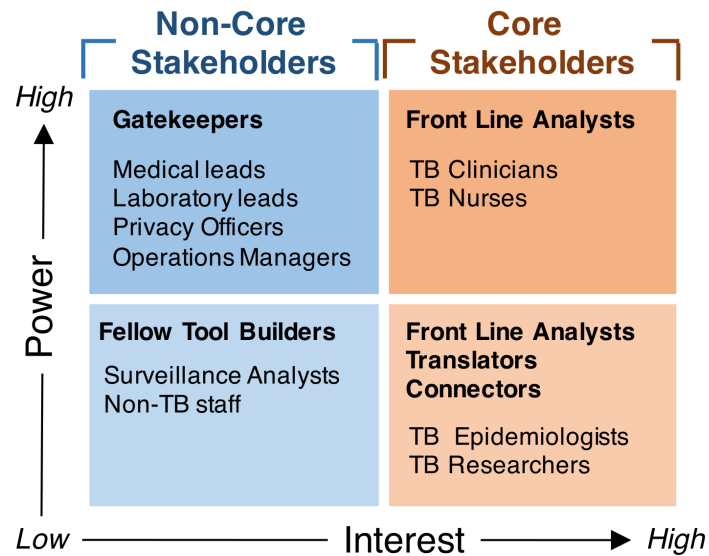


Figure 2.2: *Power Interest Matrix for identifying detailed roles during the cast phase.* Stakeholders are categorized into core and non-core groups according to their interest in project outcomes, and also as having low or high power. The specific roles identified in Section 2.5 are included here as a concrete example.

argue for specific methods including a staged design process with generation of synthetic data as a stepping stone for access to the real data. Figure 2.1 presents a summary of these recommendations.

2.4.1 Defining Stakeholder Roles

The cast stage of the DSM pre-condition phase recommends that collaborators be cast as acting in one or more of several possible specific roles to help researchers identify the ways that relevant stakeholders might become involved in a project: front-line analyst, gatekeeper, translator, connector, or fellow tool-builder.

Recommendation 1: *Classify stakeholders according to power over and interest in project outcomes.* We argue that this classification should be

extended to further improve stakeholder identification and management: these roles should be further stratified according to the amount of power over and interest in the project outcomes for each stakeholder. Using a power-interest grid can help identify stakeholders – particularly gatekeepers – that may not be immediately obvious; for example, individuals who are not directly involved in a project but who can affect the project through their role in assuring compliance with regulatory or organizational constraints. Stakeholders that have high interest in a project’s outcomes, whether low- or high-power, typically form a core group with whom researchers closely collaborate; these core stakeholders will be actively involved in visualization design and evaluation (both formative and summative) and they also supply the motivation and needs for a visualization solution. Indeed, a visualization project may be initiated through these high-interest stakeholders.

Here, we do not prescribe a specific type of formative evaluation methodology, but note that much of the evaluation studies proposed in visualization research, including interviews, questionnaires, think-out-loud, and laboratory experiments, are targeted toward these core stakeholders. Non-core stakeholders are those with whom researchers do not collaborate directly and who are thus classified as low-interest. Often, stakeholders with high power, but low interest in project outcomes are those that must be consulted with in order to access data and get approval to conduct the research; the DSM classifies these stakeholders as Gatekeepers. Gatekeepers can be individuals that oversee the appropriate access and use of data, both at the outset and throughout a project, or an institutional review board that provides initial approval for data access and use. While this quadrant of high-power, low-interest stakeholders are unlikely to participate in visualization design processes, individual gatekeepers (but not entire review boards) should be included in *at least* guideline checking formative evaluations [3], to confirm compliance with regulatory and organizational constraints. Finally, there are stakeholders with low interest and low power in visualization project outcomes. These individuals may have an intellectual interest in project outcomes, such as

other researchers building analytical tools; while these individuals will not take part in either design or evaluation they may form useful allies in the institution and inform researchers about external constraints.

Recommendation 2: Actively manage communication with stakeholders.

While DSM does indicate that poor rapport can be a potential pitfall to a project's success (PF-9) [103], it does not provide explicit guidance towards managing communication with stakeholders. Ineffectively managing stakeholder communications can impact the discovery of regulatory or organizational constraints, which in turn impacts the validity of evaluations and could even lead to premature termination of the project. Good communication with stakeholders is also critical for carrying out formative evaluations with core stakeholders and guideline checking evaluations with gatekeepers of prototypes developed through the staged design process (Section 2.4.3).

We recommend using the power-interest grid of Recommendation 1 as the framework for managing stakeholder communications. For core stakeholders, communication can be informal and will be more frequent than with non-core stakeholders. For non-core stakeholders with high power over a project's outcomes, we recommend more formal communication. Some institutions will already have policies in place for communication templates and the timeliness of those communications, but when such guidelines are not available, we recommend a formal, plain-language brief that is distributed to these stakeholders. These briefs may be more frequent at the beginning of the project, especially if there is uncertainty around the nature of agile development methods and the design study framework, and may become less frequent over time. Communication briefs should emphasize the findings of evaluations carried out during the design process. Effective communication with stakeholders can have the added benefit of improving institutional awareness of visualization research, which may make future projects easier to conduct.

2.4.2 Generation of Additional Artifacts

Conducting pre-design studies [103] [12] to assess a project’s feasibility and to identify regulatory and organizational constraints is important. In its original form, the DSM recommends going directly from the cast stage to the discover stage of the core phase, but we argue that this transition is premature and recommend an explicit propose stage between the two. This new propose stage entails creating additional project artifacts that help to guide formative evaluations of user needs, in addition to identifying regulatory and organizational constraints. These artifacts are in addition to the task and data abstractions and the prototypes that already form part of the DSM’s core phase.

Recommendation 3: Create a formal proposal document. The most important of these artifacts is a project proposal that summarizes the evidence gathered in the pre-design studies and consultation with high-power stakeholders into a single document. This proposal document should be assessed by both researchers, core stakeholders, and Gatekeepers, before proceeding to the core phase of the DSM.

Throughout various stages of formative evaluation during the design process, this document can serve as the basis for the guideline checking that will be carried out with Gatekeepers [3]. Institutions may have specified proposal templates but if a proposal template does not exist, we recommend communicating –at minimum – the project’s scope, including user needs, known external constraints, data requirements and uses, who is involved and what they will be doing, and a brief description of the design process and evaluation procedures. This proposal will typically be refined through a process of discussion with stakeholders. Although the DSM encourages researchers to backtrack to any of the proceeding steps without requiring any checkback loop explicitly, our extension proposes that the completion of the final proposal document should trigger a *required* revision of the winnow

stage, as shown by arrow 1 in Figure 2.1. The goal is to evaluate whether the project can be completed in a timely manner and is mutually beneficial to stakeholders and researchers.

Recommendation 4: Create a summary document at the end of a project.

At the end of the project we recommend creating a summary document that expresses – in plain language – the ways in which the project addressed a relevant domain problem in light of external constraints. The project conclusion document is meant to complement the initial project proposal by highlighting the resulting mutual benefits of the project for both researchers and stakeholders. A research paper describing the project outcomes in terms suitable for an academic audience of other researchers who grapple with visualization design and evaluation issues is not a suitable stand-in for this document, which is aimed at a very different audience with different concerns. In some cases, the process of abstraction that was undertaken by the visualization researcher needs to be inverted so that the solution can be described in domain-specific terms in a way that makes sense to the intended audience. However, this conclusion document can be helpful for educating stakeholders on the processes and relevance of visualization research [66], especially if the document emphasizes how the results of various evaluation studies are in line with individual stakeholder needs and also institutional policies. It has the potential of laying out important groundwork so that future visualization research projects are easier to conduct.

2.4.3 Methods

Once researchers and stakeholders have an understanding of users needs as well as external constraints, both should be integrated into the visualization design and evaluation process.

Recommendation 5: Use a staged design process. The staged design

model [71] proposes incremental prototype development through a series of stages, making it possible to progressively gain access to users and resources that may not be accessible at a project's outset and to accommodate changes in the stakeholders' context and environment that arise over a project's life cycle. Each stage consists of requirements-gathering and prototype development to produce a minimum viable product with progressively improving fidelity. The model should conclude with a formal evaluation that specifically demonstrates whether the tool and development process is in compliance with regulatory and organizational constraints, in addition to meeting stakeholder needs.

At the end of each design stage, we highly recommend that researchers and collaborators explicitly evaluate together whether or not it is feasible to proceed to the next stage of development, as shown by arrow 2 in Figure 2.1. By proactively checking on feasibility in this way, initially unforeseen constraints that arise later in the project are surfaced as early as possible, to minimize later adverse impact on researchers such as a loss of access to data or people. Using a staged design process also allows researchers to plan and prioritize minimal viable products, some of which may be valid visualization research contributions in themselves – even if a project is terminated ahead of the originally planned schedule.

Recommendation 6: Use synthetic data early on if real data is not immediately available. As discussed in Section 2.2.2, some industries have concerns around hypothesis-generating research related to both the agile design process and the types of insights that can and should be drawn from data.

Stakeholders in these industries may *want* visualization tools to support hypothesis generation of individual level data, but nevertheless may wish to impose limits on the types of uncontrolled exploration a user can conduct [96]. At a project's outset, it may not be clear yet how to operationalize

such limits, which puts researchers and stakeholders in the difficult position of potentially violating regulatory constraints. These constraints can make collaborators wary of sharing real data at a project's outset, thus impeding the launch of a potential collaboration with visualization researchers. One way to overcome this constraint is to use synthetic data in early design stages and gradually earn the trust necessary to gain access to real data in later stages. Synthetic data is never a perfect substitute for real data because it lacks nuances that may be of interest of stakeholders; consequently, the use of synthetic data affects the validity of evaluations of a prototype's utility. For example, synthetic data is often very clean, avoiding the problems of missing or erroneously entered data that are often present in real data; while such noise can be simulated, the scope of possible errors may be difficult to fully understand and incorporate in synthetic data generation. The nuances of supporting users in handling dirty data might therefore be absent from a design process and evaluation process where only clean data is used. In spite of these limitations, synthetic data can nevertheless an effective means to demonstrate a tool's functionality and to allow researchers and stakeholders to have concrete discussions about what aspects of functionality should be limited.

By graduating from synthetic to real data and modifying the rigor of evaluations over time, what may be lost in initial evaluation validity can be gained in collaborators' trust. Starting with synthetic data can be a viable alternative to giving up on the project during early stages due to initial regulatory and organizational constraints.

2.5 Case Study: Healthcare

In this section we provide a concrete example of how to interpret our recommendations through a case study in healthcare, in an approach similar to Winters *et al* [126]. Case studies provide an opportunity to dive deep into a specific domains to provide insights into a phenomenon that may be trans-

ferrable to other domains [37], and their benefits for visualization research has been argued by Shneiderman and Plaisant in their ethnographically informed proposal for multi-dimensional in-depth long-term case studies (MILCs) [108].

Healthcare systems comprise two disciplines – clinical medicine and public health – that must work together to improve the health of both individuals and populations. Public health focuses on prevention and control activities, while clinical medicine focuses on diagnosis and treatment [51]. While clinical medicine tends to be the domain of specialist health care providers such as clinicians, nurses, and pharmacists, public health professionals are more diverse. In addition to the aforementioned providers, their roles include, but are not limited to, epidemiologists, statisticians, researchers, politicians, and other community leaders.

In some cases these two disciplines can operate nearly independently of one another, but in others they must work more closely together to deliver patient care. The world of communicable disease prevention and control is an example of the latter, where disciplines must share knowledge and make decisions together – clinicians guide the management of individual patients with a disease, while public health authorities manage the disease at a population level. Although they must work together, the different traditions informing public health and clinical medicine mean that there is often a knowledge translation gap, where the knowledge and data generated by each discipline is siloed, ultimately affecting the ability of these disciplines to work together [131].

Visualization tools can help stakeholders in public health and clinical medicine to more readily share knowledge and insights that support decision making at patient and population levels. But in order to be most effective, visualization researchers need to operate within the bounds of the significant regulatory constraints that apply to healthcare and healthcare data, as well as the organizational constraints in healthcare, which can differ between public

health and clinical medicine.

2.5.1 Constraints in Healthcare

Regulatory Constraints. The law distinguishes between primary and secondary use of health data [98]. Primary uses of health data are those associated with the direct and immediate care of a patient, while secondary uses are all other uses that do not directly contribute to a patient's care. This category includes all research using health data. While the law does not prevent the secondary use of health data, it does place restrictions on such usage that are meant to balance an individual's right to privacy and confidentiality while simultaneously stimulating progress in public health and clinical medicine. Oversight and implementation of these regulatory constraints is not consistent across different institutions [98].

Organizational Constraints. It is recognized that the secondary use of health data is a ubiquitous and necessary practice, but data access models vary considerably and are not transparent, which affects research productivity [98]. Many institutions are wary of uncontrolled secondary use of data [96], in which any researcher can explore any manner of hypothesis in a dataset without clear benefit to the patient. While exploratory hypothesis-generating research is important, it is a hotly debated as a practice because it is ultimately the patient, and not the researcher, that bears the full burden of accidental data disclosure.

Researchers who request access to health data are often required to have a well-formed hypothesis at the project outset, in addition to outlining their analytical methods. As was discussed in Section 2.2.2, these restrictions on hypothesis-generating research affect not only the functionality of data visualization tools, but also the application of agile-like methods for developing them. Aside from organizational practices that enforce regulatory constraints, there also exist hierarchical and political structures that can result in protectionist tendencies toward data. These protectionist tendencies

can arise because a particular individual is responsible for stewarding the appropriate use and interpretation of health data, or because researchers are hesitant to share data that was costly and time-consuming to obtain.

2.5.2 Lessons Learned in Developing a TB Decision Support Tool

Our proposals for integrating constraints into the visualization design and evaluation grew out of a specific project in a highly-regulated healthcare domain.

Application: Tuberculosis Prevention and Control. Of the many communicable diseases managed by a public health agency, tuberculosis (TB) is one of the most interesting. It has a long history of infecting humans, with TB found in the remains of mummies and tales of “consumption” a popular theme within popular culture [29]. Despite this long history, medicine has not yet succeeded in eliminating TB. In 2012 alone, there were 8.6 million new cases of symptomatic TB and 1.3 million deaths worldwide, and as much as 1/3 of the world’s population is thought to be infected with a latent, asymptomatic form of the disease [132]. New strategies to manage existing cases and prevent future ones are clearly needed. Opportunities for designing and delivering new interventions to combat TB are available through exploring and mining patient-level data in electronic health records, population-level data in disease registries, and even molecular data describing pathogenic microbes rather than human individuals.

Collaboration Context. We report and reflect upon a collaboration with stakeholders involved in TB prevention and control at the British Columbia Center for Disease Control (BCCDC). Our goal was to build a decision support tool to facilitate our users’ routine workflows and to allow exploratory analysis in support of new intervention development. We did not set out to construct a fail-safe healthcare application; rather, we set out to collabo-

ratively explore how visualization of our stakeholders' data could support decision making. At the start of our collaboration, armed solely with existing visualization design guidelines, we were often reacting to previously unknown regulatory and organizational constraints rather than proactively mitigating them – and at one point faced the risk that the project would not move forward.

At the outset of our collaboration, we engaged with a small group of stakeholders at the BCCDC that consisted of clinicians, nurses, epidemiologists, and researchers. This group had worked together extensively in the past, and had a history of productive prior research collaborations. We engaged in discussions with them about a project that explored the utility of data visualization to provide multiple perspectives on the spread of TB through the province of British Columbia over time. The insights this group of stakeholders would gain from the tool would help inform future policies and practices in TB prevention and control. Our discussions around the project and its objectives were informal, and the data we had intended to use for tool development had received prior approval for research use. With a promising collaboration on our hands, we began to engage in discussion with these stakeholders about the data types in use at BCCDC and the ways our stakeholders used these data for both routine and high-level policy decision-making.

Discovering Lurking Constraints. While we were focused on assessing our stakeholders' needs and their primary research question, we confronted the first regulatory and organizational constraints that would temporarily suspend our project's progress. Over the course of our project, the BCCDC had changed the way it gathers Public Health data, and how use of this data for research was to be governed. Not only were data approval policies changing, but so too were the individuals responsible for the approvals (referred to internally as data stewards). As part of taking on their mandate, the new TB data steward took stock of current research projects, and flagged

our visualization project for re-assessment. His concern was that the project did not clearly outline how it may be directly beneficial to patients and so had the potential to be deemed unethical. Although an ethics committee had reviewed and approved the use of our data for secondary purposes, the new data steward indicated that we needed to provide a more detailed justification for our specific project before we could continue.

Identifying Additional Gatekeepers. Neither we nor our collaborators had anticipated this intervention by the data steward. As we began to gather information about the necessary next steps to take in order to continue our project, we sought to understand other aspects of the organizational structure and identify other gatekeepers that might further impede our project's progress. We took on the exercise of creating a power-interest grid (*Recommendation I*) and over time we stratified our TB stakeholder group as follows:

- **High Interest, High Power** *Front-line Analysts (TB clinicians and nurses)*: Data for individual patients was primarily collected and controlled by and accessed through clinicians and nurses. With a strong interest in using data to develop new policy and practice, these individuals formed part of our core stakeholder group.
- **Low Interest, High Power** *Gatekeepers (Departmental Medical Leads, Laboratory Leads, Privacy Officers, and Operations Managers)*: Both medical and laboratory leads must sign off on data usage, though they may not be directly involved in TB control or invested in our project outcomes. Privacy officers and operations managers also enforce regulatory processes. One particularly powerful, but difficult to reach, stakeholder was the organization's IT department, as they controlled the users' workstations and permissions for software installation. These individuals did not form part of our core stakeholder group.
- **High Interest, Low Power** *Front-line Analysts and Connectors (TB*

epidemiologists and researchers): In our study, researchers had control over the use of the pathogen-level molecular data they had generated and epidemiologists could advise us on the use of patient-level case data, but neither class of stakeholder had the authority to sign off on data usage beyond the molecular data. Still, as integral parts of the TB control team, they were interested in our project outcomes, and were part of our core stakeholder group.

- **Low Interest, Low Power Fellow Tool Developers (Non-TB analysts):** Other groups around the BCCDC were interested in visual analytic tools for their own applications outside of TB, but were outside of our core stakeholder group.

We established a rough communication plan (*Recommendation 2*) to engage with these stakeholders in order to proactively identify important constraints moving forward. Often our communications were one-on-one discussions, but when availability afforded it, we conducted large group meetings with both core and non-core stakeholders.

Finding Constraint Impact on Functionality. As we identified different stakeholders, we learned of more organizational constraints that would affect the functionality of the decision support tool we intended to build. We learned that our tool should not support what might, at first glance, seem to be obviously useful data wrangling functionality such as merging multiple datasets, correcting data errors, or entering missing data. There were institutional policies in place that governed how and by whom multiple datasets could be merged because of concerns around privacy – as more datasets are linked together, there is a higher likelihood of potentially re-identifying patients. Furthermore, institutional procedures were also in place to correct errors or handle missing data in a systematic way, and again were carried out only by select individuals.

Given the constraints that precluded data wrangling, we recognized that our

tool would function best as a data viewer that could alert core stakeholders to missing or incorrect data but not permit them to change the underlying dataset. Furthermore, our tool needed to flexibly handle whatever data and data types different core-stakeholders were permitted to access, ranging from clinicians and nurses allowed to access individual patient data, to generalist users who should only be shown aggregate data. These additional requirements affected functional requirements and served to constrain our design space.

Finding Constraint Impact on Real Data Access. Some stakeholders unfamiliar with the design process considered it odd that we had not already established the visual and interaction design choices for our decision support tool. They also found it unusual that we intended to conduct a research project to figure out what those design choices should be. Thus, our research was initially perceived by some as an uncontrolled use of secondary data (Section 2.2.2), and several Gatekeepers were unwilling to allow us to use real data at the outset. We thus considered at length how to develop a strategy that would gain these users' trust in our research methods.

Finding Constraint Impact on Tool Integration. Through several stakeholders, we also learned about the impact of the information technology (IT) group's policy that workstation environments should be locked down. A lengthy approval process was required to install new software or host custom web applications on institutional servers. Accessing web applications for data analysis was also prohibited because data could not leave institutional servers. Part of the reason for these constraints is that the IT group manages workstations in many healthcare settings, including not only research workstations but also those used in clinical care, resulting in very restrictive workstation policies.

One tool that could be used in the existing constrained environment – and indeed was widely used by BCCDC epidemiologists – was R. Although

the version of R available on workstations was outdated and could only be updated by IT, we knew there were plans to update it, and decided that a R-based tool would be a viable implementation solution that fit into BCCDC's existing organizational infrastructure.

Changing Strategies for Emerging Constraints. The identification of these constraints and our assessment of their impacts on our decision support tool's functionality, utility, and stakeholder adoption allowed us to reformulate our project's trajectory. We prepared a project proposal for our core-stakeholders and gatekeepers that outlined clearer objectives for our tool in light of the various regulatory and organizational constraints we identified (*Recommendation 3*). Importantly, we also indicated how stakeholders would be involved in evaluating our compliance with these constraints.

Building Trust Through Staged Design. We planned for a staged design process based upon different datasets (*Recommendation 5*).

The first stage of design would use only data available in routinely collected administrative datasets, while later stages would combine this data with laboratory and contact network (who was exposed to an infectious individual) datasets. In this way, we would produce minimum viable products for the most commonly used dataset first, and less commonly used datasets later. Although we could not use the real data, we had access to the structure and aggregate statistics of the real datasets because they were made public through BCCDC's annual reports. As much as we were able to, we based our synthetic datasets off of the real data (*Recommendation 6*). We hypothesized that if stakeholders were enthusiastic about how the decision support tool could visualize their most commonly used dataset, albeit as demonstrated by synthetic data, that this demonstration may encourage them to move toward using the tool with real data.

We conducted focus groups and developed paper prototypes during the first

design stage to gather user requirements and marry those to known regulatory and organizational constraints. The inability to install our tool on stakeholder workstations led us to rely on chauffeured demos [66], using a workstation with a more current version of R, to conduct evaluations at the conclusion of the design stage. We gathered qualitative evaluations of the tool’s perceived utility and the validity of our design choices. To evaluate compliance with regulatory and organizational constraints, we worked closely with BCCDC’s privacy officer (Guideline checking evaluations).

Although not rigorous, our evaluation gave stakeholders an opportunity to see what a decision support tool that visualizes TB data could do and how it could help them. Furthermore, instead of discussing abstract notions of how this tool may or may not be beneficial to patients in the long term, we could engage in more concrete discussion with stakeholders – especially Gatekeepers – about what functionality was appropriate and what was not. We summarized the design and evaluation progress, highlights, and outcomes of our collaboration at a larger group meeting following the conclusion of the first design stage. We emphasized how a visualization tool could responsibly incorporate regulatory and organizational constraints that are meant to safeguard patient data, and demonstrated this capacity by emphasizing the results of formative evaluations that various stakeholders had participated in. The success of this initial stage has initiated concrete discussion by both core–stakeholders and gatekeepers toward evaluating the tool using real data. Thus, what could have been a failed start due to unforeseen initial constraints has evolved into a viable project with organizational support for its continuation.

Chapter 3

Evidence Based Design:

Applying a Design Study Methodology to the Redesign of a Whole Genome Sequencing Clinical Report

Design is not just what it looks like and feels like. Design is how it works
— Steve Jobs

¹ Microbial genome sequencing is now being routinely used in many clinical and public health laboratories. Understanding how to report complex genomic test results to stakeholders who may have varying familiarity with genomics including clinicians, laboratorians, epidemiologists, and researchers is critical to the successful and sustainable implementation of this new technology; however, there are no evidence-based guidelines for designing such a report in the pathogen genomics domain. Here, we describe an iterative, human-centered approach to creating a report template for communicating tuberculosis (TB) genomic test results. We used Design Study Methodology (DSM) a human centered multi-stage approach drawn from the information visualization domain to redesign an existing clinical report. We used expert consults and an online questionnaire to discover various stakeholders needs around the types of data and tasks related to TB that they encounter in their daily workflow. We also evaluated their perceptions of and familiarity with

¹This chapter has been previously published [25]:

A. Crisan, G. McKee, T. Munzner, and J. L. Gardy. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. PeerJ, 6:e4218, Jan. 2018. doi:10.7717/peerj.4218

genomic data, as well as its utility at various clinical decision points. These data shaped the design of multiple prototype reports that were compared against the existing report through a second online survey, with the resulting qualitative and quantitative data informing the final, redesigned, report. We recruited 78 participants, 65 of whom were clinicians, nurses, laboratorians, researchers, and epidemiologists involved in TB diagnosis, treatment, and/or surveillance. Our first survey indicated that participants were largely enthusiastic about genomic data, with the majority agreeing on its utility for certain TB diagnosis and treatment tasks and many reporting some confidence in their ability to interpret this type of data (between 58.8% and 94.1%, depending on the specific data type). When we compared our four prototype reports against the existing design, we found that for the majority (86.7%) of design comparisons, participants preferred the alternative prototype designs over the existing version, and that both clinicians and non-clinicians expressed similar design preferences. Participants articulated clearer design preferences when asked to compare individual design elements versus entire reports. Both the quantitative and qualitative data informed the design of a revised report, available online as a LaTeX template. We show how a human-centered design approach integrating quantitative and qualitative feedback can be used to design an alternative report for representing complex microbial genomic data. We suggest experimental and design guidelines to inform future design studies in the bioinformatics and microbial genomics domains, and suggest that this type of mixed-methods study is important to facilitate the successful translation of pathogen genomics in the clinic, not only for clinical reports but also more complex bioinformatics data visualization software.

3.1 Introduction

Whole Genome Sequencing (WGS) is quickly moving from proof-of-concept research into routine clinical and public health use. WGS can diagnose infections at least as accurately as current protocols [40, 67], can predict

antimicrobial resistance phenotypes for certain drugs [11, 85, 120] with high concordance to culture-based testing methods, and can be used in outbreak surveillance to resolve transmission clusters at a resolution not possible with existing genomic or epidemiological methods [80]. Importantly, WGS offers faster turnaround times compared to many culture-based tests, particularly for antimicrobial resistance testing in slow-growing bacteria.

As reference microbiology laboratories move towards accreditation of WGS for routine clinical use, the community is turning its attention toward standardization developing standard operating procedures for reproducible sample handling, sequencing, and downstream bioinformatics analysis [13, 43]. Reporting genomic microbiology test results in a way that is interpretable by clinicians, nurses, laboratory staff, researchers, and surveillance experts and that meets regulatory requirements is equally important; however, relatively little effort has been directed toward this area. WGS clinical reports are often produced in-house on an *ad hoc*, project-by-project basis, with the resulting product not necessarily meeting the needs of the many stakeholders using the report in their clinical and surveillance workflows.

3.1.1 Human-Centered Design in the Clinical Laboratory

The information visualization, human-computer interaction, and usability engineering fields offer techniques and design guidelines that have informed bioinformatics tools, including Disease View [32] for exploring host-pathogen interaction data and Microreact [4] for visualizing phylogenetic trees in the context of epidemiological or clinical data. Although the public health community is beginning to recognize the potential role of visualization and analytics in daily laboratory workflows [18] these techniques have not yet been applied to routine reporting of microbiological test results. However, work from the human health domain particularly the formatting and display of pathology reports, where standardization is critical [64] sheds light on the complex task of clinical report design.

Valenstein reports four principles for organizing an effective pathology report: use headlines to emphasize key points, ensure design continuity over time and relative to other reports, consider information density, and reduce clutter [114], while Renshaw *et al.* note that when pathology report templates were reformatted with numbering and bolding to highlight required information, template completion rates rose from 84 to 98% [95]. Fixed, consistent layout of medical record elements, highlighting of data relative to background text, and single-page layout improve clinicians ability to locate information [82], while information design principles, including visually structuring the document to separate different elements and organizing information to meet the needs of multiple stakeholder types, can reduce the number of errors in data interpretation [133].

Work in the electronic health record (EHR) and patient risk communication domains has also provided insight into not just the final product but also the process of effective design. Through quantitative and qualitative evaluations, research has shown that some EHRs are difficult to use because they were not designed to support clinical tasks and information retrieval, but rather data entry [133]. Reviews of the risk communication literature note that, while many visual aids improve patients understanding of risk [136], the design features that viewers preferred namely simplistic, minimalist designs were not necessarily those that led to an accurate interpretation of the underlying data [2]. Together, these gaps indicate a need for a human-centered, participatory approach iteratively incorporating both design and evaluation [53, 54].

3.1.2 Collaboration Context COMPASS-TB

The COMPASS-TB project was a proof-of-concept study demonstrating the feasibility and utility of WGS for diagnosing tuberculosis (TB) infection, evaluating an isolate's antimicrobial sensitivity/resistance, and genotyping the isolate to identify epidemiologically related cases [85]. On the basis of

Mycobacterium Whole Genome Sequencing Report from MGIT Positive Samples

Not for diagnostic use

01/02/1915

Sample Details			
Sequencing Location	Oxford	Date received in Lab	
Local Lims Specimen ID	123456789	Run date	01/01/19150115
Guid	123456-79aab-910abr-15243hg		

Organism Identification	
Predicted/closest match	
TBCOMP/microti	100%
TBCOMP	100%
TBCOMP/TB	96.77%
TBCOMP/tuberculosis-canettii	35.71%
MACCOMP	21.21%

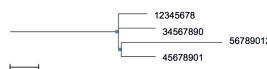
Sample/Sequencing Quality			
Total reads (~millions)	Mapped %	No reads mapped (~millions)	Coverage %
4.73	99.47	4.7	91.99

Resistance Summary						
INH	RIF	EMB	PZA	QUI	SM	AG
U	S	S	S	S	S	S

Resistotype					
Drug	Mutation	Nucleotides	Support (ACGT)	Source – (RTTotal)	Prediction
INH	katG_A727T	GCC->ACC	(160/0/1/0) (0/164/0/0) (0/167/0/0)	Unclassified	UNK

Relatedness			
NB: This data may be added or updated at a later date			
Nearest neighbour(s)			
Sample -Plate Name	Date received in Lab	Centre	No. of SNPs apart
123456789		Oxford	0
34567890	1900-01-01		10
45678901	1015-01-31	Oxford	15
56789012		London	8

The alignment width is 285. Multiply this number by the tree metrics.



Comments

Authorised
Signature: _____ Print name: _____
Position: _____ Date: _____

Figure 3.1: An initial COMPASS-TB report design.

COMPASS-TB's results, Public Health England (PHE) has implemented routine WGS in the TB reference laboratory [89]; however, this requires changing how mycobacteriology results are reported to clinical and public health stakeholders. The COMPASS-TB pilot used reports designed by the project team, but as clinical implementation within PHE progressed, team members expressed an interest in redesigning the report (Figure 3.1) to facilitate interpretation of this new data type and align laboratory reporting practices with the needs of multiple TB stakeholders.

We undertook a mixed-methods and iterative human-centered approach to inform the design and evaluation of a clinical TB WGS report. Specifically, we chose to use Design Study Methodology [103] an approach adopted from the information visualization discipline. When using a Design Study Methodology approach, researchers examine a problem faced by a group of domain specialists, explore their available data and the tasks they perform in reference to that problem, create a product in our case a report, but, in

the more general case, a visualization system to help solve the problem, assess the product with domain specialists, and reflect on the process to improve future design activities. Compared to an *ad hoc* approach to design, Design Study Methodology engages domain specialists and grounds the design and evaluation of the visualization system in tasks in this case TB diagnosis, treatment, and surveillance as well as data. It is this marriage of data and tasks to design choices, informed by real needs and supported by empirical evidence, that results in a final product that is relevant, usable, and interpretable.

Here we describe our application of design study methodology to the COMPASS-TB report redesign. Targeting clinical and public health stakeholders with at least some familiarity with public health genomics, we show how evidence-based design can be incorporated into the emerging field of clinical microbial genomics, and present a final report template, which may be ported to other organisms. We also recommend a set of guidelines to support future applications of human-centered design in microbial genomics, whether for report designs or for more complex bioinformatics visualization software.

3.2 Materials and Methods

3.2.1 Overview of Design Study Methodology

The Design Study Methodology [103] is an iterative framework outlining an approach to human-centered visualization design and evaluation. It consists of three phases Precondition, Core Analysis, and Reflection that together comprise nine stages. The Precondition and Reflection phases focus on establishing collaborations and writing up research findings, respectively, and are not elaborated upon further here. We describe our work within each of the three stages of the Core Analysis phase: Discovery, Design, and

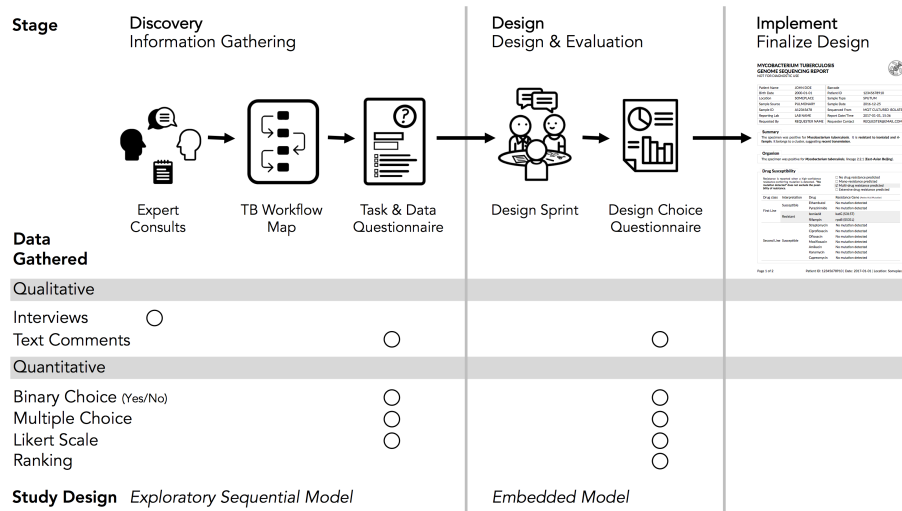


Figure 3.2: Our human-centered design approach. The Core Analysis phase of the Design Study Methodology consists of Discovery, Design, and Implementation stages. Using this methodological backbone, we collected and analyzed data using mixed-methods study designs in the Discovery and Design stages, which informed the final TB WGS clinical report design.

Implementation (Figure 3.2). We define *domain specialists* in this case as the TB stakeholders - clinicians, laboratorians, and epidemiologists - who regularly use reports from the reference mycobacteriology laboratory in their work.

Our research was reviewed and approved by the University of British Columbias Behavioural Research Ethics Board (H10-03336). All data were collected through secure means approved by the university and were de-identified for analysis and sharing. Anonymized *quantitative* results from each of the surveys and the analysis code are available at <https://github.com/amcrisan/TBReportRedesign> and in Appendix A. We also provide the full text of our survey instruments in Appendix A.

3.2.2 Discovery Stage

In the Discovery stage, we first gathered qualitative data through expert consults to identify the data types used in TB diagnosis, treatment, and surveillance tasks; we then gathered quantitative data through an online survey to more robustly link particular data types to specific tasks. This staged approach to data gathering is known as the exploratory sequential model [22].

Our expert consults took the form of semi-structured interviews with seven individuals recruited from the COMPASS-TB project team, the British Columbia Centre for Disease Control (BCCDC), and the British Columbia Public Health Laboratory (BCPHL). The interview questions served as prompts to structure the conversation, but experts were free to comment, at any depth, on the different aspects of TB diagnosis, treatment, and surveillance. We took notes during the consults in order to identify the tasks and data types common to TB workflows in the UK and Canada, as well as to determine which tasks could be supported by WGS data.

Informed by the expert consults, we drafted a Task and Data Questionnaire (text in Appendix A) to survey data types used across the TB workflow (see Figure 3.3 for a list of data types), the role for WGS data in diagnosis, treatment, and surveillance tasks, and participants confidence in interpreting different data types. The questionnaire primarily used multiple choice and true/false type questions, but also included the optional entry of freeform text. The questionnaire was deployed online using the FluidSurveys platform and participants were recruited using snowball and convenience sampling for a one-week period in July, 2016. For questions pertaining to diagnostic and treatment tasks, we gathered information only from participants self-identifying as clinicians; for the remaining sections of the survey, all participants were prompted to answer each question.

Only completed questionnaires were used for analysis. For questions pertaining to participants' background, their perception of WGS utility, and

their confidence interpreting WGS data, we report primarily descriptive statistics. To link TB workflow tasks to specific data types, we presented participants with different task-based scenarios related to diagnosis, treatment, and surveillance and asked which data types they would use to complete the task. For each pair of data and task we assigned a consensus score depending on the proportion of participants who reported using a data type for a specific task: 0 for fewer than 25% of participants, 1 for 25-50%, 2 for 50-75%, and 3 if more than 75% of participants reported using a specific data type for the task at hand. Consensus scores for a data type were also summed across the different tasks. Freeform text, when it was provided, was considered only to add context to participant responses.

3.2.3 Design Stage

The Discovery stage revealed which data types to include in the redesigned report, while the goal of the Design stage was to identify how it should be presented. We used a Design Sprint event to produce a series of prototype reports, which were then assessed through a second online questionnaire. This survey collected quantitative data on participants preference for specific design elements, with participants also able to provide qualitative feedback on each element a type of embedded mixed methods study design [22]

The Design Sprint was an interactive design session involving members of the University of British Columbias Information Visualization research group, in which teams created alternative designs to report WGS data for the diagnosis, treatment, and surveillance tasks. Teams developed paper prototypes [66] [119] of a complete WGS TB report and, at the completion of the event, presented their prototypes and the rationale for each design choice. The paper prototypes were then digitally mocked up, both as complete reports and as individual elements (see the results in Figure 3.4 and Figure 3.5); these digital prototypes were standardized with respect to text, fonts, and sample data where appropriate and used as the basis of the second online

survey.

In the Design Choice Questionnaire (text in Appendix A), we evaluated participants preferences for individual design elements, comparing the options generated during the Design Sprint as well as the initial COMPASS-TB report design, which we hereafter refer to as the control design. As with the first survey, the questionnaire used FluidSurveys, with participants recruited using snowball and convenience sampling. Individuals who had previously participated in the Data and Task Questionnaire were also invited to participate. The survey was open for one month beginning September 10, 2016 and was reopened to recruit additional participants for one month beginning January 5, 2017, as part of the registration for a TB WGS conference hosted by PHE. Only completed surveys were analyzed.

We used single-selection multiple-choice, Likert scale, and ranking questions to assess participant preferences. For multiple-choice and Likert scale questions, we calculated the number of participants that selected each option and report the sum. For questions that required participants to rank options we calculated a rescaled rank score as follows:

$$rescaledrank(D_i) = 1 - \frac{P^{-1} \sum_{p=1}^P R_p - 1}{N - 1}$$

where for each design choice (D_i), $i = \{1 \dots N\}$ where N is the total number of design choices, $R = \{1 \dots N\}$ is a raw rank (rank selected by a participant in the study), and $P = \{1 \dots P\}$ is the total number of participants. In our study, 1 was the highest rank (most preferred) and N was the lowest rank (least preferred) option. As an example, if a design, D_1 , is always ranked 1 (greatest preference by everyone), the sum of those ranks is P , resulting in a numerator of 0 and a rescaled rank score of 1; alternatively, if a design, D_2 , is always ranked last (N), the sum of those ranks will be $P * N$, and a rescaled rank score of 0. Thus, the rescaled rank score ranges from 1 (consistently ranked as first) to 0 (consistently ranked last). This transformation from

raw to rescaled ranks allows us to compare across questions with different numbers of options, but is predicated on each design alternative having a rank, which is why this approach was not extended to multiple choice questions.

To contextualize rescaled rank scores, we randomly permuted participants' scores 1000 times and pooled the rescaled rank scores across these iterations to obtain an average score (intuitively and empirically this is 0.5 for the rank questions and $\frac{1}{N}$ for multiple choice questions) and standard deviation. For each design choice, we plotted its actual rescaled rank score against the distribution of random permutations, highlighting whether the score was within ± 1 , 2, or 3 standard deviations from the random permutation mean score. The closer a score was to the mean, the more probable that the participants preferences were no better than random. We also calculated bootstrapped 95% confidence intervals for both rank and multiple choice type questions by re-sampling participants, with replacement, over 1000 iterations.

3.2.4 Implementation Stage

By combining the results of the Design Choice Questionnaire with medical test reporting requirements from the ISO15189:2012 standards, we developed a final template for reporting TB WGS data in the clinical laboratory. We used deviation from a random score, described above, as an indicator of preference, selecting design elements 3 or more standard deviations from a random score. When there was no strongly preferred element, we explain our design choice in the Design Walkthrough (Appendix A). We also considered consensus between clinicians and non-clinicians, and defaulted to clinician preferences in instances of disagreement as they are the primary consumers of this report. The final prototype is implemented in Latex and is available online as a template accessible at: <http://www.cs.ubc.ca/labs/imager/tr/2017/MicroReportDesign/>.

3.3 Results

Expert consults, the Task and Data Questionnaire, and the Design Choice Questionnaires recruited a total of 78 participants across different roles in TB management and control (Table 3.1).

Table 3.1: Total study participants across different stages of the Design Study Methodology.

	Expert Consults		Task and Data Questionnaire		Design Choice Questionnaire	
Stage	Discovery		Discovery		Design	
Data Collected	Qualitative		Quantitative		Qualitative & Quantitative	
Participants	N (% survey total)		N (% survey total)		N (% survey total)	
Clinician	2	29%	7	40%	13	25%
Nurse	1	14%	3	18%	5	9%
Laboratory	2	29%	3	18%	8	15%
Research	0	0%	1	6%	8	15%
Surveillance	1	14%	3	18%	8	15%
Other*	1	14%	0	0%	12	21%
Total	7	100%	17	100%	54	100%

3.3.1 Experts Emphasized Prioritizing Information and Revealed Constraints

The objective of our expert consults was to understand how reports from the reference mycobacteriology laboratory are currently used in the day-to-day workflows of various TB stakeholders, including clinicians, laboratorians, epidemiologists, and researchers, and what data types are currently used to inform those tasks. Tasks and data types enumerated in the interviews were used to populate downstream quantitative questionnaires; however, the interviews also provided insights into how stakeholders viewed the role of genomics in a clinical laboratory.

Amongst the procedural insights, stakeholders frequently reported that the biggest benefit of WGS over standard mycobacteriology laboratory protocols was to improve testing turnaround times and gather all test results into a sin-

gle document, rather than having multiple lab reports arriving over weeks to months. Several experts emphasized that these benefits can only be realized if the WGS analytical pipeline has been clinically validated. Although our study team included a clinician and a TB researcher, two surprising procedural insights emerged from the consultations. First, multiple experts from a clinical background emphasized that this audience has extremely limited time to digest the information found on a clinical report. In describing their interaction with a laboratory report, one participant noted that *10 seconds [to review content] is likely, one minute is luxurious* while others described variations on the theme of wanting bottom-line, actionable information as quickly as possible. This insight profoundly shaped downstream decisions around how much data to include on a redesigned report and how to arrange it over the report to permit both a quick glance and a deeper dive. Second, experts indicated that laboratory reports were delivered using a variety of formats, including PDFs appended to electronic health records, faxes, or physical mail. This created design constraints at the outset of the project our redesigned report needed to be legible no matter the medium, ruling out online interactivity, and needed to be black and white.

3.3.2 Experts Vary in Their Perception of Different Data Types

At the data level, we observed that the experts had differing perceptions of data types and desired level of detail between clinicians and non-clinicians, perhaps reflecting the clinicians procedural need for rapid interpretation. Clinicians emphasized the importance of presenting actionable results clearly and omitting those that were not clinically relevant for them. For example, when presented with the sequence quality data on the current COMPASS-TB report (Figure 3.1) metrics reflecting the quality of the sequencing run and downstream bioinformatics analysis interviewees did not expect the lab to release poor quality data, given the presence of strict quality control mechanisms. ISO15189:2012 standards require some degree of reporting around the measurement procedure and results, but this insight suggested

such data might best be placed later in the report in a simplified format, or described in the report comments. Similarly, experts were also divided on the interpretability and utility of the phylogenetic tree in the epidemiological relatedness section of the current COMPASS-TB report, with clinicians noting that the case belonging to an epidemiological cluster would not impact their use of the genomic test results.

Experts also disagreed about the level of detail needed for WGS data, and this appeared to depend upon whether the expert was a clinician as well as their prior experience with WGS through the COMPASS-TB project. For example, one expert indicated that “*clinicians are wanting to know which mutations conferred resistance*”, while another noted that they “*dont use these [mutations] right now routinely, so its not that relevant*”. When asked to comment on the resistance summary table in the current COMPASS-TB report (Figure 3.1), clinicians were concerned about the use of abbreviations for both drug names and susceptibility status leading to misinterpretation, and many were uncertain how to use the detailed mutation information in the resistotype table.

	WGS equivalent	DIAGNOSIS TASKS				TREATMENT TASKS			SURVEILLANCE TASKS					TOTAL SCORE
		Diagnose Latent TB	Diagnose Active TB	Reactive vs New Infection	Characterize Transmission Risk	Choose Meds	Choose Tx Duration	Assess Response to Tx	Guide Contact Tracing	Report to Public Health	Define a Cluster	Connect Case to Existing Cluster	Guide Public Health Response	
Patient Identifier	Same	3	3	3	3	3	3	3	2	1	1	1	1	26
Sample Collection Date	Same	3	3	2	3	3	3	3	1	1	1	1	1	24
Patient Prior TB Results	Same	3	2	3	3	3	3	3	1	1	1	0	1	23
Speciation	Speciation	1	3	2	3	3	3	3	2	1	1	1	1	23
Sample Type (sputum, fine needle aspirate etc.)	Same	2	3	2	3	3	3	3	1	1	1	0	1	22
Culture results	NA	1	3	2	3	3	3	3	2	1	1	0	1	22
Sample Collection Site (lymph node, lung etc.)	Same	2	3	2	3	3	3	3	1	1	0	0	1	21
Acid Fast Bacilli Smear	Speciation	2	3	2	3	2	3	3	1	1	1	0	1	21
Resistotype	Predicted DST	0	2	3	1	3	3	2	2	1	1	1	1	19
Phenotypic DST	Predicted DST	0	2	3	2	3	3	2	1	1	1	0	1	18
Chest x-ray	NA	3	3	2	3	0	2	3	1	0	0	0	0	17
Report Release Date	Same	2	2	1	2	2	2	2	1	0	1	0	1	15
Requester IDs	Same	2	2	2	2	2	2	2	1	0	0	0	0	15
Interpretation or comments from reviewer	Same	2	2	1	2	2	2	3	1	0	0	0	0	15
Predicted DST	Predicted DST	0	2	2	1	3	3	2	1	0	1	0	0	15
MIRU-VNTR	SNPs	0	2	3	1	1	1	1	1	1	1	1	1	13
Cluster Assignment	Same	0	2	2	1	1	1	0	1	1	1	1	1	11
SNP/variant distance	SNPs	0	1	2	1	1	1	0	1	1	1	1	1	10
Phylogenetic Tree	Same	0	2	1	1	1	1	0	1	0	1	1	1	9
Reviewer ID	Same	1	1	1	1	1	1	1	1	0	0	0	0	8
TST results	Speciation*	3	1	1	1	0	0	0	1	0	0	0	0	7
IGRA results	Speciation*	3	1	1	1	0	0	0	1	0	0	0	0	7
Lab QC	WGS Specific	0	1	2	1	1	1	0	1	0	0	0	0	7
Spoligotype	SNPs	0	1	1	1	0	0	0	0	0	0	0	0	3
RFLP	SNPs	0	1	1	1	0	0	0	0	0	0	0	0	3

Degree of Consensus: High (3) Some (2) Low (1) Very low (0)

Figure 3.3: Extent of consensus between TB workflow tasks and available TB data. Results are redundantly encoded using colour and a numerical value to represent the degree of consensus between participants around using a specific data type to carry out a specific task.

3.3.3 WGS Data is Vital, but Some Lack Confidence in its Interpretation

The expert consults provided a detailed overview of the tasks and data associated with TB care, allowing us to create a draft workflow outlining the TB diagnosis, treatment, and surveillance tasks coupled to the supporting data sources and data types (Figure A.1²). This workflow was used to design the Task and Data Questionnaire.

Of the 17 participants responding in full to the Task and Data Questionnaire (Table 3.1), most were from the United Kingdom (88%) and most reported professional experience and formal education in infectious diseases and epidemiology (Table A.1³). Participants were less likely to report education at the masters or doctoral level in microbial genomics, biochemistry, or bioinformatics (Table A.1). Fewer than half (47.1%) of participants had participated in TB WGS projects, but all (100%) participants were enthusiastic about the role of microbial genomics in infectious disease diagnosis, both today (47.1%) and in the near future, pending clinical validation (52.9%).

When queried about their potential future use of molecular data, whether WGS, genotyping, or other, participants indicated they foresaw themselves consulting, often or all the time, data on resistance-conferring mutations (82.3% of participants), MIRU-VNTR patterns (88.2%), epidemiological cluster membership (76.5%), single nucleotide polymorphism/variant distances from other isolates (64.7%), and WGS quality metrics (58.8%) (Table A.2). However, of the 14 different data types queried, the majority of participants only felt confident in interpreting four (MIRU-VNTR, drug susceptibility from culture, drug susceptibility from PCR or LPA, genomic clusters) - most participants only felt somewhat confident, or not confident at all, interpreting the other data types (Table A.3).

Moving from confidence in their own interpretation of laboratory data types

²This figure and all others with the prefix A are presented in Appendix A

³This table and all others with the prefix A are presented in Appendix A

to confidence in the utility of WGS data in general, the majority of participants were confident that information contained within the TB genome *can* be used to correctly perform organism speciation (76.5%), assign a patient to existing clusters (70.0%), rule out transmission events (64.7%), and to a lesser extent were confident TB WGS could be used to identify epidemiologically related patients (58.8%) and predict drug susceptibility (52.9%) (Table A.4). The majority of participants thought genomic data *may* be able to inform clinicians of appropriate treatment regimens (100%) and identify transmission events (94.1%); however, participants showed mixed consensus toward whether genomic data could be used to monitor treatment progress for TB (47.2%) or diagnose active TB (52.9%).

3.3.4 Respondent Consensus Suggests a Role for WGS in Diagnosis and Treatment Tasks

To examine which data types were being used to support diagnosis, treatment, and surveillance tasks in the workflow, we assigned a numerical score reflecting respondent consensus around each data type-task pair (Figure 3.3). We found greater consensus around the data types that participants would use in diagnosis and treatment tasks, but little consensus around the data they would use for surveillance tasks, contrasting with participants previously stated support for using WGS or other genotyping data for understanding TB epidemiology. Overall, the most frequently used data types included administrative data (patient ID, sample type, collection site, collection date) and results from current laboratory tests (solid or liquid culture, smear status, and speciation), which together were used primarily for diagnosis and treatment. Prior test results from a patient were deemed important; however, the earlier expert consults indicated that such data was difficult to obtain and unlikely to be included in future reports.

We also queried participants perceptions of barriers impacting their workflow, with the majority of participants (83.3%) reporting issues with both

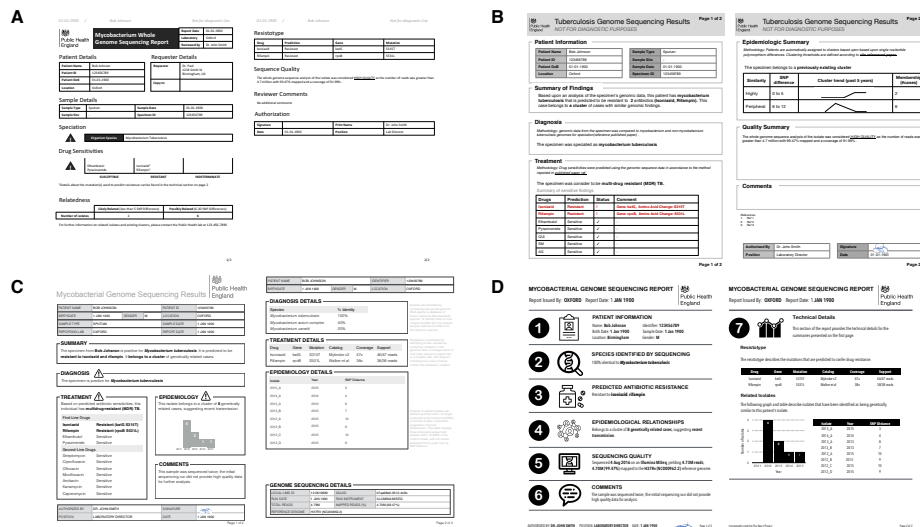


Figure 3.4: Digital mockups of complete report prototypes generated during the design sprint

the timeliness of receiving TB data from the reference laboratory and the distribution of test results across multiple documents (Table A.5) a finding that corroborated the procedural insights from the expert consults.

3.3.5 Prototyping Via a Design Sprint Produces a Range of Design Alternatives

Equipped with an understanding of how WGS data might be used in the various TB workflow tasks, we embarked on the Design stage of the design study methodology. A Design Sprint event involving study team members and information visualization experts resulted in four prototype report designs (Figure 3.4) and various isolated design elements (Figure 3.5). Although each prototype used different design elements for the required data types, when the prototypes were compared at the end of the event, common themes emerged. These included: presenting data in an order informed by the workflow data related to diagnosis, treatment, then surveillance; placing actionable, high-level on the front page, with additional details on the over

Original Report Element

1 Resistance Summary							2	4	5
INH	RIF	EMB	PZA	QUI	SM	AG			
U	S	S	S	S	S	S	3		

Tested Design Elements

- Alternative titles [Q8]**
 - A - Drug Resistance
 - B - Drug Sensitivity
 - C - Drug Susceptibility
 - D - Treatment
- Drug name format [Q9]**
 - A - 3 letter abbreviation (Ex. INH)
 - B - Full Name (Ex. Isoniazid)
 - C - Show me everything (Ex. Isoniazid (INH,H))
 - D - The are equally informative
- Susceptibility status format [Q10]**
 - A - 1 letter abbreviation (Ex. S,R,U)
 - B - Full Name (Ex. Susceptible, Resistant, Unknown)
 - C - They are equally informative
- Order or categorization [Q16]**

A – Drugs by category

Drug Susceptibility	
Drug	Prediction
Sensitive	Ethambutol, Pyrazinamide
Resistant	Isoniazid, Rifampin
Indeterminate	-

B – Listed by drug

Drug Susceptibility	
Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

C – Summary Sentence

Drug Susceptibility	
The specimen was resistant to isoniazid and rifampin, and sensitive to ethambutol and pyrazinamide	

D – Drugs by category bin

Drug Susceptibility		
Isoniazid Rifampin	Ethambutol Pyrazinamide	
Resistant	Sensitive	Indeterminate

E – Abbreviation listed by drug

Drug Susceptibility			
INH	RIF	EMB	PZE
R	R	S	S

5. Summary statement [Q13]

A - None

Drug Susceptibility	
Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Resistant
Pyrazinamide	Resistant

B - Summary sentence

Drug Susceptibility	
Based on predicted antibiotic mutations, the individual has multidrug resistant TB	
Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Resistant
Pyrazinamide	Resistant

C – Tick boxes

Drug Susceptibility	
Mono-resistant <input type="checkbox"/>	
Multidrug-resistant (MDR) <input type="checkbox"/>	
Extremely Drug Resistant (XDR) <input checked="" type="checkbox"/>	
Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Resistant
Pyrazinamide	Resistant

Figure 3.5: Isolated design component. The original report element, highlighted in red, is broken down into isolated design elements, each of which was tested independently in the report design survey. In this example, the original resistance summary yields five different alternative wordings and design elements.

page; and using both an overall summary statement at the beginning of the report as well as brief summary statements at the beginning of each section.

To drill down and determine which design elements best communicate the underlying data, we isolated individual design elements (Figure 3.5) and classified them as wording choices for example, which heading to use for

a given section of the report or design choices, such as layout, the use of emphasis, and the use of graphics (Table A.6).

3.3.6 The Design Choice Questionnaire Quantifies Participant Preferences for Specific Design Elements

We next developed an online survey, the Design Choice Questionnaire, to assess stakeholders preferences for both specific design elements and overall report prototypes. The distribution of public health roles amongst survey participants is presented in Table 3.1; all but 11 participants (20%) actively worked with TB data. Participants were employed by Academic Institutions (35.2%), Hospitals (24.1%), and Public Health Organizations (33.3%), with only 7.4% of participants being employed in some other sector. The majority of participants were from the UK (59.2%), while 11.1% were from Canada; the remaining 29.7% were drawn from the United States (6.5%), Europe (14.8%), Brazil (2.8%), India (2.8%), and Gambia (2.8%)

We first examined participants' preference for specific wording and design elements (Figure 3.6A,B), comparing elements arising from the prototypes to those used in the existing COMPASS-TB report, which acted as a control. Notably, of the 15 wording and design elements queried, in only two cases was the control design preferred over a design arising from one of the prototypes (note that one query did not compare to a control). Furthermore, in 8 out of 15 queries (Q6, Q8, Q9, Q10, Q12, Q17, Q5, Q18) participants showed strong preferences, wherein the to preference was +3 or more standard deviations from the mean for *both* clinicians and non-clinicians. Figure A.1 provides a version of Figure 3.6 with confidence intervals and indicates concordance between strong preferences and non-overlapping confidence intervals.

The findings from the analysis of wording elements (Figure 3.6A) showed that participants preferred complete terms to abbreviations, such as writing out isoniazid as opposed to INH or H, or resistant as opposed to R, and

that both clinicians and non-clinicians were in agreement over the preferred vocabulary for section headings. Interestingly, wording questions related to the treatment task yielded the widest range of rankings.

Clear preferences were also observed for information design elements, again largely concordant between clinicians and non-clinicians (Figure 3.6B). Participants preferred elements that drew attention to specific data, such as summary statements, shading, and tick boxes, and many participants preferred that sections be prioritized, with less important details relegated to the second page of the report. However, there was less consensus around how much detail to include and where. The majority of participants indicated that genomic data pertaining to resistance-conferring mutations should be included (Figure 3.6B; Q11), but were divided as to which data should be included and where. Most (85%) wanted to know the gene harboring the resistance mutation (i.e. *katG*; *inhA*), but only half wanted details of the specific mutation (50% wanted the amino acid substitution, 46% wanted to know the nucleotide-level change). We did not test any design elements displaying the strength of the association between the mutation and the resistance phenotype; however, we will add this to a future version of the report pending receipt of the final mutation catalog from the ReSeqTB Consortium.

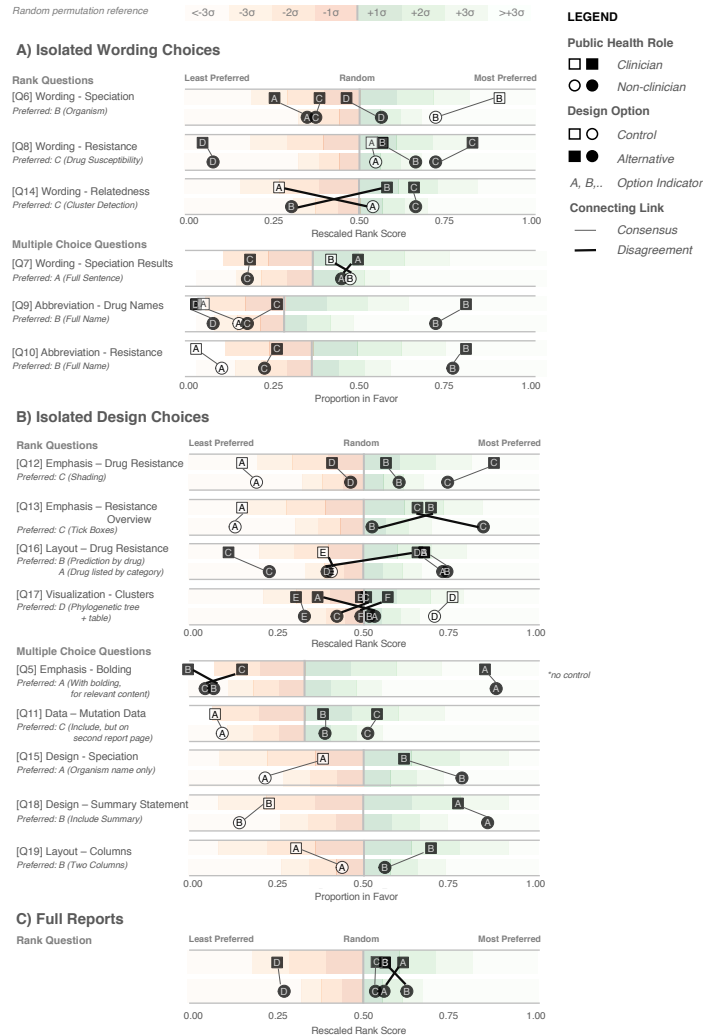


Figure 3.6: Design Choice Questionnaire results. Responses are grouped according to question type: wording (A), design choices (B), and full reports (C), and partitioned into clinician participants (squares) and non-clinician participants (circles). Responses are coloured according to whether they are the control design from the original report (white) or an alternative design devised in the design sprint (black). Lines connect options between clinician and non-clinicians preferences, with thicker crossing lines showing discordance between the two groups and vertical lines showing concordance in preferences. Rescaled rank scores are shown against a reference of random permutations (see Methods), with scores closer to 1 indicating the most preferred response. Specific questions are indicated with Q; the participants questions are shown in Table A.6

Interestingly, while both clinicians and non-clinicians reported similar rankings for most design elements, one element showed an unusual distribution of scores the visualization for showing genomic relatedness and membership in a cluster. While both groups of participants preferred a phylogenetic tree accompanied by a summary table, which is the current COMPASS-TB control design, the other four options appeared to be ranked randomly, with rescaled rank score close to 0.5, suggesting that none of the alternative options were particularly good.

We also had participants rank their preferences for the four prototype designs (Figure 3.6C). While all participants ranked Prototype D as their least preferred choice, many citing that the images used were too distracting, clinicians and non-clinicians varied in their ranking of the other three options, with clinicians preferring option A and non-clinicians preferring B. However, qualitative feedback collected for this question revealed that participants found comparing individual elements easier than comparing full reports.

3.3.7 Qualitative Data Affords Additional Insights into Report Design

The qualitative responses in the Design Choice Questionnaire raised important points that would otherwise not have been captured by quantitative data alone. For example, the importance of presenting drug susceptibility data clearly emerged from the qualitative responses. Participants indicated that *the report must call attention [to] drug resistance* and expressed concern that the abbreviation of drug names and/or predicted resistance phenotype could lead to misinterpretation and pose risks to patient safety, stating that *not all clinicians [are] likely to recognize the abbreviations* and *[using the full name] reduces the risk of errors, especially if new to TB*. When choosing how to emphasize predicted drug susceptibility information (shading, bolding, alert glyphs, or no emphasis), some participants suggested *shading draws the quickest attention to [resistance]* and that *with presbyopia, resistance*

can be easily missed and therefore shading affords greater patient safety, but other participants indicated drug susceptibility, rather than resistance, should be emphasized: *not sure that resistant should be shaded better to shade sensitive drugs in my view and it would be better to highlight what is working instead of highlight what is not working*. We opted to highlight resistance given the low incidence of drug-resistant TB in the UK and Canada, which were the primary application contexts. Some reported concerns as to whether such emphasis was possible with current electronic health records, including *[bolding or shading] may not transfer correctly and shaded [text] wont photocopy well*, which prompted us to test both printing and photocopying of the resulting report.

The issue of clinicians having little time to interact with the report, raised in both the expert consults and the Task and Data Questionnaire, also became apparent in the qualitative responses to the Design Choice Questionnaire, such as *the best likelihood of success will [come] from the ability to draw attention to someone scanning the document quickly*. However, participants perceptions of which design choices best promoted rapid synthesis varied. Some preferred summaries in the form of check boxes “*[a] tick box is the most straightforward way to summarize it. Reading a summary sentence will probably take longer*” and “*the check boxes provide an at-a-glance result*” while others preferred additional commentary “*interpretation is important; but tick boxes alone lack the necessary nuance required for interpretation*” and that “*tick boxes may cause confusion when clinicians read XDR without realizing that option is not selected. Ideal to add a comment about resistance*”. To address this concern we added a “No drug resistance predicted” option to the check-boxes (absent from the survey design options), and included shading elements to emphasize the drug susceptibility result.

The qualitative responses to Q17 (Figure 3.6B) provided further insight into the uncertainty around how best to represent genomic relatedness suggestive of an epidemiological relatedness. Some participants felt that data related to surveillance tasks should not appear in a report that is also meant for

clinicians, either because it wasn't relevant to this audience *[this data] should not appear in the report. It should only be given to field epi and researchers. Overloading the clinical report would be deteriorating and not useful for a clinician* or because they were uncertain about its interpretation *cluster detection would be fine for those who already know what a cluster is and my patient's isolate is 6 SNPs from someone diagnosed 3 years ago. What is the clinical action?*.

Of the design choices for cluster detection, several participants articulated that many of the options, including the control, *[included] too much information and [were] unnecessary for routine diagnosis/treatment*. However, others felt that the options did not provide sufficient detail and offered alternatives, such as *if you can combine the phylogenetic tree with some kind of graph showing temporal spread that would be perfect. Adding geographical data would be a really helpful bonus too..* This is an area of reporting that requires further investigation and was not fully resolved in our study.

Finally, participants were candid about those design options that did not work well—for example, of the report design with many graphics (Figure 3.6A, option D), participants indicated it was *distracting; looks like a set of roadworks rather than a microbiology report* and that it was important to *keep it simple*. Their feedback also revealed when our phrasing on the survey instruments was unclear.

3.3.8 Developing a Final Report Template

A Original Report

Mycobacterium Whole Genome Sequencing Report from MGIT Positive Samples

Not for diagnostic use

01/02/1915

Sample Details			
Sequencing Location	Oxford	Date received in Lab	
Local ID	123456789	Run date	01/01/19150115
Specimen ID			
Guidid	123456-79aab-910abr-15243hg		

Organism Identification

Predicted/closest match	
TBCOMP	100%
TBCOMP	100%
TBCOMP/TB	96.77%
TBCOMP/tuberculosis-canettii	35.71%
MACCOMP	21.21%

Sample/Sequencing Quality

Total reads	Mapped %	No reads mapped	Coverage %
(-millions)		(-millions)	
4.73	99.47	4.7	91.99

Resistance Summary

INH	RIF	EMB	PZA	QUI	SM	AG
U	S	S		S	S	S

Resistotype

Drug	Mutation	Nucleotides	Support (ACGT)	Source - (R/Total)	Prediction
INH	katG_A727T	GCC->ACC	(1600/1600)	Unclassified	UNK
			(0/167/0/0)		

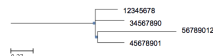
Relatedness

NB: This data may be added or updated at a later date

Nearest neighbour(s)

Sample #	Plate	Date received in Lab	Centre	No. of SNPs apart
123456789			Oxford	0
34567890		1900-01-01	Oxford	10
45678901		1015-01-31	Oxford	15
56789012			London	8

The alignment width is 285. Multiply this number by the tree metrics.



Comments

Authorised

Signature: _____ Print name: _____

Position: _____ Date: _____

B Revised Report

MYCOBACTERIUM TUBERCULOSIS GENOME SEQUENCING REPORT NOT FOR DIAGNOSTIC USE



Patient Name	JOHN DOE	Barcode	
Birth Date	2000-01-01	Patient ID	12345678910
Location	SOMEPLACE	Sample Type	SPUTUM
Sample Source	PULMONARY	Sample Date	2016-12-25
Sample ID	A12345678	Sequenced From	MGIT CULTURED ISOLATE
Reporting Lab	LAB NAME	Report Date/Time	2017-01-01 15:36
Requested By	REQUESTER NAME	Requester Contact	REQUESTER@EMAIL.COM

Summary

The specimen was positive for *Mycobacterium tuberculosis*. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.

Organism

The specimen was positive for *Mycobacterium tuberculosis*, lineage 2.2.1 (East-Asian Beijing).

Drug Susceptibility

Resistance is reported when a high-confidence resistance-conferring mutation is detected. "No mutation detected" does not exclude the possibility of resistance.

Drug class	Interpretation	Drug	Resistance Gene (from Acetaminophen)
First Line	Susceptible	Ethambutol	No mutation detected
		Pyrazinamide	No mutation detected
	Resistant	Isoniazid	katG (S315T)
		Rifampin	rrsB (S531L)
Second Line	Susceptible	Streptomycin	No mutation detected
		Clarithromycin	No mutation detected
		Ofloxacin	No mutation detected
	Resistant	Moxifloxacin	No mutation detected
		Amikacin	No mutation detected
		Kanamycin	No mutation detected

Page 1 of 2

Patient ID: 12345678910 | Date: 2017-01-01 | Location: Someplace

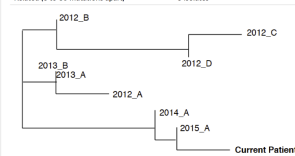
MYCOBACTERIUM TUBERCULOSIS GENOME SEQUENCING REPORT NOT FOR DIAGNOSTIC USE



Cluster Detection

The current isolate was clustered with previously sequenced isolates, suggesting recent transmission.

Relatedness	Number of prior matching isolates
Closely Related (< 5 mutations apart)	2 isolates
Related (6 to 30 mutations apart)	4 isolates



Assay Details

Sample ID	A12345678	Barcode	
Sequencer	ILLUMINA HISEQ 2500	Method	WGS
Pipeline	RESEQTB V.3.2C	Reference	H37RV

Comments

No additional comments for this report.
Standard Disclaimer: Low frequency hetero-resistance below the level of detection by sequencing may affect typing results. The interpretation provided is based on the current understanding of genotype-phenotype relationships.

Authorised

Signature: _____ Name: _____

Position: _____ Date: _____

Page 2 of 2

Patient ID: 12345678910 | Date: 2017-01-01 | Location: Someplace

Figure 3.7: Original and revised reports. The revised report uses empirical evidence gathered through multiple stages of a human centered design process. Note that the image in the upper corner of the revised report is a placeholder for an organizational logo.

There are no prescriptive guidelines around integrating our quantitative data, qualitative data, and ISO15189:2012 reporting requirements; thus, we have attempted to be as transparent and empiric as possible in justifying our final design (Figure 3.7). A more thorough walkthrough is presented in Appendix A, and here we highlight selected choices. The final prototype is implemented in Latex and is available online as a template accessible at:

<http://www.cs.ubc.ca/labs/imager/tr/2017/MicroReportDesign/>

We first incorporated ISO15189:2012 requirements (see Appendix A) into the final report template and then turned to the preferences expressed in the Design Choice Questionnaire. Overall, information was structured to mirror the TB workflow diagnosis, treatment, then surveillance. We chose to limit bolding to relevant information, and used shading to highlight important and actionable clinical information, under the rationale that appropriate use of emphasis could facilitate an accurate and quick reading of the report, with detailed information present but de-emphasized.

In two instances, our design decisions deviated from participant preferences: we opted to use one column instead of two, and we presented detailed genomic resistance data on the first page of the report, rather than the second page. A single column was chosen as all of the information ranked as important by participants could be presented on a single page without the need to condense information into two columns. Because many of the resistotype details of the original report, such as mutation source and individual nucleotide changes (Figure 3.1), were not included in the revised report, it was possible to present all of the participants' desired data in a single table on one page.

A draft of the final design was presented to a new cohort of TB stakeholders at a September, 2017 expert working group on standardized reporting of TB genomic resistance data. Through a group discussion, subtle changes to the report were made, including updating some of the language used (for example, replacing occurrences of the word “sensitive” with “susceptible”),

adding the lineage to the Organism Section, and adding additional fields to tables describing the sample, and the assay, such as what type of material was sequenced (pure culture, direct specimen) and what sequencing platform was used.

3.4 Discussion

Microbial genomics is playing an increasingly important role in public health microbiology, and its successful implementation in the clinic will rely not just on validation and accreditation of WGS-based tests, but also in how effective the resulting reports are to stakeholders, including clinicians. Using Design Study Methodology, we developed a two-page report template to communicate WGS-derived test results related to TB diagnosis, drug susceptibility testing, and clustering.

To our knowledge, this project is the first formal inquiry into human-centered design for microbial genomics reporting. We argue that the application of human-centered design methodologies allowed us to improve not only the visual aesthetics of the final report, but also its functionality, by carefully coupling stakeholder tasks, data, and constraints to techniques from information and graphic design. Giving the original report a “graphic design facelift” would not have improved the functionality, as some of the information in the original report was found to be unnecessary, presented in a way that could lead to misinterpretation, or did not take into account stakeholder constraints. For example, interviews and surveys revealed procedural and data constraints our study team had not anticipated, including the limited time available for clinicians to read laboratory reports and the need for simple, black and white formatting amenable to media ranging from electronic delivery to fax. These findings were critical to shaping the downstream design process. Furthermore, in nearly every case, study participants preferred our alternative design elements, informed by empirical findings in the discovery stage, over the control elements derived from the original report. Our ap-

proach also suggested that some participants are not confident in their ability to interpret certain types of genomic data. As WGS moves towards routine clinical use, it is clear that successful implementation of genomic assays will also require complementary education and training opportunities for those individuals regularly interacting with WGS-derived data.

Although human-centered information visualization design methodologies are commonly used in software development, it could be asked whether they are warranted in a report design project. One advantage of tackling the simpler problem of report design is that it allows us to demonstrate Design Study Methodology in action and link evidence to design decisions more clearly than with a software product. We also collected data with the intention of applying it to the development and evaluation of more complex reporting and data visualization software that we plan to create. Similarly, others can use our approach or our data to inform the design of simple or complex applications elsewhere in pathogen genomics and bioinformatics.

The exploratory nature of this project brings with it certain limitations. First, our participants were identified through convenience and snowball sampling within the authors networks, and thus are likely to be more experienced with the clinical application of microbial genomics. While this is appropriate for the context of our collaboration, in which our goal is redesigning a report for use by the COMPASS-TB team and collaborating laboratories, it does limit our ability to generalize the findings to other settings. WGS is only used routinely in a small number of laboratories, and even if its reach were larger, these may be settings where English is not the first language used in reporting clinical results, or where written text is read in different ways both of which would affect our design choices. Second, we did not have *a priori* knowledge of the effect sizes (i.e. extent of preferential difference for each type of question) in the Design Choice Questionnaire, making sample size calculations challenging. Had *a priori* effect sizes been available, the study could be powered, for example, for the smallest or average effect size. To avoid mis-characterizing our results, we have relied

on primarily descriptive statistics, without tests for statistical significance, and assert that our findings are best interpreted as first steps toward a better understanding how information *and* visualization design can play a role in reporting pathogen WGS data. However, when confidence intervals were calculated for the results of the Design Choice Questionnaire, we observed that non-overlapping confidence intervals separated user preferences as well as the *deviation from a random score* metric that we primarily used in our analysis. We argue the latter is a useful measure for exploratory studies without clear *a priori* knowledge of effect sizes for proper sample size calculations. Finally, we did not undertake a head-to-head experimental comparison between the original report design and the revised design. While this comparison had been planned at the outset of our project, the results of the Design Choice Questionnaire showed such a clear preference for the alternative designs when comparing isolated components that we concluded there was no need for such a final test as it would yield little new evidence.

For researchers wishing to undertake a similar human-centered design approach, we have summarized our primary findings into three experimental guidelines and five design guidelines. These guidelines arose from our experience throughout this report redesign process, but are intended to apply generally to the process of designing visualizations for microbial genomic data or other human health-related information.

The three experimental guidelines reflect the areas of the design methodology that we found to be particularly important in our data collection and analysis as well as the final report design process. First, **design around tasks**. It is tempting to simply ask stakeholders what they want to see in a final design, but many of them will not be able to create an effective end product because design is not their principal area of expertise. However, stakeholders know very well what they do on a daily basis and can indicate data that are relevant to those specific tasks and can indicate in which areas they require more support. The role of the designer is to marry those tasks, clinical workflows, and constraints into design alternatives. Depending on the tasks and context,

many design alternatives might be possible, making use of colour, more complex visualizations, or interactivity. In other situations, such as the one presented here, design constraints limit the range of prototypes that can be generated. Second, **compare isolated components, and not just whole systems**. Here we use system to mean either a simple report or a more complex software system. Comparing whole systems can overload an individual's working memory, meaning they may rely on heuristics such as preferences around style or distracting elements, when assessing and comparing full systems [104]. Presenting isolated design elements and controlling for non-tested factors (i.e. font, text) can reduce the burden on working memory and isolate the effect of design alternatives. Finally, **compare against a control whenever possible**. If a prior report or system exists, or if there are commonly agreed upon conventions in the literature or field, it is useful to compare novel designs against an existing one. More generally, comparison of multiple alternatives is the most critical defense against defaulting to *ad hoc* designs and the most important step of our human-centered design methodology.

Our five design guidelines reflect techniques from information visualization and graphic design that we used in an attempt to improve the readability of the report and balance different stakeholder information needs. First, **structure information such that it mimics a stakeholder's workflow**. In this case, the report prioritizes a *clinical* workflow, and this workflow is reflected in the report's design through the use of gestalt principles [75] – treating the whole as greater than the sum of its parts. Specifically, we group related data and order information hierarchically, so that the document is read according to the clinical narrative we established in the Discovery phase. Second, **use emphasis carefully**. Here, bolding, text size, and shading were reserved to highlight important data and were not applied to aesthetic aspects of the report design. Third, **present dense information in a careful and structured manner**. Stakeholders should not have to search for relevant information a cognitively expensive task [19] that can result in

information loss [107]. Through the combination of gestalt, visual hierarchy, and careful use of emphasis, it is possible to present a lot of information by creating two layers: a higher-level “quick glance” layer and a more detailed lower layer. The quick glance layer should contain the relevant and clinically actionable information and should be visually salient (i.e “pop-out”), while the detailed layer should be less visually salient and contain additional information that some, but not all, stakeholders may wish to have (based on their tasks and data needs). Fourth, **use words precisely**. Specific terminology may not be uniformly understood or consistently interpreted by stakeholders, particularly when the designer and the stakeholders come from different domains, or even when individuals in the same domain have markedly different daily workflows, such as bioinformaticians and clinicians. Finally, **if using images, do so judiciously**. Images can be distracting when they do not convey actionable information relevant to the stakeholder.

3.5 Conclusions

We applied human-centered design methodologies to redesign a clinical report for a reference microbiology laboratory, but the techniques we used drawn from more complex applications in information visualization and human-computer interaction can be used in other scenarios, including the development of more complex data dashboards, data visualization or other bioinformatics tools. By introducing these techniques to the microbial genomics, bioinformatics, and genomic epidemiology communities, we hope to inspire their further use of evidence-based, human-centric design.

Chapter 4

Adjutant:

**an R-based Tool to Support Topic Discovery for
Systematic and Literature Reviews**

Research complete — Terran Adjutant

¹ Adjutant is an open-source, interactive, and R-based application to support mining PubMed for literature reviews. Given a PubMed-compatible search query, Adjutant downloads the relevant articles and allows the user to perform an unsupervised clustering analysis to identify data-driven topic clusters. Following clustering, users can also sample documents using different strategies to obtain a more manageable dataset for further analysis. Adjutant makes explicit trade-offs between speed and accuracy, which are modifiable by the user, such that a complete analysis of several thousand documents can take a few minutes. All analytic datasets generated by Adjutant are saved, allowing users to easily conduct other downstream analyses that Adjutant does not explicitly support. We used Adjutant in the methodology presented in the subsequent Chapter 5 to cluster the genomic epidemiology research literature with the intention of sampling and classifying data visualization strategies within different topic clusters.

¹This chapter has been previously published as an *Application Note* [26]:
A. Crisan, T. Munzner, and J. L. Gardy. Adjutant: an R-based Tool to Support Topic Discovery for Systematic and Literature Reviews. *Bioinformatics*, 35(6):10701072, 08 2018.
[doi:10.1093/bioinformatics/bty722](https://doi.org/10.1093/bioinformatics/bty722)

4.1 Introduction

Literature reviews, whether systematic or not, can necessitate the analysis of thousands of documents to derive relevant topics, which can quickly become overwhelming [83]. Software that implements text-mining techniques, such as Abstrackr [94], EpiphaNet [21], or Retro [134], as well as libraries within various programming languages, can help to streamline the literature review process by identifying topics relevant to the user. Yet in spite of the availability of such tools, automation is still not routinely used to support literature reviews [106], perhaps due to existing tools’ intensive compute requirements and the need for extensive user input. To address these limitations, we have developed Adjutant, an R package with an associated Shiny application that supports the quick derivation and exploration of topic clusters within a PubMed document corpus. Adjutant’s objective is to provide a rapid overview of the corpus’ topic structure with minimal overhead, facilitating an individual’s literature review. Like the military rank from which it draws its name, Adjutant is intended to support an individual’s expert knowledge, rather than to supplant it.

4.2 Implementation Details

Adjutant is primarily designed to be used as a graphical user interface (GUI) that guides a user through a series of steps to query, cluster, explore, and sub-sample documents from a PubMed query. The GUI is deployed through R’s Shiny framework, such that Adjutant requires no specialized hardware and all of the analysis takes place on the user’s own computer. Adjutant’s workflow is also visible to a user as part of the R package and can thus be integrated into an R Script, bypassing the GUI altogether. In this section we will briefly describe the various steps within Adjutant’s workflow. We refer the reader to Appendix B for specific details of the implementation.

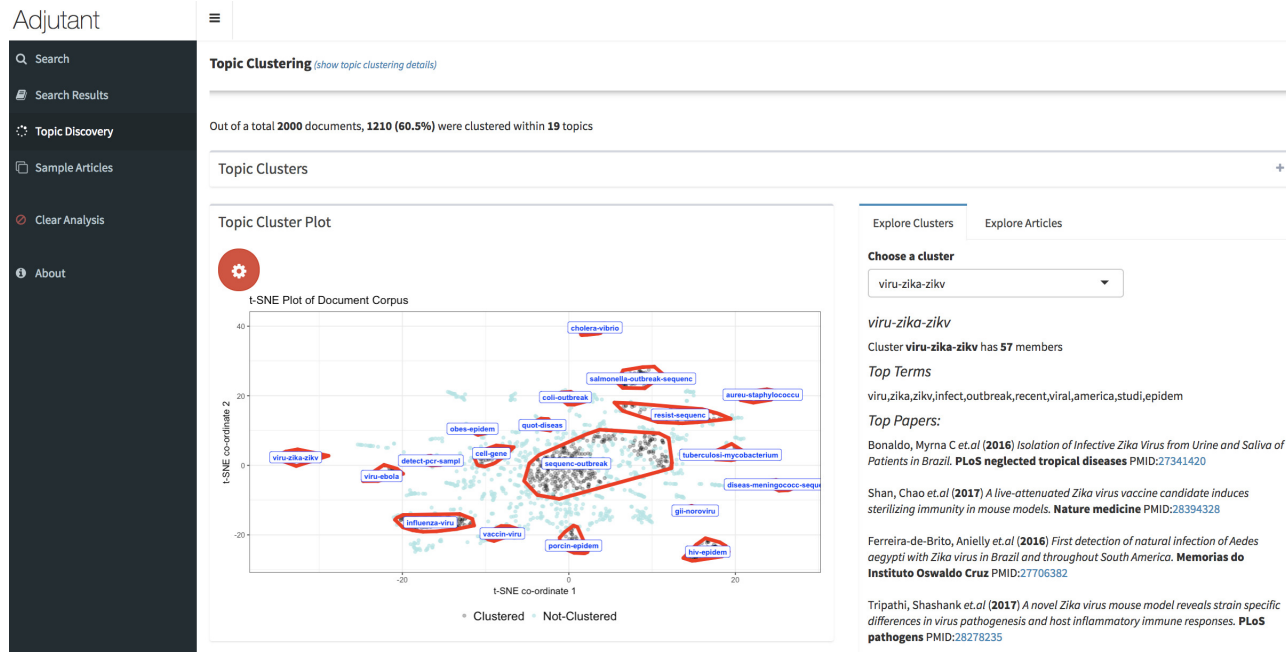


Figure 4.1: The Adjutant user interface after running the unsupervised topic clustering on 2000 documents from the the query ' (outbreak OR epidemic OR pandemic) AND genom*' run on June 6, 2018. The GUI consists of four tabs that guide the user through the analysis steps.

Querying PubMed and Preparing a Document Corpus. Adjutant runs PubMed-compatible queries through the Entrez API and downloads article data and metadata from the NCBI servers to a user’s own computer. The time it takes to download the data can vary depending upon the bandwidth of the Internet connection and number of documents. The resulting document corpus is then run through the textmining workflow specified in [109], with some modifications (see Appendix B). The workflow begins by using article titles and abstracts to derive a document term matrix (DTM), in which articles are rows, stemmed single words (terms) are columns, and the “term frequency inverse document frequency” (tf-idf) is the relevant analytic metric. The tf-idf metric is a statistic that weighs how important a term is for a particular article relative to other articles in a corpus and is a commonly used metric in text analysis.

Unsupervised Topic Clustering. Unsupervised topic clustering is carried out by first dimensionally reducing the data using t-SNE [115] and then clustering with hdbscan (a hierarchical density based, spatial clustering; [15]). The t-SNE algorithm is routinely applied to text data [16] and we choose to use hdbscan for clustering because it has a much more intuitive parameter of minimum cluster size rather than the more common, and less intuitive, number of topics in the corpus; Adjutant performs a greedy search to select a good setting for hdbscan’s minimum cluster size parameter. The hdbscan scan algorithm allows us to generate clusters of different sizes and exclude articles that do not easily belong to any one cluster. Applying t-SNE ahead of hdbscan speeds up the analysis, in line with Adjutant’s goals, but at the cost of some accuracy [122]. The goal of Adjutant’s unsupervised clustering procedure is to allow a user to get the gist of the document collection structure, emphasizing ease of use over pure accuracy. In Appendix B we provide further details on Adjutant’s implementation and an evaluation of its unsupervised clustering techniques.

Sampling and Automatically Saving Analysis Results. A user may wish to export all the data or some subset of it, either before or after clustering, in

order to read articles for further analysis. In the 'Sampling' tab of the GUI, users may choose between downloading all or some subset of the data. If only downloading a subset of the data, users may do so by ranking articles by citation count or year, or randomly sampling articles with the option to weight by year or citation count. If clustering has been performed, subsets can be obtained from across the topic clusters. Adjutant also automatically saves analysis documents into R-compatible formats that can be that can be reloaded and reexamined within Adjutant at a later date or be used in other downstream analyses that Adjutant does not itself support.

4.3 Usage Scenario

Appendix B contains several detailed examples of Adjutant usage scenarios in both notebook and video form. Users unfamiliar with the R environment or who wish to interactively explore their data can use the Adjutant GUI (Figure 4.1). The GUI guides users through the analysis steps from query to clustering, provides users with an overview of their search results in the 'Search Results' tab, allows them to generate and explore topic clusters in the 'Topic Discovery' tab, and to export all their data or some subset in the 'Sample Articles' tab. Advanced R users can bypass the GUI altogether and use Adjutant's underlying methods in their own R Script. We have also implemented Adjutant's source code in a modular format and have included extensive documentation, such that advanced users are also free to modify various aspects of Adjutant's workflow while still leveraging its GUI.

4.4 Conclusion

Adjutant is an R-based application that supports literature reviews by enabling users to quickly visualize and explore the topic structure of a set of PubMed-derived documents. Its R-based architecture enables users to access a wide range of analysis tools. Like the military rank from which it draws its name, Adjutant is intended to support an individual's expert knowledge, rather than to supplant it.

Chapter 5

GEViT:

A Systematic Method for Surveying Data Visualizations and a Resulting Genomic Epidemiology Visualization Typology

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it — John W. Tukey

¹ Data visualization is an important tool for exploring and communicating findings from genomic and healthcare datasets. Yet, without a systematic way of organizing and describing the design space of data visualizations, researchers may not be aware of the breadth of possible visualization design choices or how to distinguish between good and bad options. We have developed a method that systematically surveys data visualizations using the analysis of both text and images. Our method supports the construction of a visualization design space that is explorable along two axes: why the visualization was created and how it was constructed. We applied our method to a corpus of scientific research articles from infectious disease genomic epidemiology and derived a Genomic Epidemiology Visualization Typology (GEViT) that describes how visualizations were created from a series of chart types, combinations, and enhancements. We have also implemented

¹This chapter has been previously published [24]. As indicated in preface, the names of the chart combinations have been modified from the original publication.

A. Crisan, J. L. Gardy, and T. Munzner. A Systematic Method for Surveying Data Visualizations and a Resulting Genomic Epidemiology Visualization Typology: GEViT. *Bioinformatics*, 35(10):16681676, 09 2018. doi:10.1093/bioinformatics/bty832

an online gallery that allows others to explore our resulting design space of visualizations. Our results have important implications for visualization design and for researchers intending to develop or use data visualization tools. Finally, the method that we introduce is extensible to constructing visualizations design spaces across other research areas.

5.1 Introduction

Genome sequencing is becoming an integral part of modern infectious disease diagnostics [85] and epidemiology [35, 92]. When genomic and/or phylogenetic data are combined with clinical and epidemiologic data routinely generated by public health laboratories and programs, the resulting analyses support a variety of public health professionals, including clinicians, epidemiologists, researchers, and policymakers, in their real-time decision-making around treatment, surveillance, and outbreak response. However, this new data-driven approach to public health also introduces interpretability challenges—it is difficult to succinctly and accurately represent such multi-variate and high-dimensional data, particularly when many stakeholders do not routinely work with the genomic or phylogenetic data these analyses rely upon. These challenges arise not only late in an investigation, when attempting to communicate the results of an analysis, but also in the early phases of a project, such as initial data exploration and model-building [47].

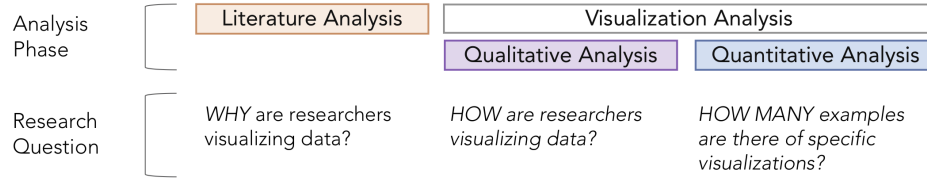
Data visualization is an important means to address interpretability challenges, and one which is increasingly being used in genomic epidemiology. Tools including nextstrain [48] and Microreact [4] use developments in web technologies to produce sophisticated, interactive data visualizations that allow users to explore and interact with public health phylogenetic data in an epidemiological context. Other tools, such as treeviewer [55], GenGIS [88], or libraries such as PhyloCanvas (<http://phylocanvas.org/>) also allow researchers varying degrees of freedom to generate visualizations blending phylogenetic trees with other metadata. As more and more visual-

ization tools and libraries are being developed for genomic epidemiology, it is an appropriate moment at which to assess the type of visualizations being generated and used in public health genomic studies in order to inform the design of future visualization tools.

When analyzing existing data visualizations, the concept of a visualization design space becomes important. This design space is defined as the combinatorial space of data visualizations afforded by graphical marks (points, lines, and areas) that convey information through their aesthetic properties (position, colour, size, shape, texture), which are also referred to as channels in the information visualization research literature [79]. There have been explicit attempts to describe visualization design spaces and share them via web galleries, such as SetVis [1], TreeVis [101], Visualizing Health (<https://www.vizhealth.org>), and BioVis Explorer [58], but these were not created through a process as systematic as what we propose and thus do not serve to provide insight into current practice in a specific research community. Collections of visualizations also arise implicitly from search engine results, including Google, PubMed, or Semantic Scholar image searches, but these lack a systematic taxonomy and ontology describing the visualizations themselves. It is only through organizing the visualizations created by a research community within a design space that common visualization practices become apparent and better practices can be suggested.

Here, we present a method for the systematic analysis of a visualization design space. By employing this structured approach to both generating and analyzing a suite of visualizations within the context of public health genomic epidemiology, we reveal current data visualization practices common to this domain. We are able to identify those visualization designs that could be better supported through new software tools or improved to make them more effective, as well as areas of the design space that are currently underused. This methodological contribution can be applied to visualization design spaces in domains beyond public health genomic epidemiology; here we describe its application in a specific domain as an additional contribution.

A General Method Overview



B Application of our Method to Infectious Disease Genomic Epidemiology

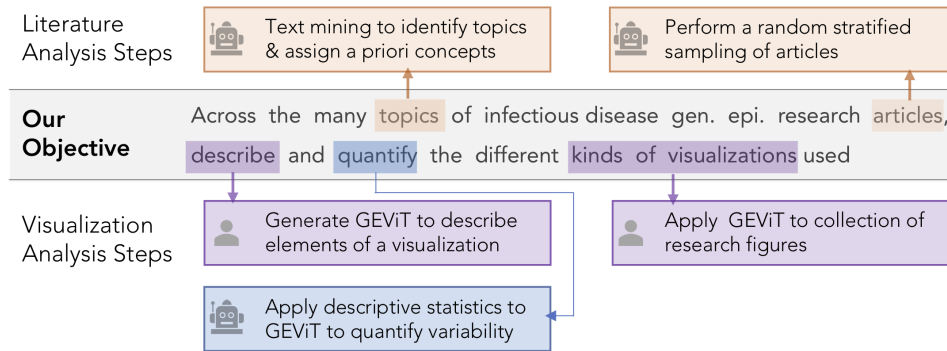


Figure 5.1: Method and application overview. **A)** Constructing and systematically analyzing a visualization design space requires analysis of both the literature and visualizations themselves, using qualitative and quantitative approaches. **B)** Automated steps, as indicated by the robot icon, are used in literature analysis to identify articles in genomic epidemiology and the topics those articles address. Manual steps, as indicated by the human icon, are used in the analysis of visualizations derived from those articles, followed by further quantification with automated statistical approaches. See Figure C.1 and C.2 for more details

We present the **Genomic Epidemiology Visualization Typology (GEViT)**, and we provide a web-based platform for exploring GEViT that researchers, bioinformaticians, and software developers can use to inform their own genomic epidemiology data visualization practice.

5.2 Methods

5.2.1 Developing a Method for the Systematic Analysis of Data Visualizations

Data visualizations are often challenging to analyze because, unlike images of real-world objects, visualizations in the scientific literature are abstractions devised by researchers to convey a combination of concepts. For example, phylogenetic trees display genomic data in an evolutionary context, and can be further enriched to show metadata about the sampled sequences and/or organisms and the underlying evolutionary processes. Visualizations vary across research contexts, and can be described using the nested model for visualization design and analysis [78], which deconstructs a data visualization into four layers: the *why* – a research or domain problem that a data visualization supports; the *what* – the data that needs to be visualized and the specific tasks performed using the data and visualization, such as finding trends or communicating a specific finding; the *how* – the visual design and interactivity; and the algorithmic implementation of the visualization.

We have constructed a method for the systematic analysis of data visualizations that specifically articulates and then attempts to connect the visualization research problem (*why*) with the visualization design (*how*) – this goal is possible because we can meaningfully capture and label these elements of a data visualization through a systematic analysis based on image and textual analysis. Our method consists of an initial literature analysis phase followed by a visualization analysis phase, resulting in a visualization design space in which images are classified according to their *why* and their *how*. The literature analysis phase (Figure 5.1) automatically analyzes text from a corpus of research articles to identify the topic of a data visualization – *why* it was created – as we assume that different topics are likely to yield different visualization designs. In the current instantiation of this method, we also use the literature analysis phase to perform a random stratified sampling of

articles to select a reasonable subset of visualizations for the subsequent visualization analysis phase, which requires a human-curated inventory of each image. In this phase, we iteratively apply open and axial qualitative coding techniques to the set of images harvested from the sampled articles. The iterative qualitative coding phases [20] ultimately yield a set of hierarchical taxonomies that we collectively refer to as a visualization typology and that allows us to articulate how visualizations are created (Figure 5.1). Further detail around the methods employed during both phases are provided below as well as in Figure C.1 and C.2².

Our specific application of this method to articles and images from the infectious disease genomic epidemiology context resulted in the Genomic Epidemiology Visualization Typology (GEViT) a structured way of describing a collection of visualizations that together form a visualization design space. As a research community publishes new data visualizations, these can be annotated using the typology and added to the design space, and may even result in the addition of new terms to the typology if the image includes new elements of visual design.

5.2.2 A Systematic Analysis of Data Visualizations from the Infectious Disease Genomic Epidemiology Research Literature

Literature Analysis

We developed an R package called Adjutant [26], described in detail elsewhere, to support our literature analysis. Here, we briefly describe how Adjutant’s functionalities are used to search, prepare, and cluster articles in order to derive a representative subset of documents for the visualization analysis phase.

Search Terms. We searched for articles related to infectious disease ge-

²These figures and all others with the prefix C are presented in Appendix C

nomie 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*, combining their results and retaining only unique records for further analysis. We also manually included cancer genomics articles that were known to us to use phylogenetic trees in their analysis.

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

Unsupervised Topic Clustering. We used the t-SNE and hdbscan algorithms to perform an unsupervised clustering using the DTM. We used the Barnes-Hut implementation of t-SNE [115], which allows for some acceleration at the cost of accuracy, with the perplexity parameter set to 100; otherwise default parameters of the R package implementation were used [60]. We then used hdbscan [15] on the t-SNE co-ordinate to derive the topic clusters; we show in our earlier work on Adjutant [26] that this order of operations yields relevant results. Clusters are sensitive to the minimum number of cluster points (minPts) parameter supplied to the hdbscan, thus 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 combined results

from the minPts 75 and 150 analyses. Each cluster is assigned a topic by using the two most frequent terms within the cluster. Following topic clustering, we validated our clusters using an external list of human pathogens (Table C.1), assessing the correspondence between pathogen names and cluster topics.

Linking To A *Priori* Concepts. Before conducting the unsupervised clustering, we discussed what results we might expect given our knowledge of research activities in the public health genomic epidemiology community. This initial discussion produced a set of 23 *a priori* concepts that we categorized into three groups: *genomic concepts*, including drug resistance, genome, genotype, molecular biology, pathogen characterization, phylogeny, and population diversity; *epidemiology concepts*, including clusters, disease reservoirs, geography, outbreaks (at international, community, and hospital levels), surveillance, transmission, vaccine, and vectors, and *medical concepts* (clinical, cancer, diagnosis, outcome, and treatment). Following the clustering, we identified bigrams that occurred in at least ten articles within a pathogen topic cluster and between at least 10% of the other pathogen topic clusters, and manually assigned those bigrams to an *a priori* concept (Table C.2) for example, the bigram "vancomycin resistance" was assigned to the *a priori* concept of "drug resistance". Assignments were validated by internal discussion among the research team, including a genomic epidemiology expert.

Document Sampling. To produce a manageable, diverse, and systematically derived dataset for the human-curated visualization analysis step, we performed random stratified sampling on our document corpus, sampling one document for each *a priori* concept within each of the automatically derived topic clusters. Each sampled article was examined and either considered acceptable for further analysis or rejected. Most articles were rejected because they did not contain any figures; other reasons for rejection included: full text article not accessible; article not in English; article was about a

laboratory or bioinformatics technique and not an epidemiological scenario; no human data; or the article was a review rather than original research. For each rejected article, we resampled two additional articles, choosing one 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 C.3 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, or illustrations without underlying data. We also included a small number of missed opportunity tables—stand-alone tables that we felt could have been visualized, most frequently matrices of numbers or large tables of patient metadata where each row consisted of a patient.

5.2.3 Visualization Analysis

Extracted figures and tables were analyzed using iterative open and axial qualitative coding techniques. Originally derived from the use of Grounded Theory in sociology, psychology, and anthropology [20], qualitative coding methods are now being used in human-computer interaction [57] and information visualization research [17]. Qualitative coding involves iteratively examining data and assigning it to some category. The categories themselves are refined and can take on hierarchical relationships through different cycles of the coding process (see appendix C), and were informed here by concepts from visualization theory and terminology [79].

Here, we analyzed whole figures separately—we did not decompose multi-part figures in order to understand the potential interplay between panels within

a figure. We began by creating a taxonomic code describing the types of charts present in different figures. We next examined how different types of images were combined to show different aspects of the data and thus created a chart combination taxonomy. Finally, we created a taxonomy that captured how basic chart types were enhanced to encode additional information. We refer to the collection of taxonomic code sets for chart types, combinations, and enhancements that were derived from this document corpus of genomic epidemiology research articles as GEViT. We conducted three rounds of qualitative coding, in which we reviewed figures and made additions or changes to GEViT; by the third round of coding, there were too few additional modifications to warrant a subsequent round.

Creating an Explorable Visualization Design Space

We used the results of the literature and the visualization analysis phases to produce an explorable visualization design space, which is freely available at <http://gevit.net>. The images presented gallery are used under Fair Use copyright terms and we provide links back to the original source publications.

5.3 Results

5.3.1 Literature Analysis

Literature Mining Showed Article Clusters According to Pathogens

We assembled a document corpus of 17,974 articles pertaining to infectious disease genomic epidemiology research published in the past 10 years (Figure 5.2). 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,

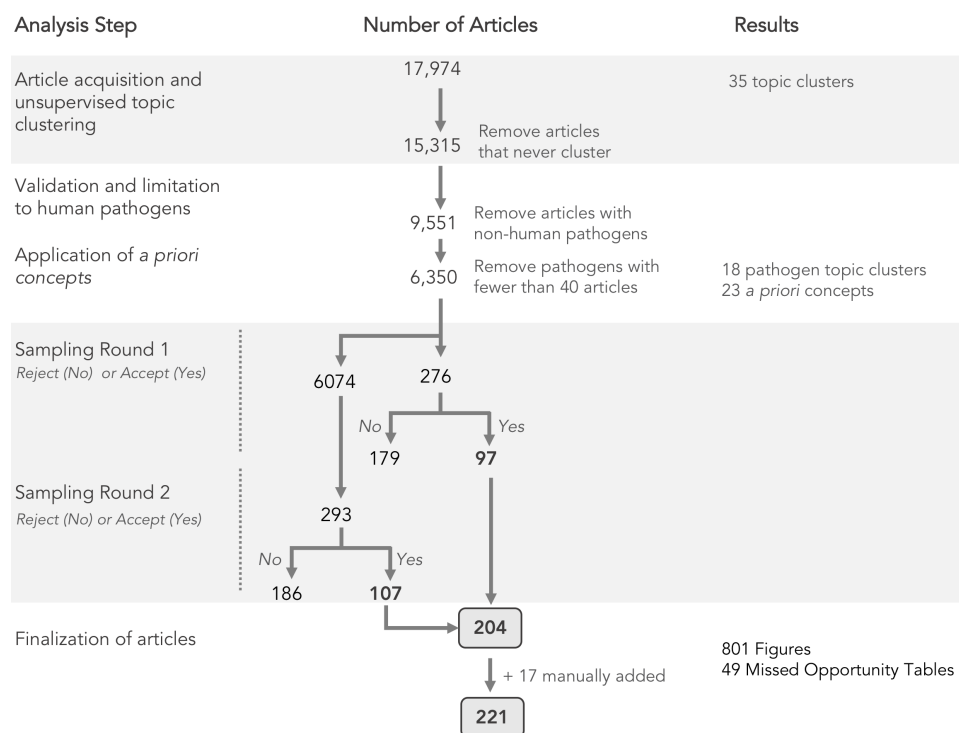


Figure 5.2: Summary of literature analysis steps and document sampling

or never being clustered under any parameter settings (Figure 5.3A). 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 5.3B). 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 C.1³, Figure 5.3C), 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” clus-

³These tables and all others with the prefix C are presented in appendix C

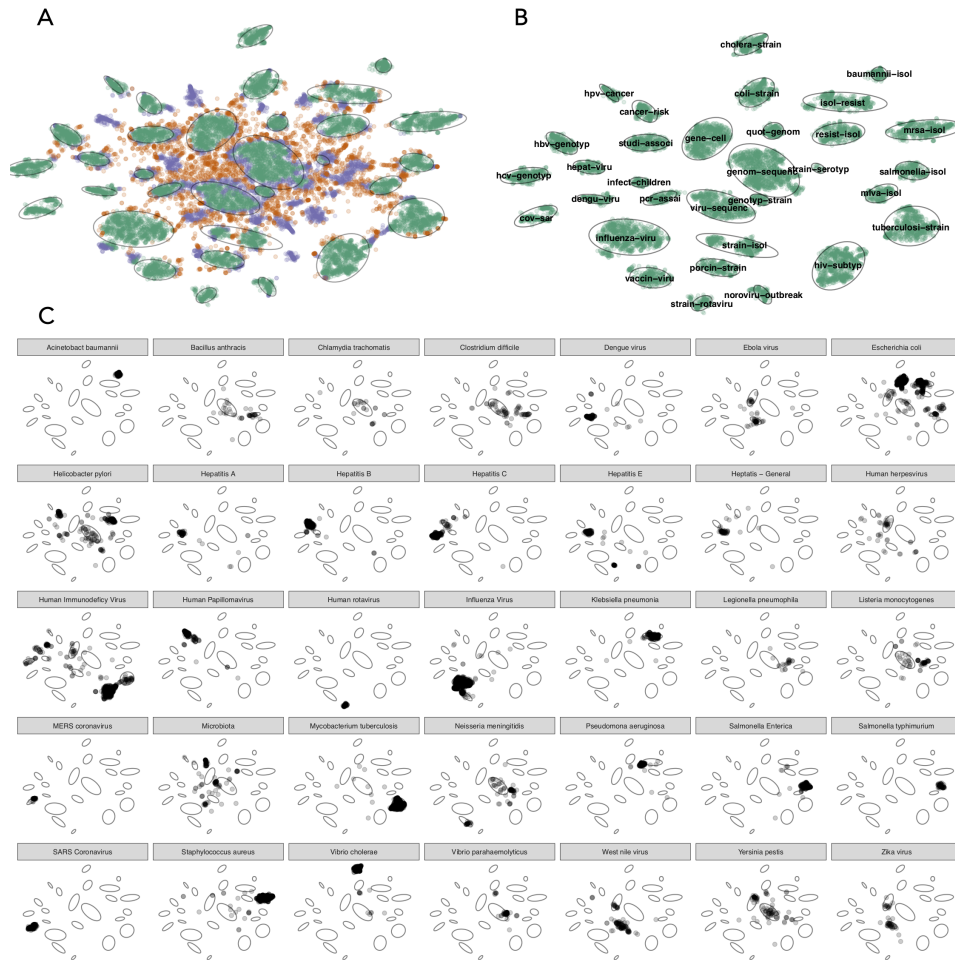


Figure 5.3: Summary of literature analysis results. **A)** Documents were classified according to whether they were part of a cluster (green), unclustered under current parameter settings (purple), or never clustered (orange). The 32 cluster boundaries were automatically determined and are shown as light 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.

ter). Some notable exceptions are *Escherichia coli*, *Helicobacter pylori*, and Human Immunodeficiency Virus in addition to having their own defined cluster, these terms also appeared in other clusters, suggesting co-infections or another phenomenon. We also found that clusters with more generic names (for example “viro-sequenc”, or “geno-sequenc”) contain pathogens that likely had too few articles to form their clusters, possibly reflecting recently emerged pathogens (i.e., Zika, Ebola) with a less extensive research history. 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 (Table C.4).

Clusters were Manually Mapped 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, pathogen-agnostic concepts, such as drug resistance, surveillance, and outbreak investigation. In order to allow researchers to investigate the connections between these familiar *a priori* concepts and the literature-derived clusters, we linked them together manually. We mapped a total of 23 *a priori* concepts to 404 bigrams. We found that *a priori* concepts did not occur uniformly across pathogen clusters (Figure C.3A) 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” (Figure C.2).

Stratified Sampling by 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 mainly from cancer research these are clearly

tagged in our analysis for a total of 221 articles (Table C.1) from which we extracted a total of 770 figures, including a small number (45) of “missed opportunity” tables.

5.3.2 Visualization Analysis

GEViT A Genomic Epidemiology Visualization Typology

Using the analysis set of figures from the sample documents, we used iterative open and axial coding techniques to devise a systematic way to describe how data visualizations are constructed (see appendix C). We began by classifying the types of charts in figures, then classified how charts were combined, and then classified how charts were enhanced. We found that these three descriptive axes allowed us to sufficiently describe all visualizations in our dataset of figures. For each of these descriptive axes, we also derived a hierarchical 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 variance in the resulting visualization design space.

Chart Types in GEViT

We identified eight classes of chart types that form the basis of the data visualizations in our dataset (Figure 5.4): Common Statistical; Colour (statistical charts that intrinsically depend on hue or brightness to convey data); Relational; Temporal; Spatial; Tree; and Genomic. We compiled a taxonomy of common chart names to classify specific instances of chart types within 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 chart 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) (Figure C.4). The frequency of tables, either alone or in combination with another chart type, is a notable finding, indicating missed opportunities for visualization. Our findings also suggest that only a small portion of the available design space is typically used.

Chart Combinations in GEViT

The majority of figures were composed of a single chart type (40.1%), but we observed distinct and common patterns of combining chart types to create more complex, and often linked, multi-part figures (Figure 5.5). **Spatially aligned** charts (20.3%) contained multiple chart types that were aligned along a common horizontal or vertical spatial axis for example, a heatmap and dendrogram are aligned along a horizontal axis to jointly convey clustering information. 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. **Visually aligned** combinations (13.5%) used multiple different chart types that were visually linked for example, using a common colour, shape, or even connection marks to denote some property of the data across the different charts, but are not spatially aligned. Finally, **unaligned** combinations (8.8%) comprise multiple chart types, but where there is no spatial or visual link between charts these likely were combined into a single figure due to manuscript

space restrictions. It was not always straightforward to distinguish between some instances of visually aligned and unaligned, and in such cases, we resolved the ambiguity in favor of the latter classification. We also observed instances of **Complex combinations** (11.9%), in which visualizations used two of the previously described chart combinations. Phylogenetic Trees were the chart type mostly commonly combined with other chart types.

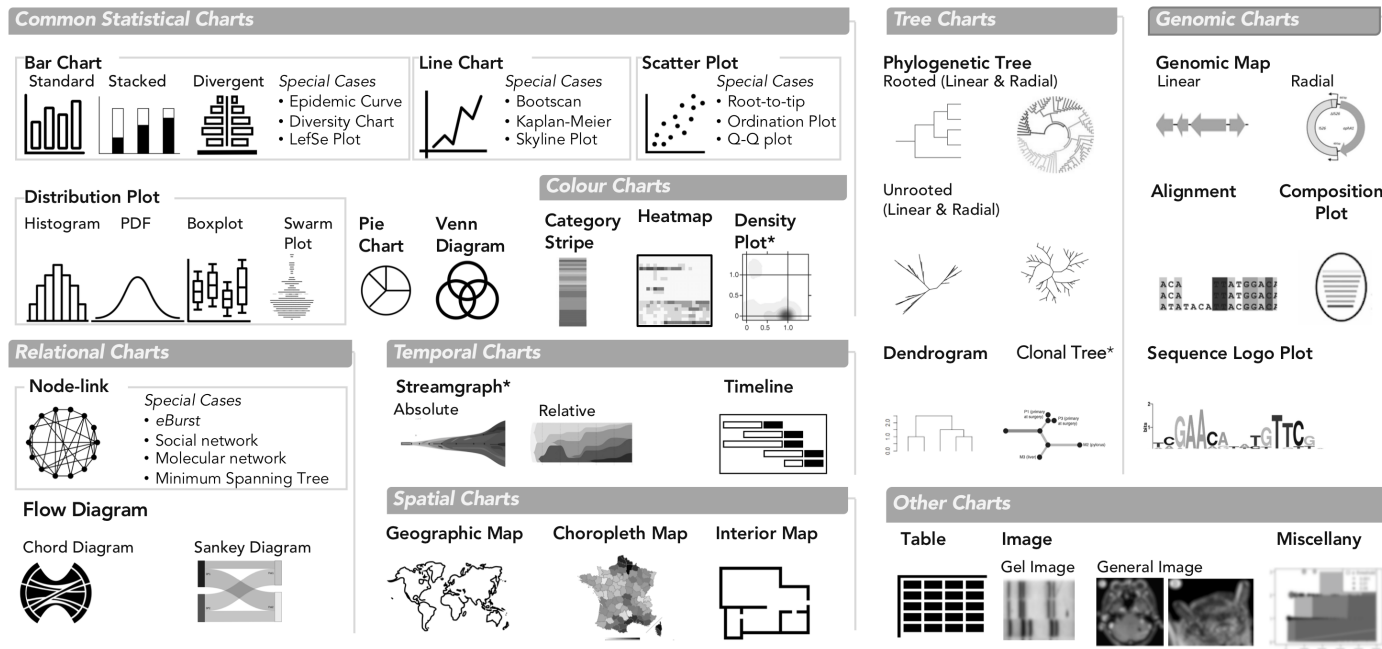


Figure 5.4: Chart Types in GEViT. We used common names for chart types and separated them into eight main classes and also one 'Other' class. Special cases of chart types were defined only when there were multiple instances of the same specific chart across our dataset. Chart types with an asterisk mark (*) indicate that they were included in the analysis through manually added articles.



















Combination Type	# of chart types	# of charts	Alignment Type	Example
Simple	1	1	NA	 OR  OR 
Spatially Aligned	Many	1	Horizontally or Vertically	 AND  = 
Small Multiples	1	Many	Chart Type & Data	 AND  AND 
Visually Aligned	Many	Many	Visual, but not spatial	 AND  AND 
Unaligned	Many	Many	NA	 AND  AND 
Complex Combinations	Many	Many	Context dependent	 AND  AND 

Figure 5.5: 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. Complex combinations are an amalgamation of the above five chart types for example, a spatially aligned visualization that is represented as a small multiple and also linked another chart type.

Chart Enhancements in GEViT

Lastly, we noted that standard chart types were often enhanced to add meta-data 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). Graphical marks are points, lines, areas, and text, which are endowed with the aesthetic properties of size, shape, colour, and texture that can be modified to encode data (Figure 5.6A). For example, a phylogenetic tree encodes evolutionary relationships inferred from nucleic acid or protein data as lines of some calculated length (Figure 5.6B). These lines are often black; however, they can be re-encoded to incorporate data from some additional source for example, colouring lines according to geographic regions. It is also possible to add marks to the base chart type for example, adding coloured point marks to a trees leaf positions (Figure 5.6B), or adding linear brackets and text to delineate or otherwise annotate groups. We did not consider axis text, titles, or data labels to be added marks, subsuming them as constituent parts of the base

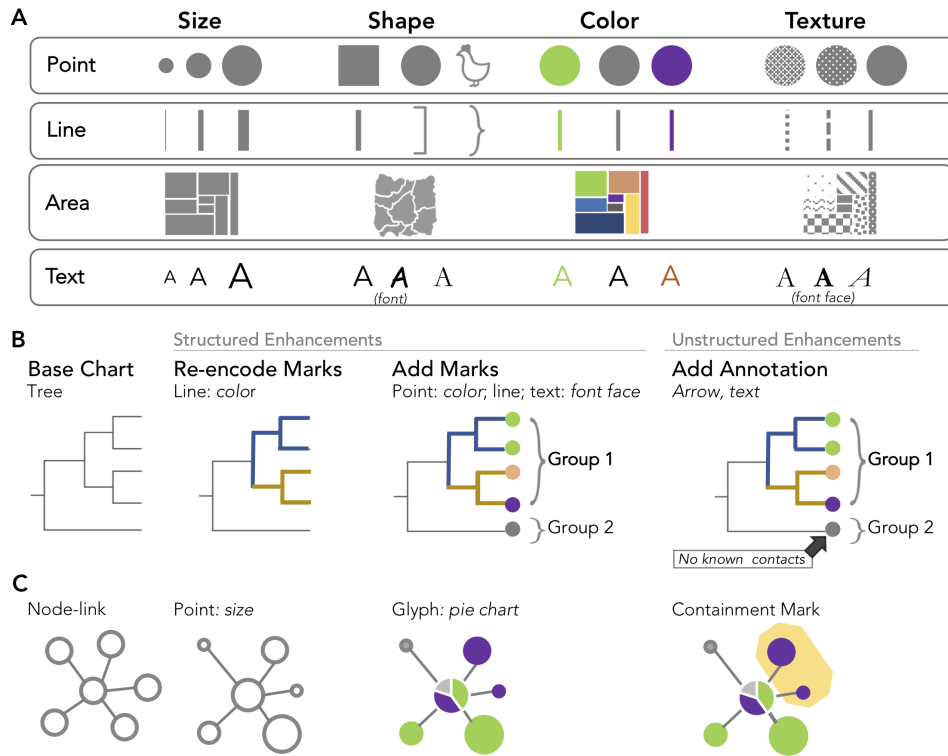


Figure 5.6: Chart Enhancements in GEViT. A) Our characterization of marks and their associated aesthetic properties is based on longstanding conventions in the visualization literature [72, 79] with roots in Bertins Semiology of Graphics [6]. Illustrative examples are shown for B) a tree and C) node-link chart types.

chart type.

It is also possible to add more complex types of marks. 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 spatially aligned 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 5.6B) to communication proportions within 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 instances, we resolved ambiguities by choosing structured enhancement when analyzing figures.

In our dataset, we observed that most base chart types 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 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.

The GEViT Gallery: an Interactive Exploration of the Visualization Design Space

We created a browsable gallery, available at <http://gevit.net>, that allows others to explore the genomic epidemiology visualization design space we created, examine the results of the literature analysis, and browse our GEViT taxonomic code sets. Visitors to the GEViT gallery can browse visualizations across different pathogen types, contextual tags derived from

a priori concepts, and data or arbitrary terms found in the figure captions. Clicking on an individual figure within the gallery reveals its construction via GEViT. Users are also able to browse visualizations based upon GEViT’s taxonomic codes to see the myriad applications for certain chart types, combinations, and enhancements. In our analysis of the data visualizations we also identified examples of good and bad visualization design practice.

5.4 Discussion

Data visualizations are important outputs of many scientific investigations and, when viewed collectively, merit close study to reveal common, good or bad, or missed practices in visualization design. Here, we describe a method for systematically studying data visualizations from the scientific literature, using both text and image data to articulate both why a visualization was created and how it was created from various chart types, combinations, and enhancements. We applied this method to a literature corpus from the domain of infectious disease genomic epidemiology, resulting in the **Genomic Epidemiology Visualization Typology (GEViT)**, and we created a web-based GEViT gallery, allowing users to browse the visualization design space generated through the application of our methodology and to see how elements of our typology can be used to describe a given image.

The typology aspect of our work is similar in spirit to the Grammar of Graphics proposed by Wilkinson [125] and modified and instantiated by Wickham within ggplot2 [123]. While that prior work focuses on low-level details of chart implementation ours uses a higher level of abstraction, using whole chart types as a basis. This higher level of abstraction is more appropriate for exploring and describing the design space used by a community. Our work also differs from existing ontology-based efforts to describe a research domain, such as GenEpiO [46]. GEViT might be considered as the data visualization equivalent to the structured vocabulary that an ontology provides; however, it does not describe the relationships among entities as an

ontology would. However, with future work, incorporating visual typologies into ontologies like GenEpiO is possible.

The present study used data visualizations from articles within the published research literature and did not include visualizations intended for public consumption, not published in peer-reviewed journals. This choice was pragmatic. First, it bounded the search space for our analysis. Academic articles are often accessible through specialized literature repositories, whereas including public-facing visualizations would have required extensive web scraping. Furthermore, research articles are relatively structured, making them an ideal substrate for topic modelling. Limiting our analysis to peer-reviewed scientific literature also bounded the content of the sampled images. There is shared technical knowledge within a research community, meaning that most users can interpret a visualization without additional assistance, whereas visualizations designed to communicate a concept to a more general audience often incorporate additional explanation or background information (although many of these more general, public-facing visualizations likely begin as images created in the academic research context). The typology we developed is extensible to more unstructured, non-academic data visualizations used for public communication, and it would be interesting to compare such a design space with the one we present here. An important limitation imposed by our literature search strategy is that we only included the final data visualization used to communicate some research finding – we do not have access to those data visualizations that researchers created during their internal data analysis process. Our own experiences in public health genomics research and developing data visualizations to share our research findings suggest that the visualizations used during an analysis and those used to communicate final results do not substantially differ, but confirmation of this conjecture would be a good subject for future work.

Another limitation of our current method is the requirement for a human to manually carry out aspects of the visualization analysis. Although this process was time consuming, this inclusion of the human in the loop was

crucial to understand what aspects of each visualization were necessary to delineate a design space; current machine learning methods are not capable of generating such a result. Developing a semi-automatic method that combines some automatically created decomposition of visualizations with human judgement as part of the analysis loop through future work would accelerate the process of refreshing and maintaining the GEViT resource.

5.4.1 Implications of our Findings for Visualization Design

By creating a visualization design space, we not only capture current common practices in our research domain, but we also reveal gaps and areas that require additional attention. While we found some instances of bespoke, effective, and aesthetically appealing data visualizations, the systematic nature of our method to exploring visualization choices reveals that across pathogens and *a priori* concepts, visualization design choices are quite homogenous, and the quality of visualizations varies substantially. We had expected greater variability, given that different pathogens have different transmission routes, are responsive to different interventions, and exist in different environmental, zoonotic, and human contexts. However, phylogenetic trees are the dominant visualization choice, often with additional contextual data included as tree labels or as accompanying tables. This dominance may impact effective knowledge translation in the genomic epidemiology domain, as the interpretability and utility of trees is unclear among public health decision-makers who have limited experience with genomic data [25].

Although our finding of design homogeneity is not surprising, it also underscores how a lack of awareness of design alternatives leads to ineffective data visualizations. For example, geographic data is often encoded as text rather than an alternative mark or an explicit visual representation. The pervasive use of text in genomic epidemiology visualizations stands in contrast to recommendations from the information visualization research literature, where the use of text as a mark type is discouraged. Reading text requires

more working memory and thus imposes a high cognitive load, whereas the goal of most visualizations is to reduce cognitive load by leveraging human perceptual systems to interpret information through the encoding of data as marks and aesthetic properties [79]. Our finding that text was often used as a mark and was endowed with aesthetic properties like colour and variable font faces and sizes suggests that researchers are aware of the power of visual channels, but not necessarily the choice of an effective mark. We also note that many visualizations tended to show all of the data, rather than exploring alternatives that visually summarize data at multiple levels of detail.

Our work highlights opportunities for further work on areas where the genomic epidemiology research community could be better supported in designing data visualizations. Bioinformaticians and software developers can use GEViT to evaluate whether the tools they are creating afford the visual expressivity that infectious disease researchers need to communicate their research findings. Phylogenetic trees are evidently important, but there is a need for better tools that allow researchers to explore alternative visualizations and to more effectively encode tabular metadata onto trees and other visualizations. Although our study did not reveal how researchers create their data visualizations, our experience in the genomic epidemiology research community suggests that many chart- or tree-generating packages, some in R, are often used in conjunction with Power Point or Adobe Illustrator to compose complex visualizations that include chart combinations and enhancements. Software tools or libraries that support more expressive generation of visualizations can lower the barrier to generating data visualizations, reduce the overreliance on text, expand the use of combinatorial charts, and contribute to more reproducible research by creating more informative visualizations in which data is not obscured.

We also suggest that our findings might inspire developers to create alternatives to existing common design choices, and that our gallery of visualizations gives such developers a resource with which they can empirically test their new visualization design against existing choices. This empirical

approach to testing a new visualization will help move the community further away from the *ad hoc* approach to visualization development, where design choices are heavily biased by individual preferences. As more work is done to explore and test new visualization designs, GEViT will incorporate these designs, potentially resulting in the addition of new typological terms. It will also be interesting to explore how GEViT might be used to suggest visualizations to researchers, as is currently done with common statistical charts in tools like Tableaus “Show Me” feature [69], Google Sheets’ “Chart Suggestions”, or in novel systems like Draco [76].

5.4.2 Implications of our Findings for the Genomic Epidemiology Community

Data visualization can be an important tool for translating scientific results to a group of experts working in a common domain but with varying backgrounds. This situation is often the case in public health genomic epidemiology, where microbiologists, computational biologists, clinicians, epidemiologists, healthcare administrators, and others often come together around a specific issue. By making individuals aware of data visualization conventions used by the community through the GEViT gallery, we hope to assist researchers who struggle to visually communicate their research findings by providing both inspiration and a framework for reasoning about data visualizations that will assist as they develop their own data visualization practice. We have tagged examples in the GEViT gallery with good and missed opportunities to provide some guidance, but these labels are assigned by our subjective reasoning as data visualization experts and have not been empirically validated. Future work in this area might include quantitative evaluation of the efficacy of particular visualizations, and ultimately more sophisticated guidance around visualization design and analysis in the public health context.

5.5 Conclusion

Through a systematic method, we have delineated the visualization design space used in infectious disease genomic epidemiology. We provide both a concrete terminology for describing data visualizations and a gallery of visual inspiration, the combination of which we hope will provide guidance to visualization tool developers and to researchers looking to create their own visualizations. Mostly importantly, our work demonstrates that is possible to think systematically and rigorously about data visualizations and that there exist open, complex, and interesting problems in visualizations design and analysis, where the potential impacts on research domains such as public health are profound.

Chapter 6

minCombinR:

Coordinating Chart Combinations with Minimal Specifications

The ability to simplify means to eliminate the unnecessary so that the necessary may speak — Hans Hofmann

¹ Domain experts routinely use coordinated combinations of static charts to communicate their findings. In a prior study we identified and quantified different strategies for combining multiple types of charts and we assessed that existing data visualization tools and charting libraries did not provide consistent support for these strategies. While some domain experts could programmatically generate these combinations on their own, many more would be left to resort to manually combining charts through post-processing, which can introduce errors and impact the reproducibility of analyses. We have developed minCombinR, an R-based toolkit that implements a minimal specification syntax for a wide range of chart types and their combinations. minCombinR attempts to balance ease-of-use and expressivity with the ability to link to nuanced data types and analyses that users perform. Through minCombinR, domain experts use a consistent specification syntax for creating four types of combinations (small multiples, spatially aligned, colour aligned, and unaligned), across nineteen distinct chart types ranging from common statistical graphics to phylogenetic trees to geographic maps to

¹This chapter has been submitted for publication [27]:

A. Crisan, S. Fisher, S. Kasica, J.L. Gardy, and T. Munzner (2019). minCombinR: Coordinating Chart Combinations with Minimal Specifications.

images. We demonstrate the capacities of minCombinR using both real and synthetic datasets from multiple domains. We have implemented minCombinR using a bottom-up approach to system development, using an existing domain-specific visualization typology to inform a general architectural design. Our system, analysis artifacts, and bottom-up approach straddle a middle ground between a general top-down approach and specific bespoke systems tailored to the domain as undertaken in design studies.

6.1 Introduction

Many domain experts have analysis needs that require understanding linkages between information appearing across different types of charts generated from heterogeneous data sources. Visualization researchers have long advocated supporting view coordination through interactive techniques such as brushing [5] to link between charts. The substantial power and flexibility of interactive linking between views has led to their support in many visualization systems such as Vega/Vega-lite [100], stand-alone analysis systems such as Tableau [112], and notebook environments [8]. Despite the utility of interactivity for exploration, there are many situations where static charts are the only option for presentation; most notably, in traditional publications such as research journals. With a few recent exceptions, such as the online Distill journal, static charts continue to dominate both the analysis process and the communication process of reporting findings to the scientific community and for general consumption. Although the coordination of static charts is often possible with sufficient effort in current software systems, it could be made easier with better infrastructure support.

Moreover, we note the many costs to interactivity, both temporal and cognitive [61]. A particular concern in scientific analysis settings is potential threat to reproducibility. When users conduct analyses through interactive interfaces they may forget the details of their past actions and draw conclusions without a full awareness of how they arrived at them, making it

difficult to assess their validity and robustness. Despite active research toward data and analysis provenance [111] to address some of the challenges introduced by interactivity, it is unclear whether these methods handle the complex heterogeneous data types and analysis procedures that are used in scientific analyses. As an alternative to interaction, users may manually post-process results to achieve some type of coordination that their tools do not support. This common practice [7] is problematic for scientific reproducibility because it can be a source of inadvertent errors, or even intentional manipulation of findings.

In this paper, we address the challenge of facilitating coordinated static combinations for two or more complex chart types. While there exists some support for combined charts in various systems, toolkits, and charting libraries, static combinations are not consistently supported or sufficiently expressive of the full range of combinations a user could produce, and at present may require considerable coding effort from the user to be achieved. Furthermore, combinations have been principally explored for statistical chart types, with little previous work addressing how combinations for more complex chart types should be supported. We build on a recent analysis of a domain-specific visualization design space that produced a visualization typology for describing and enumerating the types of charts, their combinations, and enhancements, that experts in the domain of genomic epidemiology commonly used for communicating their scientific results [24]. Its findings highlight the lack of support for creating coordinated chart combinations.

To address this unmet need, we have developed a toolkit called minCombinR that supports the coordinated combination of static complex chart types in R. In the design and implementation of this system, we demonstrate how we can use domain-specific visualization typologies to inform a bottom-up approach to development of data visualization systems. While rooted in a domain, this approach identifies real world problems and challenges that are used to inform, rather than constrain, the architectural design in a way that is generalizable. Our bottom-up approach is in contrast to the more

frequent top-down approach to system development, where a generic system is implemented first and later tailored or optimized for specific applications.

We contribute minCombinR, a system that supports minimal specification syntax for singular charts and their combinations in R. Our implementation of minCombinR is based on a visualization typology that we previously developed for a specific domain (see Section 6.2), but we have architected our system in a way that is broadly extensible. Key features of the minCombinR architecture include:

- a declarative framework for gradual binding that integrates user-defined and programmatically-derived specifications,
- support for a broad variety of singular chart types including spatial maps, trees, node-link networks, genomic charts, timelines, and annotated images in addition to common statistical charts, and
- control-flow algorithms for enforcing positional and colour consistency within small multiple, spatially aligned, colour aligned, and unaligned combinations of charts.

We also contribute analysis artifacts for the minimal specification requirements of different chart types and their combinability that could be used in future systems with a different algorithmic framework.

6.2 Domain Motivation and Design Decisions

Our previous work in the domain of genomic epidemiology visualization was a major motivation for minCombinR, and inspired our three major design decisions.

6.2.1 GEViT Findings

We undertook a systematic review of public health researchers studying microbial genomic epidemiology to identify common practices, assess good strategies, note areas for improvement, and generate evidence for the development of new idioms that could better address a common set of stakeholder tasks [24]. We developed a novel method for systematically appraising the design space and used it to generate a Genomic Epidemiology Visualization Typology (GEViT) that we used to annotate a dataset of roughly 800 figures extracted from research articles written by domain experts. GEViT attempted to summarize *why* a visualization was generated by using text analysis to ascertain its creation context and also *how* a visualization was constructed through the development and then application of GEViT to our dataset of figures. We also strategically sampled across the design space so that we could quantify *how many* instances there were of specific types of visualizations. We refer the reader to the prior publication for more details on methods and the justification of this approach [24].

GEViT summarized an expert-defined visualization design space and broke down how visualizations were constructed according to individual chart types, enhancements to those chart types by adding or re-encoding marks and channels, and finally how these individual chart types were combined. For combinations, we observed four distinct patterns of chart combinations that were in common use: spatially aligned, colour aligned, small multiples, and unaligned (formerly presented as composite, many type linked, small multiples, and many types general, respectively). Overall, our analysis revealed a complex combinatorial design space, but also quantifiably showed us that only a very limited range of that design space was explored. Perhaps the most surprising finding was how much data existed in text in the form of complex labels or tables that were deliberately *added to a visualization and existed as a part of the figure*. We also made some qualitative assessments as we reviewed these figures in the form of speculations about the process

used to create them. While it is challenging to prove definitively, we noted signs of post-processing when experts were trying to enhance or combine chart types. From these qualitative findings and our knowledge of existing visualization systems, we hypothesized that experts were not well supported to create the kinds of visualizations they wished using existing software, and resorted to putting data into text and doing extensive post-processing to achieve their objectives.

Here, we take a step toward helping users address their data visualization challenges. Rather than introduce new visual idioms, we determined that it would be a prudent first step would be to develop a system to help users generate and link together charts based on existing visual idioms more easily, within a platform that is integrated with their complex and diverse analysis needs.

6.2.2 Design Decisions

We identify a set of design decisions that address the limitations of existing systems. These decisions are originally motivated by the findings in the GEViT study, but we conjecture they are relevant for many other domains.

- **D1: Support a broad variety of chart types.** Experts produced complex chart types from multiple heterogeneous sources of data, so helping experts to visualize their data necessarily involves helping them to integrate multiple different chart types.
- **D2: Automatically harmonize static charts into consistent combinations.** Many previous systems focus on single charts, but with insufficient support for users who want to maintain consistency across multiple charts [91]. We would like combinations to be “first-class citizens” that are easy to generate from simple and minimal specifications, with toolkit support to enforce the consistency required by each combination type. While considerable visualization research has been

devoted to view coordination through linked interaction, combinations of static charts remain surprisingly difficult to generate.

- **D3: Integration with analysis.** Experts use complex data within nuanced analysis pipelines; a visualization system that is easy to integrate with these existing pipelines will lower the barriers for its use and improve the reproducibility of their work.

The GEViT findings led us to choose four types of combinations to support in minCombinR. **Small multiple combinations**, also known as faceted charts, are a well known approach where multiple instances of the same chart type are each used to show different partitions of a dataset. **Spatially aligned combinations** feature vertical or horizontal alignment where a common attribute is used to encode spatial position within each chart, and moreover the spatial arrangement between charts is also constrained to be consistent along the same axis. **colour aligned combinations** impose a consistent colour scheme across common attributes in different chart types. Finally, **unaligned combinations** have no constraints to share common attributes, and simply show multiple chart types in any kind of arrangements.

6.3 Related Work

We situate the minCombinR system in the context of prior research on general-purpose visualization systems as well as specific applications for visualization in biology. In comparing the contributions of these systems to our own we consider the following factors:

- **Ease of use:** how hard or easy it may be for experts to use this system. Our quantitative proxy for ease of use is the amount of code a user has to write, where minimal code corresponds to high ease of use and a system that requires extensive coding has a lower ease of use.
- **Expressivity:** the breadth of a design space that can be supported by a

system. We consider expressivity across a large class of design spaces, only one of which is GEViT.

- **Links to analysis:** the breadth of analysis procedures that a system can link to within an integrated programming environment. We consider both pre-existing analyses that are built into the system itself and customized analysis procedures that an expert could implement without leaving the environment.

Although experts could vary considerably in their technical skills, we make our assessments assuming that the average expert is able to write some code, is familiar with their own data, and has knowledge of their data analysis pipelines. We group prior work into the following categories for analysis: stand-alone applications, charting libraries (specifically with JavaScript, Python, and R), and domain-specific tools for microbial genomic epidemiology since we do make specific claims in our results for this domain.

6.3.1 Stand-Alone Applications

The most widely used of the many existing stand-alone applications used for data visualization are Excel, Tableau, and PowerBI. Tableau and especially Excel are commonly used among public health experts and are thus an important baseline to compare against. Although all of these applications require some learning time, they are both designed to be and generally perceived to be easy to use. The monolithic architecture of these applications can facilitate view coordination, but also imposes limitations to expressivity. The majority of existing stand-alone applications support tabular data and the generation of common statistical charts such as bar charts and scatterplots, and some types of geographic chart types such as choropleth maps or scattercharts with maps as a base. These tools also primarily support generating singular chart types with varying degrees of interactivity, with some support for linked views and small multiples (Excel excepted). Extending these applications to support different data types or more complex visual idioms

is possible in some cases, for example through Tableau’s “extensions” and PowerBI’s “marketplace”, where the development and integration of novel visual idioms can be shared with the wider community. These paths to increased expressivity are technically possible, but the ease of use for coding them is very low. Finally, these stand-alone applications support analysis to various degrees. Excel is a quite powerful analysis tool, while Tableau and PowerBI support some analyses natively and also make use of extended features to integrate other types of analysis or even link to analytic scripting languages like Python or R.

This class of systems all involve a fundamental architectural design choice that there is a major separation between two regimes: default functionality that is built in, vs. extensibility only through plugins. In contrast, minCombinR architecture does not have this pronounced distinction, so functionality can be extended in a more integrated way.

6.3.2 Charting Libraries and Packages

Systems implemented as graphics or charting libraries exist widely within many programming languages. Compared to stand-alone applications all of these systems have a lower ease of use, with varying level of difficulty depending on the library. Within JavaScript, D3 [9], Vega [99], and Processing.js [39] have lower ease of use compared to Vega-Lite [100] or Plotly [90] (also instantiated within Python and R), considered in terms of the number of lines of code to be written. Chart combinations are primarily supported through interactive linked views and brushing, although Vega-Lite does support some types of combinations (concatenation, layering, and faceting) that are akin to minCombinR combinations. Furthermore, Vega-lite can perform an automatic resolution of scales, but its scope appears to be more limited than minCombinR’s automatic scale derivation. Finally, the link between analysis and visualization with JavaScript libraries is tenuous. Observable notebooks [8] have made it easier to perform browser-based analysis and

more libraries are being developed to support analysis in the browser. However, at this time the more common paradigm is to use another language such as Python for analysis, with output in a format that can be readily used by JavaScript libraries for visualization.

Programming languages that are analysis driven also contain their own libraries for data visualization, although as with JavaScript tools the ease of use is lower than with stand-alone applications. Python has matplotlib [56], bokeh [86], and seaborn [121] as its most commonly used visualization libraries. Recently, Altair [117] has been developed on top of Vega-Lite to support data visualization for Python. Within the R language, both the base graphics and ggplot [124] are widely used. These two libraries are built on different graphics systems that exist within R, so they are not easily integrated. R also has a means of incorporating JavaScript libraries and there are active ports of visual idioms created in D3 and even Vega-Lite. As with the JavaScript libraries, these analysis languages primarily produce static chart types (although interactivity is possible), and have variable support for combinations; there appears to be the greatest support for small multiples.

Graphics and charting libraries present systems with ultimate flexibility and a more integrated link to analysis relative to stand-alone applications. However, this flexibility does come with high cost in terms of ease of use: we have observed that many more complex chart types and combinations are not easy to generate. We have built minCombinR on top of R, a common analysis scripting language, to be able to directly integrate our toolkit with analysis while lowering the burden of creating coordinated combinations of multiple complex chart types.

6.3.3 Domain-Specific Tools

A third class of systems is domain-specific tools for biological data. Generally, these tools are meant to support very specific tasks, which limits their expressivity. They are also connected to specific data analysis pipelines or

are only able to take in already-completed results for display. Amongst the community of microbial genomic epidemiology researchers, Nextstrain [48] and Microreact [4] are the most commonly used tools. These browser-based applications are connected to complex data analysis pipelines and are meant primarily to communicate findings. Phylogenetic trees are by far the most commonly use chart type in this community and a variety of tools exist such as TreeViewer (stand-alone, [101]), Baltic (python, [33]), ape (R, [87]), and ggtree (R, [135]). There are also a number of stand-alone tools outside of the scope of genomic epidemiology that support the analysis and visualization of genomic data in particular. The most pertinent to the scope of our design space is Stack’n’flip [113], which implements an interactive method for linking information across heterogeneous biological datasets but does not cover the full breadth of chart types and combinations as minCombinR. While these toolkits reflect the current state of the art for this domain, they are limited in their expressivity compared to minCombinR.

6.4 Design of minCombinR

We describe the major challenges of creating a toolkit from a typology and present an architecture to address them. We then walk through each architectural layer in detail.

6.4.1 From Typology to Toolkit

The typology that inspired us was created by analyzing existing images. Its description of how individual static charts can be combined into coordinated combinations offers no guidance on how to support these capabilities with an easy to use software architecture that leverages existing charting libraries. GEViT describes charts in terms of some standard configuration that can be augmented by adding additional marks, or changing existing marks. The idea of creating a single chart and then changing it to co-ordinate with others does

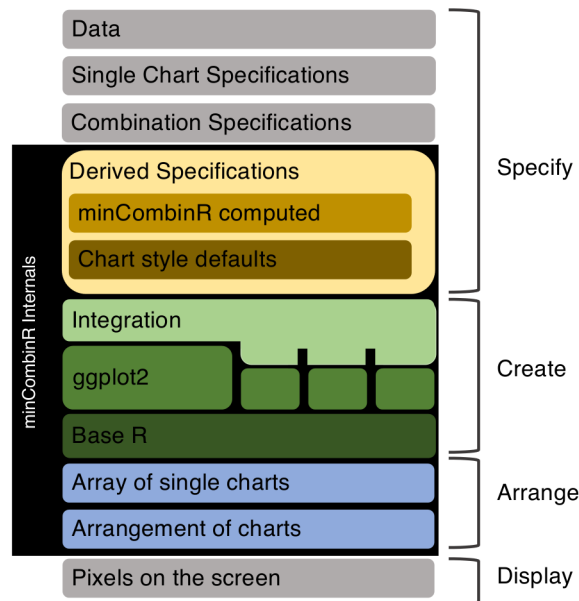


Figure 6.1: Architectural layers of minCombinR.

not map easily onto existing charting libraries, which render a singular chart as an immutable *box of pixels* and leave the burden of coordinating multiple charts to the user. While it is trivial to concatenate charts together into an unaligned combination, the more interesting cases require **automatic chart harmonization** where consistency is enforced between visual encodings, which must occur in advance of rendering the combination.

We solve this problem with a declarative approach that relies on **gradual binding**: specifications are generated and modified in discrete stages, after which a final specification is passed along to existing imperative chart libraries for rendering. The user declares only minimal initial specifications for singular charts and for a requested combination. The system then derives additional specifications that enforce the necessary consistency within each chart, so that all of the rendered boxes of pixels can be trivially concatenated together into a large box of pixels with the desired characteristics.

Specifications are a declarative language that tell minCombinR what to gen-

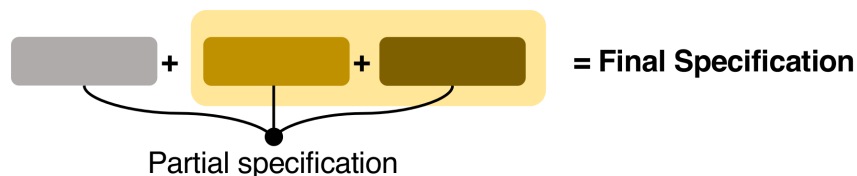


Figure 6.2: User and derived partial specifications. A final specification is a combination of partial specifications from the user (grey), derived specifications from minCombinR (tan), and stylistic specifications inherited from different charting libraries (brown).

erate. A key design choice is that users specify singular charts by designating a chart type, which is easy for domain users who typically reason about visualizations at the level of choosing between different chart types. Although recently the language of marks and channels (called visual attributes within R) has become somewhat more widely used beyond visualization research, thanks to the Grammar of Graphics [125] and D3 [9], making such choices effectively requires a level of knowledge about, interest in, and attention to visualization best practices that is not realistic to assume for most domain experts. Systems such as Vega-Lite [100] and ggplot [124] that require specifications of marks and channels have a higher specification burden for users.

6.4.2 Gradual Binding Architecture

The architecture of minCombinR is composed of four layers (Figure 6.1): *specify*, *create*, *arrange*, and *display*. The **specify** layer takes initial partial specifications from a user for both singular charts and requested combinations, and integrates them with computationally derived specifications that align the appearance and scales of charts automatically. The **create** layer generates singular charts from those specifications, wrapping around existing chart libraries. The **arrange** layer will position singular charts according to the combination specification. The **display** layer renders the arranged charts to the user. All of the create and arrange layers are internal to minCombinR;

in Figure 6.1, only the grey sections in the specification and display layers are exposed to the user, while all of the coloured sections contained within the black enclosing rectangle are internal.

6.4.3 Specification

The gradual binding process starts with initial user partial specifications, first for singular chart types and then their combinations. These are augmented with minCombinR-derived specifications, and finally with style specifications inherited from underlying chart libraries (Figure 6.2).

Our goal is that the initial partial specifications that the user must provide should be as simple as possible. For each of the supported chart types, we conducted an analysis of the minimal specifications that a user must provide in order to generate it. The results are recorded as table for each chart type in Appendix D. This information is used within minCombinR to help the user generate valid specifications, both to document what should be provided and to generate warning messages when the user specification is insufficient. It is straightforward to extend the system with new chart types, but the existing minimal specifications persist unchanged.

For each type of combination, minCombinR derives additional specifications.

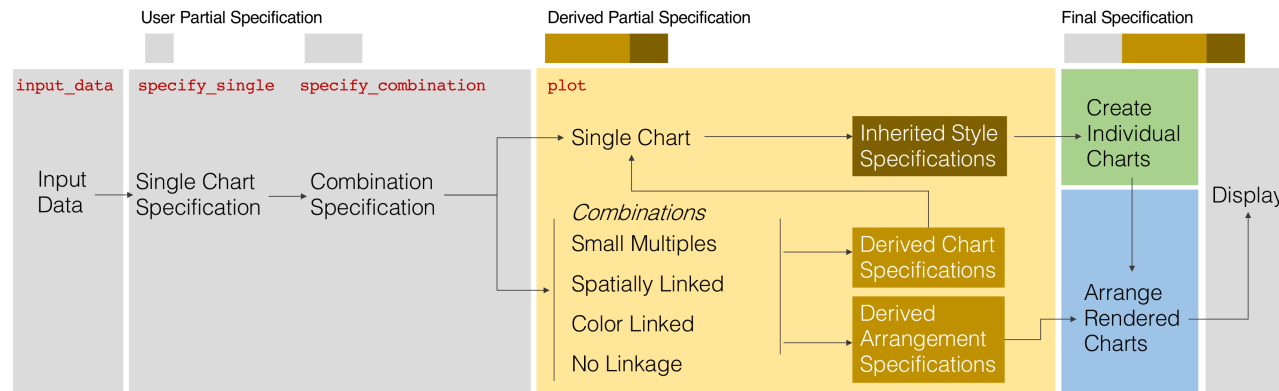


Figure 6.3: Overall flow of specifications and control in minCombinR. The sequence is colour-coded by layer as in Figure 6.1: user-facing specifications and display in grey, derived specifications in yellow, the final charting library specification in green, and arrangement in blue. The red minCombinR functions gather specification details from the user through `specify_single` and `specify_combination`, and `plot` triggers the internal computations. The process of **gradual binding** flows from initial specifications from the user, followed by those derived by minCombinR, and finally with stylistic specifications inherited from underlying packages.

Spatially aligned combinations guarantee that within charts there is a common spatial axis along either the horizontal or vertical direction, and also that the rendered charts are arranged in a linear 1D array in that direction. We conducted an analysis to determine which chart types are spatially alignable, with the results recorded as a matrix for every possible chart pair (see Appendix D). We automatically scan the positional encoding specifications of chart types to establish a shared axis to align on. This axis alignment is done in two steps. First, the algorithm will check whether charts have the same underlying data source and if so, will then establish whether the positional encoding channels contain the same variable name. In instances where data comes from different sources, minCombinR will try to detect if the attribute fields assigned to positional encoding variables between the two chart types are in fact the same. minCombinR will analyze all fields that are assigned to positional encodings of two or more chart types, and if these fields are determined to be identical, these charts are established to have a shared axis for spatial alignment. minCombinR will only discover shared axes using categorical attribute fields; fields with numerical data do not contain enough contextual information to disambiguate whether an axis is shared without additional user input.

Once a shared axis of alignment has been established, minCombinR will automatically specify the correct orientation for individual charts so that the shared axis is properly oriented with the direction of alignment. For example, if all charts are to be aligned horizontally, then the shared axis attribute field should be mapped to y-axis positional encoding. If a chart happens to have the shared axis attribute field in the x-axis instead, then minCombinR will 'rotate' the chart such that the attribute field is now mapped to the y-axis positional encoding. As a final step to facilitate the spatial alignment, minCombinR will also derive and apply a common scale to the shared axis. In our analysis of chart types and combinations, we identified a set of charts whose positional coordinates cannot be altered because it would inappropriately distort the information in the chart; these are tree, geographic

maps, and images (with some exceptions). We refer to the positionally immutable charts as *lead charts* and align all other *support charts* to the axis scales of the lead chart. All support charts are harmonized to lead chart.

In addition to automatic chart harmonization, minCombinR also provides helpful guidance by assessing when chart types are not combinable and automatically dropping any non-combinable chart types from the specification (with a warning to the user). Even when charts types are combinable, minCombinR will also automatically drop charts from the combination specification that do not have a shared axis for alignment. If the user provides a specification with more than one lead chart, minCombinR will provide suggestions for alternative specifications that are appropriate for each of the individual lead chart types. The analysis reported in Appendix D also includes the logic behind all of this automatic guidance.

Colour aligned combinations will automatically apply a common colour palette for shared attributes across multiple charts. minCombinR will assess whether all charts have some common variable to connect on, similarl to the approach to finding a shared attribute field between charts for spatially aligned combinations. However, unlike spatial combinations, which need to have positional encodings specified, colour aligned combinations can *add a colour encoding* or even *overwrite an existing colour encoding mapping*. For example, two or more chart types may have position encodings that are mapped to specific fields in the data, but a colour encoding is possible through some other attribute field that those charts share. In this case, minCombinR will simply map that attribute field to the colour encoding channel for all charts and ensure that these charts use a common colour palette. In another instance, two or more charts may already have both positional and colour encoding channels mapped to specific attribute fields in the initial partial specification, but the user’s combination specification may request that the charts should be colour aligned by a common attribute field. In this case, minCombinR will override the existing colour encoding mappings and modify the specification with a different colour coding using

the requested field. Users are warned whenever minCombinR overrides initially specified defaults.

Small multiple combinations are supported natively by many charting libraries, but minCombinR overrides those defaults in order to have a unified interface for all types of combinations. Given a chart type and faceting variable, minCombinR will derive and apply a common scale to the singular charts so that they have shared axes. Similarly to the spatial alignment analysis for positionally immutable charts, we have also identified a set of chart types (node-link diagrams, trees, and maps) where maintaining common scales is vital for correctly constructing the facets. For example, a phylogenetic tree must be constructed using all of the data because the underlying tree structure would be inaccurate otherwise, but each facet is meant to highlight a specific subset of the data. By comparison, a bar chart can be generated using only a subset of the data for each facet. minCombinR can distinguish between these two scenarios for small multiples and automatically selects the best approach for a specific chart type.

Unaligned combinations refer to situations where a user may simply wish to combine multiple charts without any constraints. The charts may share a common underlying data source but not a common axis or attribute to facilitate other types of combinations, or the user may simply need to conserve space by combining unconnected multiple subfigures together. Unaligned combinations are also already commonly supported by many charting libraries, but again we override those defaults in favor of a unified interface for combinations.

6.4.4 Creation and Integration

The architecture of minCombinR is designed to wrap around existing charting libraries within R and flexibly adapt to new ones. We achieve this flexibility in minCombinR by gradually binding user and derived specifications to a common set of generic encoding parameters that are mapped to

library-specific conventions in the integration layer. These generic parameters are standard combinations of mark and channel encodings (for example, point colour) and also more generic parameters (for example, colour without a reference to a mark). There are also internal parameters that are not exposed to the user and are employed to coordinate harmonization across combined charts, for example what the alignment axis is, if and how charts should be oriented, if and how chart axes should be scaled, and so on. This final specification is composed of encoding and internal parameters that are then passed to the relevant charting library to produce a single chart. The integration layer also manages details of the aesthetic appearance of individual charts. For example, in spatial combinations it automatically drops the axis labels for the shared axis so that it is not repeated in every chart.

As a final step, the `minCombinR` specifications are converted into a final, R package dependent, specification that is used by the R graphics systems in the display layer to render the individual chart types into an immutable box of pixels. The final output from the integration layer is an array of final specifications of size N , where N is the total number of singular charts that a user has specified. This array of individual chart specifications is passed to the arrangement layer, along with the combination specification that indicates the desired configuration.

Rendering in R entails integration challenges because R has two primary graphics systems, referred to as `base` and `grid` graphics. `minCombinR` currently wraps around `ggplot` and its extended universe of packages, which we call the *gguniverse*, all of which use `grid` graphics. Integrating `base` and `grid` graphics is possible, but difficult to engineer for in a consistently reliable and robust fashion, so the initial `minCombinR` implementation only encompasses the *gguniverse*.

6.4.5 Arrangement and Display

Arranging and rendering charts into the configuration required by the user's combination specification request is straightforward because the final specifications charts emerging from the integration layer are already harmonized, from the gradual binding of the individual chart specifications in the first layer of the minCombinR architecture.

The boxes of pixels that represent rendered individual charts are arranged in a grid. For most combinations the default grid has three columns and an arbitrary number of rows depending upon the total number of charts N . The exception is spatially aligned combinations, which in the horizontal case are arranged in a single row and N columns or in the vertical case are arranged in a single column and N rows. The order of the chart arrangements depends on the combination type. Spatially aligned charts are ordered with the lead chart first and support charts following in the order that they were specified. Both colour aligned and unaligned combinations are arranged in the order that they were specified. Small multiples are ordered according to their individual facets, by default in alphanumeric order.

The output of the arrangement layer is a large box of pixels, which is simply passed to the currently active R display device.

6.5 Implementation

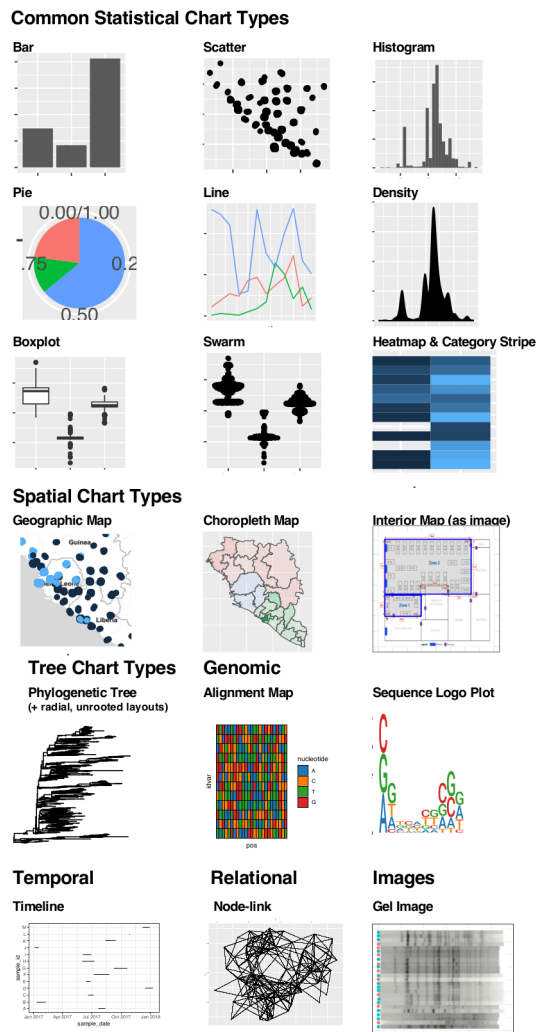


Figure 6.4: Currently implemented chart types in minCombinR.

Our initial implementation of minCombinR supports a subset of 19 of the 25 chart types that we identified in our prior GEViT study, enumerated in Figure 6.4. This choice provided both a reasonable limit for the implementation scope and a sufficiently broad context to assess the expressivity of our

approach. We have deployed minCombinR as an open-source R package available at: <https://github.com/amcrisan/minCombinR>.

We provide details on the specific chart types that are currently supported by minCombinR, along with the algorithmic implementation details of the different types of chart combinations. In Appendix D we provide a complete list of all the R base charting libraries used to generate the individual chart types. At this time, minCombinR requires that individual chart types generated by the integration layer are compatible with ggplot2 [124]; despite this compatibility, its extensions do require extra steps to facilitate the integration, as communicated by the panhandle shape of this layer in Figure 6.1.

6.5.1 User Functions and Specifications

Users can specify chart types, specify combinations, and display their results using the three main functions: `specify_single`, `specify_combination`, and `plot`, as shown in Figure 6.5 and Figure 6.3.

`specify_single` requires that users define a chart type and a data source, possibly with additional required encoding parameters depending upon the different chart types, and returns an initial partial specification. We have implemented helper functions that allow users to identify the full range of supported chart types and their minimal specification requirements. Users are warned if they do not provide the minimal specification requirements for a chart type.

`specify_combination` takes as input the single specifications of two or more charts, a type of combination (small multiple, spatially aligned, colour aligned, and unaligned), and an additional parameter specific to a chart type. For small multiples, this additional parameter is the attribute field by which to produce the facets. For spatially and colour aligned combinations, it is the attribute field used to facilitate the alignment. The function returns a

valid combination specification. As indicated in Section 6.4.3, the system provides detailed guidance with suggestions on how to achieve a valid specification in any warning message indicating the original request cannot be accommodated.

plot will take a minCombinR specification for a singular chart or a combination as input. This function also assumes that the input minCombinR specification is valid, since the prior two functions guide the user if any problems surface. The plot function is the workhorse of minCombinR. The previous two specification functions only provide initial user partial specifications and do not trigger substantial computation. When the plot function is called, chart harmonization is carried out, additional specifications are derived, and then charts are arranged, rendered, and displayed.

In addition to minCombinR specification and plotting functions, we have also implemented additional helper functions that assist users to load heterogeneous data, annotate images that they intend to combine with other chart types, and consolidate spatial data from multiple different sources. Within minCombinR's online repository are analysis notebooks that demonstrate how to use minCombinR's functions within the R environment.

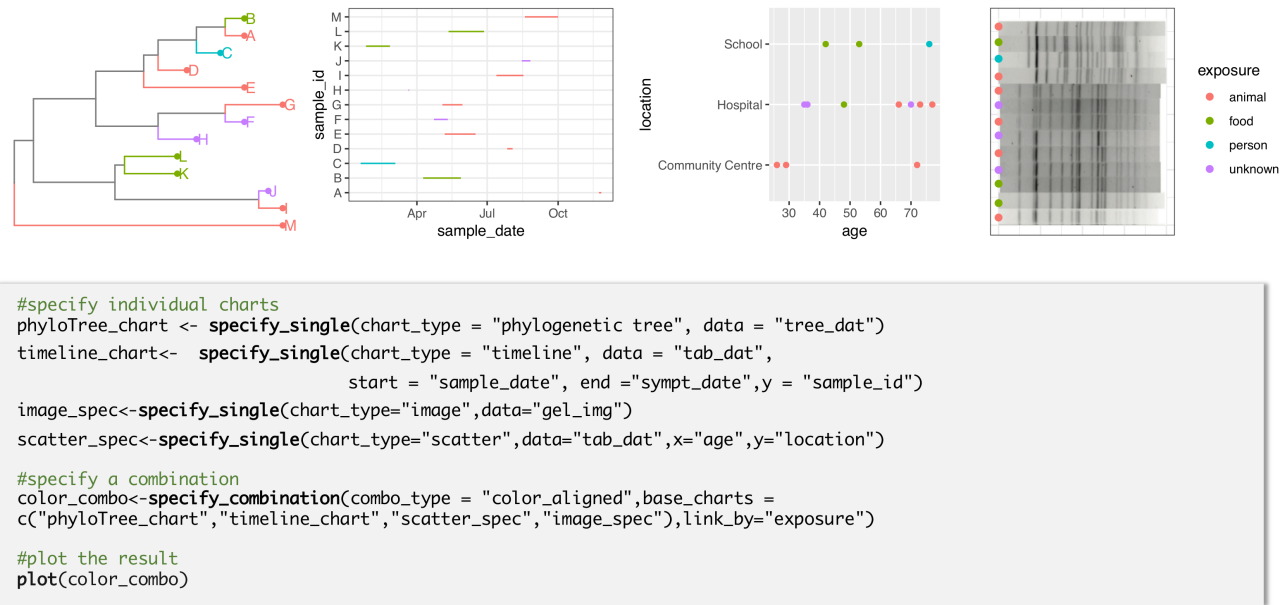


Figure 6.5: Colour aligned combination of disparate static charts. The R-based minCombinR architecture features a declarative framework for gradual binding. The concise code to create this result requires only initial partial specifications from the user for the single charts and the requested combination, with automatically enforced positional and colour consistency within and between charts computed via derived specifications after the `plot` command.

A) Unaligned

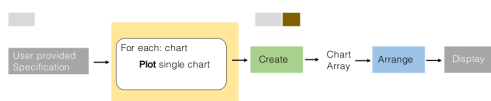
```

phyloTree_chart <- specify_single(chart_type = "phylogenetic tree", data = "tree_dat")
scatter_chart <- specify_single(chart_type = "scatter", data = "tab_dat", x = "sample_id", y = "suscept_status")
timeline_chart <- specify_single(chart_type = "timeline", data = "tab_dat",
  start = "sample_date", end = "sympt_date", y = "sample_id")
unaligned_combo_tree <- specify_combination(combo_type = "unaligned",
  base_charts = c("phyloTree_chart", "scatter_chart", "timeline_chart"))
plot(sm_combo_tree)

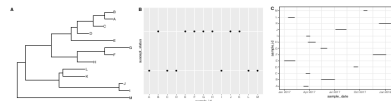
```

Code

Control Flow



Example

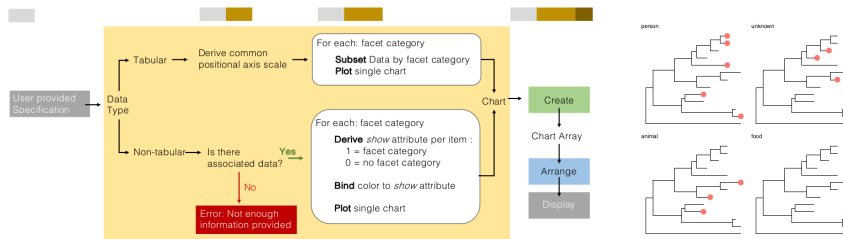


B) Small Multiple

```

sm_combo_tree <- specify_combination(combo_type = "small_multiple", base_charts = "phyloTree_chart", facet_by = "exposure")
plot(sm_combo_tree)

```

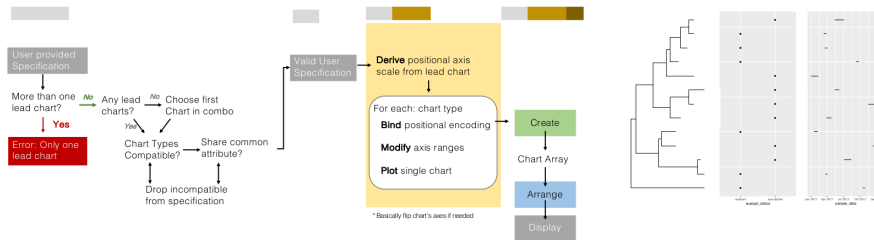


C) Spatially Aligned

```

spatial_combo <- specify_combination(combo_type = "spatial_aligned",
  base_charts = c("phyloTree_chart", "scatter_chart", "timeline_chart"))
plot(sm_combo_tree)

```



D) Color Aligned

```

cl_combo <- specify_combination(combo_type = "color_aligned", base_charts = c("phyloTree_chart", "scatter_chart", "timeline_chart"),
  link_by = "exposure")
plot(cl_combo)

```

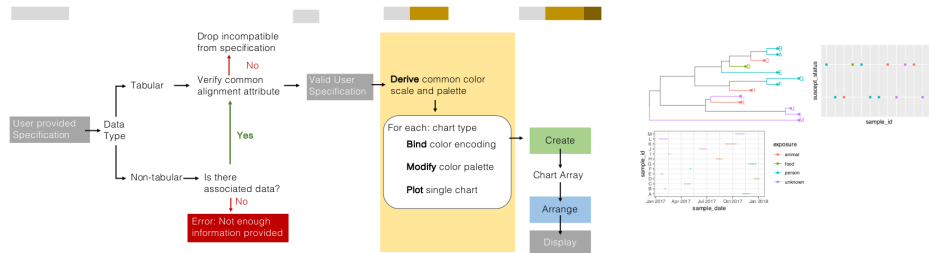


Figure 6.6: Code, control flow, and resulting displays for the four combination types.

6.5.2 Supported Data and Chart Types

minCombinR handles heterogeneous data types as input, either pre-loaded within the R environment or specified as files. In addition to the `data.frame`, a tabular data structure that is the most commonly used data type in R, minCombinR is able to use many other classes of R data types. The current implementation supports tabular, spatial, tree, image, genomic, and network data – for more detailed information please see the online repository. We have created a data object that stores input data and any associated data files together into a common data structure; for example, images may have an associated annotations file with contents that identify features within the image, which can be used to encode data onto the image and facilitate combinations.

From these different data types, minCombinR currently supports the generation of 19 different chart types (Figure 6.4), a subset of the 25 chart types that we observed in GEViT study. This difference is because some chart types did not have good support within existing R charting libraries (for example, certain types of genomic plots) and others that existed could not be reliably converted into a common graphics format. The minCombinR architecture should be able to easily accommodate those charts within the rendering and integration layer once better support exists for them in the future.

6.5.3 Combination Control Flows

In Figure 6.6 we show the code, algorithmic control flows, and displayed results for the four types of combinations, using a synthetic dataset for clarity. We now describe the individual considerations for each of the four combination algorithms pertaining to how the user is guided to generate valid specifications of chart types and combinations, the way that additional specifications are derived to facilitate chart harmonization, and importantly

how different data types are handled by minCombinR. We use a consistent colour scheme in Figure 6.3 through Figure 6.6 to connect our ideas of specifications, architectural layers, and control flows.

Unaligned combinations are the simplest because they do not require any additional derived specifications from minCombinR.

Small multiple combinations can distinguish between scenarios where the scales between facets must be consistent and others where consistency is not required. When scales must be kept consistent between facets, minCombinR will generate the whole chart within each facet and use colour to identify the subsets within the data pertaining to the individual facets (see example in Figure 6.6). We determine which class of small multiples to instantiate based upon the type of input data. From our prior analyses, tabular data is generally used to generate common statistical chart types and generally these chart types do not require a common scale across all facets. Non-tabular charts types such as tree, network, or geographic data tend to contain positional information and it may not be appropriate to show only a subset of the data. For example, it may be pertinent to preserve an entire map and show a region in context rather than only show one specific region within a facet. By default, if minCombinR detects that the input data for the a chart type is non-tabular it will require a consistent scale across the facets, unless the user chooses to override this standard behavior. An additional constraint on non-tabular data is that there must be some associated tabular data file that contains further information about the elements. For example, a spatial file for geographic data must have some associated tabular data that provide details about, say, specific regions within the geographic data and that could be used to facilitate faceting. The user is warned by an error message if no such file exists.

Spatially aligned combinations have a shared horizontal or vertical axis within each chart, and the rendered charts also are arranged in a linear list in the same direction, according to a common attribute field that occurs within

all of the charts. This type of combination is complex because it requires very precise conditions to be correctly generated. In other charting libraries where spatial combinations are facilitated as concatenations, it is up the user to identify whether charts are spatially alignable and exactly how to do so. In contrast, minCombinR shoulders that burden with the automatic derivation of specifications, and extensive guidance to the user in the form of warning messages in case of conflicting or invalid input specifications. The control flow in Figure 6.6 shows the various ways that minCombinR guides the user toward generating valid chart types. First, minCombinR will check whether the user has specified more than one lead chart and if so, will provide alternative suggestions to the user for a new specification with just a single lead chart. If none of the input charts are a lead chart then minCombinR will select the first chart in the combination specification to serve that role. All remaining support charts that share an axis with the lead chart will be retained in the specification, while those that do not will be dropped (the user will be warned of this outcome, but not required to take action). Once a valid input specification has been achieved, minCombinR will automatically rescale and reorient chart along their shared axis in the derived specification.

colour aligned combinations similarly require all chart types in the specification to share a common attribute field, and as with spatially aligned combinations will automatically drop non-compatible charts from the combination specification (again informing the user of this action through warnings but not requiring action). As was previously mentioned in Section 6.4.3, colour aligned combinations are unique in that they modify specifications of singular chart types by adding or overwriting a colour channel encoding.

6.6 Results

We show minCombinR in action on multiple datasets and compare it to existing tools.

6.6.1 Showcasing minCombinR on Different Datasets

Figure 6.4 showcases data from the 2013-2016 Ebola outbreak [34] to demonstrate minCombinR’s ability to generate complex chart types. The exceptions are the network, interior map, gel images which are from different data sources unrelated to the outbreak. Figure 6.5 and Figure 6.6 use a small synthetic public health dataset for clarity. In Appendix D and online repository also include multiple R analysis notebooks with the code required to generate these figures and many others, including many examples of combinations with the Ebola outbreak data, and separate worked examples of the frequently used `mtcars` tabular dataset. These supporting materials provide evidence that minCombinR achieves our ease of use goals by requiring only very few lines of code to accomplish complex combinations across a broad variety of single chart types.

These notebooks also show how minCombinR is able to integrate with analysis in R to enable allows users to apply a myriad of techniques to their data and then to rapidly visualize different chart types and their combinations. They also demonstrate the various helper functions that we have implemented to assist users in working with their data and preparing it for visualization, including an RStudio Shiny application that helps users annotate images and store this data automatically in order to facilitate combinations of images with other chart types.

6.6.2 Comparison to Existing Tools

We have also conducted an analysis of minCombinR capabilities compared to the three widely stand-alone tools Excel, Tableau, and PowerBI; the gg-universe and the base R libraries; the JavaScript systems D3 and Vega/Vega-Lite, the Python libraries Bokeh/holoviews/geoviews, matplotlib, seaborn, and the two domain-specific tools Microreact and NextStrain (6.7). For each of the chart types and combinations that are supported by minCombinR

we assessed whether these were easy to generate, possible but challenging, had any types unsupported, or all types were impossible to generate. We define a type as *easy* (dark green) when it is natively implemented in the stand-alone tools. For charting libraries, users can make almost anything with sufficient effort; we defined the easy case as chart/combination code being present on the official gallery pages of those libraries since even users with limited technical skills could copy and modify that code with little effort. The *possible but challenging* case (light green) is our determination that it is technically possible to generate some chart type or combination if a user has sufficient technical knowledge to do so, but these solutions have lower ease of use. For stand-alone packages, plugins or extension architectures imply that it is possible to have support even if it is not explicit. For charting libraries, we consider anything that is not part of the official gallery to be possible but challenging. The *1 + not possible* case is a documented instances where some chart type or combination is actually not possible to generate (orange), which we have determined by explicit requests for features or official documentation confirming that the type is not supported. Finally, the *None are possible* case indicates that none of the chart types in that category are supported or the combination is impossible in that system.

Our results show that R and JavaScript charting libraries provide the most extensive explicit support; by design minCombinR provides much more explicit support for chart types and combinations relative to other charting libraries and stand-alone tools. The ease of use of minCombinR manifests in the lines of code needed to write in order to achieve a desired result. For individual chart types, the advantages of minCombinR over existing charting libraries is marginal. However, for combinations, especially across multiple different types of complex charts, minCombinR's contributions are substantial. For spatially and colour aligned combinations in particular, minCombinR greatly reduces the burden to the user by minimizing the total amount of code needed to be written; with five lines a user can generate a spatial or colour alignment between two chart types, whereas it would take

considerably greater effort to achieve the same result with existing charting libraries. For the stand-alone and bespoke tools we note that many chart types and combinations are not possible to create. Bespoke tools, unsurprisingly, are the most restrictive, support a much more limited set of tasks compared to other stand-alone systems. Excel, a very popular tool in public health, has improved its visualization capacity in recent years but it still has many limitations including an emphasis on single charts. Tableau and PowerBI offer better support through both their defaults and extensions, but currently have limitations for domain-specific idioms and data. The limitations of the stand-alone and bespoke tools highlight some of the benefits of the bottom-up approach to developing minCombinR's architecture. By having an awareness of what a broad set of users in a particular domain currently want to do, we could design an architecture with sufficient flexibility that it can serve their current needs and has the potential to evolve over time to adapt to their changing needs.

		R Based			JavaScript Based		Python Based			Stand Alone			Public Health Stand Alone	
Chart Class	Chart Type	minCombinR	gg universe	Other R libraries	D3	Vega/ Vega-lite	Bokeh + holoviews + geoviews	matplotlib	Seaborn	Excel	Tableau	Power BI	Microreact	Next Strain
Single Chart	Common Statistical						not: swarm	not: swarm	not: pie				none	none
	Color									not: catStripe			none	none
	Relational								none				none	none
	Temporal								none	not: stream			only: timeline	only: timeline
	Spatial						not: intMap	not: intMap	none	not: intMap	not: intMap	not: intMap	only: geogMap	only: geogMap
	Tree						none	none	none	not: phyloTree	none	none	only: phyloTree	only: phyloTree
	Genomic					not: alignment	none	none	none	none	none	none	none	none
	Other								none	not: image	not: image	not: image	none	none
Combinations	Unaligned													
	Color Aligned													
	Spatially Aligned													
	Small Multiple													

Legend	Easy	All are possible, but challenging	1+ not possible	None are possible
--------	------	-----------------------------------	-----------------	-------------------

Figure 6.7: Related work comparison. We summarize here a detailed analysis of minCombinR's capabilities compared to R, JavaScript, Python, general stand-alone, and bespoke stand-alone approaches. Appendix D contains the the full details of this analysis including links to each example or counterexample.

6.7 Discussion and Future Work

Scientific data is becoming much more complex and data visualization systems need to adaptively respond to these changes in order to keep up with the demand for better analysis tools. Many past visualization systems have fallen into one of two primary camps: general systems developed with a top-down approach where domain specificity is intended to be handled purely by the programmers who use them, and specific bespoke visualization systems that are highly tailored for domain-specific requirements but require intensive labor to develop and maintain. We have instead explored a middle ground between these by taking a bottom-up approach to build the minCombinR system from an existing domain-specific visualization typology. The typology allowed us to identify the complex varieties of data and chart types that domain experts generate, and also the types of combinations that they were using to communicate the linkages between their heterogeneous data. This bottom-up approach led us to propose a flexible architecture based on a set of general requirements, which could be used or extended for other domain-specific instantiations.

Although we have succeeded in automating chart harmonization beyond previous work, interesting problems pertaining to specification of chart combinations remain that would benefit from further investigation. GEViT also specified an overlay spatial alignment, where two chart types are placed on top of each other, to classify figures such as a network placed on top of a geographic map to show the spread of disease. Overlay spatial alignments are clearly possible to support in that specific case, but much more complex to support generally. This combination examples highlights an important challenge of how the linkage between data and the individual pixels on the screen that make up visual idioms. The D3, Vega, and the ggplot charting libraries make use of domains and scales to demarcate data and pixel values, respectively, and much of minCombinR's automation is fundamentally about automatically transforming scales of two or more chart types based upon the

data (domain) and specification. Overlay combinations and even more complex types of spatial alignments involve much more co-ordination between the domains and scales of many different and complex chart types, which is difficult to support automatically in the general case even as many solutions exist for limited combinations of specific chart types. Individual users that are more technically savvy may also find it complex to programmatically coordinate more complex combinations of charts. We are continuing to explore how we might incorporate such a combination into minCombinR.

6.8 Conclusion

We have presented minCombinR, an R-based toolkit whose architecture is informed by a domain-specific typology. It supports coordinated combinations of a broad array of static chart types, including not only common statistical charts but also maps, trees, and genomic charts. Existing systems provide many options, but address different points than minCombinR in the trade-off space between ease of use, expressivity, and the extent to which they support analysis. In particular, no previous system supports easy generation of the full range for chart types we support or the full range of static chart coordination combinations that we support. In minCombinR we sought to architect a system that minimizes the amount of code a user must write, creating an extensible and expressive support for a broader design space that includes chart combinations as first class citizens, and that integrates readily with analysis tools.

Chapter 7

GEViTRec:

Domain-Aware Visualization Recommendation for Data Reconnaissance and Harmonization

A picture is not thought out and settled beforehand. While it is being done it changes as one's thoughts change. And when it is finished, it still goes on changing according to the state of mind of whomever is looking at it.
— Pablo Picasso

¹ Data visualization recommender systems primarily support tabular datasets through generating common statistical charts, but many more datatypes are required for current data analysis practices. These heterogeneous and multi-dimensional datasets introduce challenges that are beyond the capacity of existing recommender systems and are also overwhelming to domain experts. We first present a novel conceptual framework for data reconnaissance and task wrangling with the four phases of acquire, view, assess, and pursue to characterize what domain experts do when attempting to make sense of complex unknown data landscapes with the objective of identifying data that matches their analysis goals. This conceptual framework motivates our development of a novel general algorithm for domain-aware visualization recommendations that automatically harmonizes attribute fields between heterogeneous data sources to create an entity graph. That graph is traversed

¹This chapter has been submitted for publication [28]:
A. Crisan, J.L. Gardy, and T. Munzner (2019). GEViTRec: Domain-Aware Visualization Recommendation for Data Reconnaissance and Harmonization.

to find high-ranked paths according to connection strength between data sources, the diversity of visual encodings, and their relevance. The relevance score used to rank and select visual encodings for multiple datatypes is based on an existing quantitative analysis of a design space reflecting current common practice by experts in a specific domain. We also present the implementation of GEViTRec, an instance of our algorithm for the specific domain of genomic epidemiology, and demonstrate its results using a real-world public health outbreak dataset. Our framework and algorithm are an important step in extending the capacity of visualization recommender systems to better support domain experts as they endeavor to understand and analyze their growing and complex collections of data.

7.1 Introduction

Automatically recommending suitable visual encodings for data has been a longstanding area of visualization research dating back to the foundational APT system [68], with significant recent activity including ShowMe [69], Compass and Voyager [129], and Draco [76]. This previous work has been focused on tabular data, but current data analyses require support for a much broader set of datatypes including spatial, network, and genomic data. Stakeholders who need to integrate and analyze heterogeneous data are becoming increasingly overwhelmed by the complexity of their data, in addition to its volume. Moreover, previous recommendation systems have been architected around the assumption that stakeholders are ready for a deep dive into an existing specific dataset to conduct exploratory searches. This assumption that the data a stakeholder needs to visualize is clearly demarcated and immediately available as input to the system does not hold in many situations. Stakeholders may be faced with an unfamiliar **data landscape**: a large space of datasets that are either available to them now, or that they could gather, or that they could request from some gatekeeper who controls access. In this case, they need a system for rapid assessment to

establish the basics of what a dataset contains, and whether it is sufficient for some intended task or if more data might be needed to for their analysis goals. We call this assessment **data reconnaissance**.

We have found that stakeholders who are sufficiently unfamiliar with a data landscape typically cannot simply articulate their analysis needs. They may have many possible questions, but determining which of them have a reasonable chance of being answered depends on what data is available to them. The difficulty of understanding stakeholder needs is well known in the visualization research community, which has long advocated human-centered design methods for task elicitation. One popular methodology for untangling tasks is the design study approach [103], where bespoke visualization tools are developed for a specific domain problem through iterative rounds of task elicitation and prototype development. The two major limitations of this approach are the need to invest many months of time into the process, and the dependence on a very specific data configuration that weighs heavily in the design of the eventual solution. It is thus a poor match for assessing unfamiliar data landscapes during data reconnaissance. More generally, we assert that tools designed to support the deep investigation of a particular dataset do not necessarily support stakeholders who are attempting to rapidly glean a high-level understanding of an unknown data landscape.

We observed these unmet needs first hand while working with stakeholders in public health that were faced with the challenges of integrating different forms of heterogeneous datasets arising from genomic epidemiological investigations. Although our stakeholders were enthusiastic to use genomic data in conjunction with with other data sources, many of them reported that they were unsure about these new datasets and how incorporating them into analyses would affect their approach [25]. Although we first became aware of these challenges within the context of a specific domain, they are not unique to genomic epidemiology.

To address the challenges arising from data reconnaissance, we have devel-

oped a general data visualization recommender algorithm for heterogeneous datatypes that automatically harmonizes datasets, finds linkages between datasets, and shows those linkages through combinations of multiple coordinated visual encodings. This algorithm allows stakeholders to examine automatically generated visual encodings, without the need to provide initial partial specifications, so that they can rapidly explore a complex data landscape, gaining enough familiarity with the data that they can conjecture about possible tasks or decide what new sources of data to seek. Beyond support for heterogeneous datatypes, we also propose the new idea of a domain-aware recommendation system, that incorporates a domain-specific quantification of relevance that is used to rank visual encodings. The purpose of introducing relevance ranking is to address an existing limitation of previous recommenders that have focused on ranking visual encodings with respect to perceptual efficiency. Although ranking by perceptual efficacy works well for tabular data encoded with common statistical charts, where there has been substantial empirical study of human perceptual response, there is far less empirical evidence available for ranking visual encodings of non-tabular datatypes.

Our work presents contributions toward better characterizing the challenges that stakeholders face when analyzing unfamiliar heterogeneous data and the design of visualization systems that may help them. **Our first contribution is a framework for data reconnaissance and task wrangling, generalized to a domain-agnostic context.** Giving a concrete name and framing to the difficulties that arise in the visualization of heterogeneous multidimensional data landscapes is also intended to motivate further discussion toward how visualization methods can evolve to help stakeholders address their complex domain-specific challenges.

Our second contribution is the design of a general algorithm for domain-aware visualization recommendation that goes beyond tabular data and singular charts.

- We define and describe the properties of a *domain prevalence* visualization design space that may be used to computationally generate relevant recommendations of singular and combined charts
- We introduce the notion of data harmonization and present methods for the integration and analysis of heterogeneous and multidimensional collections of data including genomic, network, temporal, and spatial datatypes.
- We present methods for the automatic generation of visualization specifications without any initial partial specifications from the user, given only input data

Our final contribution is GEViTRec, an instance of a domain-aware visualization recommendation system targeted at the domain of genomic epidemiology. We compare the design and implementation of GEViTRec to existing visualization recommenders and discuss how such systems need to continue to evolve beyond tabular data and singular charts to support complex heterogeneous and multidimensional data.

7.2 Background

We briefly describe our prior research findings that are necessary for understanding the design decisions of our algorithm.

We have been collaborating closely with stakeholders in public health to help them analyze and visualize emerging genomic data for use in their epidemiological investigations and disease control policy making. Our stakeholders each had different expertise and consequently were familiar with different datatypes depending upon their role. Compounding the problem were challenges with data access and availability [23] that obscured the data landscape. It quickly became clear that stakeholders needed a general view of their data in order to begin to discuss what it might be used for and, importantly, for them to assess what data they still need to obtain access

to. To overcome these challenges, we decided to explore the visualization strategies of the broader genomic epidemiology research community so that we could begin to anticipate the different types of data that may exist and how they could be visualized.

In a prior study we characterized and enumerated the domain-specific data visualization strategies used by these stakeholders [24]. We created a Genomic Epidemiology Visualization Typology (GEViT) that broke down how visualizations were constructed through chart types, enhancements, and combinations. Our typology revealed 25 unique chart types within 8 categories (common statistical charts, colour, relational, temporal, spatial, tree, genomic, and other), four types of combinations (spatially aligned, colour aligned, small multiples, and unaligned), and two primary mechanisms of enhancements (adding or re-encoding marks). GEViT was developed using a corpus of approximately 18,000 research articles pertaining to genomic epidemiology that was representatively sampled to yield a set of 800 figures that informed the typology generation. We also applied text mining techniques to the titles and abstracts of all articles to derive a sense of the creation context for the sampled set of figures. Most importantly, our representative sampling strategy allowed us to enumerate these different visualization strategies and obtain a quantitative sense of their relevance and importance.

7.3 Data Reconnaissance and Task Wrangling

We provide a general framing of data reconnaissance and task wrangling as an abstraction of the processes undertaken by stakeholders attempting to explore a heterogeneous and multidimensional data landscape. This conceptual model is the underlying motivation for the formalisms and design decisions that we present in the subsequent sections.

7.3.1 Operational Definitions

Exploration is a general term that is broadly applied and consequently, captures many different complex processes. Here we distinguish **data reconnaissance** as the process of exploring an unfamiliar **data landscape**; that is, the very large space of datasets that are available but not yet understood. It includes datasets that already exist but have not been assessed, or that do not yet exist but could be gathered, or that do exist but have barriers to access. Data reconnaissance differs from *data wrangling* and *investigative exploration*, both of which require as a precondition, an existing dataset that is transformed and analyzed in depth to generate new insights.

Task wrangling is the process of progressively forming a crisper notion of both what tasks a stakeholder needs to address and whether available data is suitable for them. We follow the design study methodology (DSM) [103] definition of a clarity axis with crispness in contrast to fuzziness on its two ends, where a crisp task has a clearly defined goal with a known set of steps. The DSM also posits that task crispness should evolve over time, but assumes a clearly demarcated dataset. Task wrangling describes situations where the data itself is also fuzzy.

Data reconnaissance and task wrangling are related but distinct. A better understanding of the data landscape can help to improve the clarity of tasks, and clearer tasks provide further information about which areas of the data landscape to next pursue. We posit a chronological ordering where data reconnaissance and task wrangling come before investigative exploration and data wrangling. The objectives of data reconnaissance and task wrangling are to identify relevant data and relevant visual analysis tasks. Its findings could feed subsequent exploration and data wrangling phases. Although we reify and name these processes explicitly for the first time, there is clear evidence that previous visualization researchers have indeed faced these challenges. For example, Ghani *et al.* [44] state that the challenges of forming crisper tasks for their study are exacerbated because “multimodal social networks

are not a well established concept even in social science”. Wood *et al.* [130] discuss the complex trajectories of different domains that do not follow a “a linear progression” toward crisper tasks, but involve a more dynamic relationship between data, tasks, and understanding.

7.3.2 Conceptual Framework

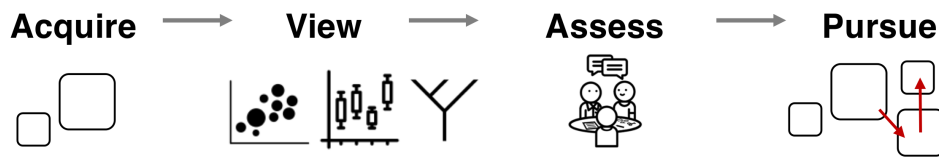


Figure 7.1: Conceptual framework for data reconnaissance and task wrangling.

The four phases of acquire, view, assess, pursue are repeated across multiple cycles of a data reconnaissance and task wrangling process. Squares represent individual data sources, and the red arrows indicate stakeholders obtaining new data informed by their assessment of existing data.

Our conceptual framework is composed of four phases: acquire, view, assess, pursue (Figure 7.4). Stakeholders **acquire** some initial data in the form of one or more heterogeneous datasets. They **view** these data to gain a sense of what these datasets are, how they may be related, and a high-level overview of what they show. Stakeholders can then **assess** these data by using the visualization results to consider whether these data meet any of their needs, whether more data may be collected, and what tasks these data could be used for. Stakeholders can then opt to **pursue** additional data sources. A stakeholder may begin a data reconnaissance and task wrangling process without a clear understanding of what data should be visualized or for what purpose. We refer to this initial phase as the **fog of war** (Figure 7.2). Through multiple cycles of the acquire, view, assess, and pursue phases, stakeholders undertake **informed data ideation** where purposefully acquiring new data provides an ever-better understanding of the data landscape. When the connections between and the utility of the different data sources is sufficient, stakeholders can conclude with a demarcated

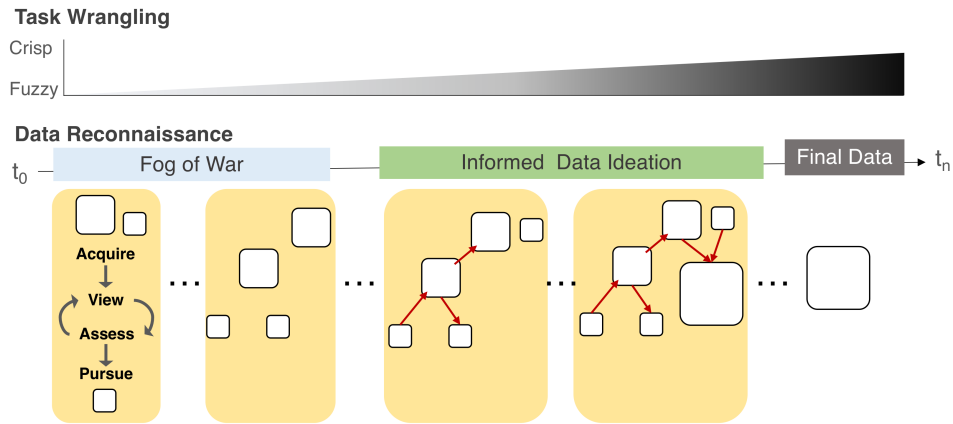


Figure 7.2: Data reconnaissance and task wrangling phases over time. Within a single cycle (yellow square) a stakeholder will perform the acquire, view, and assess phases in order to familiarize themselves with their data and determine what to further pursue. Over a period of time ($t_0...t_n$) a stakeholder will gather more datasets for analysis and form crisper ideas of the tasks these data could be used for. We mark the passage of time and growing familiarity of the data into broader phases that data reconnaissance and task wrangling process occur within: the fog of war, informed data ideation, and the demarcation of the final data.

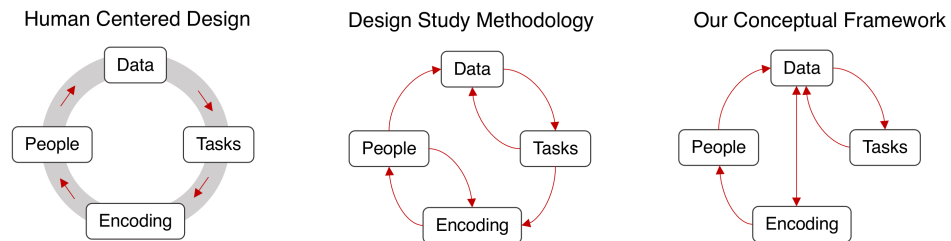


Figure 7.3: Comparing different approaches to human centered design. (a) The steps of a human centered design process [66] (b) A widely used design study methodology [103] (c) Our conceptual framework.

dataset that can be used for some specific analysis goal. This process may even be followed by a design study if a clear need for a bespoke solution emerges.

For an example of such a process in action, consider the following scenario. *When investigating a disease outbreak, epidemiologists may initially have*

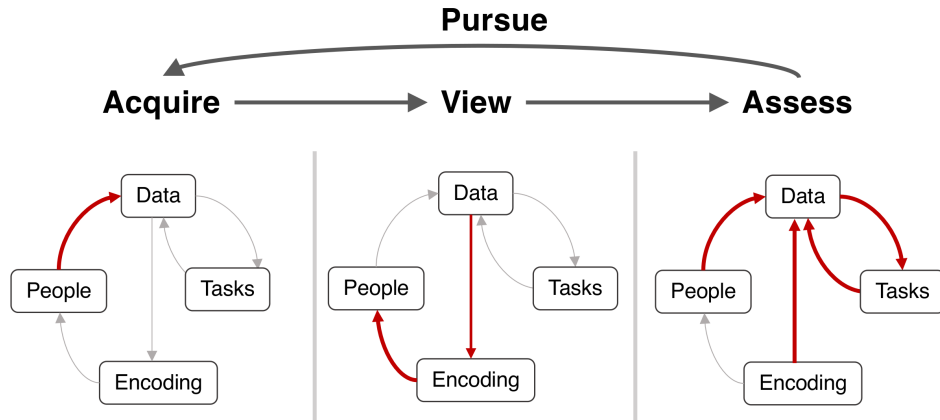


Figure 7.4: Data reconnaissance and task wrangling phase breakdown. We indicate the acquire, view, assess, and pursue phases of our conceptual framework on the human centered design loop.

some tabular data for sick individuals they have already identified, but wish to obtain additional data sources that could help their investigation. They consult with researchers to obtain new phylogenetic and geographic data that could add context to their investigations. These epidemiologists are unfamiliar with these new data and use a data visualization tool to get a quick sense of what these data are and how they may be connected together. They can see some linkages between leaf nodes in the phylogenetic tree within some geographic region. They discuss their findings with others and ideate further upon datasets that may exist to provide additional information on their particular finding or the outbreak dynamics more generally. They can use the results of their current assessment to justify the collection of more data. They continue this process until they arrive at a finalized, more complete dataset that they now investigate deeply to generate a more concrete understanding of the outbreak origin, spread, and dynamics.

We demonstrate the differences between data reconnaissance and task wrangling and two common approaches described in prior work, for human-centered design generally [66] and design study methodology specifically [103] (Figure 7.3). Broadly, all three of these approaches involve

people, data, tasks, and visual encodings. The prior approaches require that people provide some context of creation, data, and some initial set of tasks that are further refined over time and that directly inform the design of visual encodings. Our conceptual framework permits a loosely defined creation context and uses the data itself to inform the design of visual encodings in advance of any contextual task information. We employ visual encodings to *identify relevant tasks* and *motivate the pursuit and acquisition of new data*. We overlay the phases of our conceptual framework on the human-centered design loop (Figure 7.4) to indicate the goals of each phase.

7.4 Formalisms for Visualization Recommendation

To support data reconnaissance and task wrangling we propose an algorithm that is able to automatically display relevant visual encodings to domain experts and adaptively respond to new heterogeneous datasets. The algorithm that we propose uses a pre-existing domain prevalence visualization design space informed by expert definitions as a kind of prior that maps data sources to potential visual encodings. When stakeholders provide their own heterogeneous data to our algorithm, it uses that prior knowledge from the design space to select and prioritize visual encodings for display. In addition to selecting relevant visual encodings, our algorithm also identifies linkages between stakeholder’s data sources, a process that we refer to as **data harmonization**, and attempts to generate coordinated static combinations of visual encodings so that several aspects of heterogeneous and multidimensional data can be presented side by side.

We now present formalisms that are used in our algorithmic design. We present our definition of the properties of a visualization design space, a data model that defines our assumptions around heterogeneous and multi-dimensional data, and a visualization specification including what it means to automatically generate such a specification without input from the stakeholder.

7.4.1 Domain Prevalence Design Spaces

To inject domain awareness into a visualization recommendation algorithm, we create and analyze a visualization design space. The term *design space* is used broadly within the visualization research literature with many shades of meaning. Here, we define a **domain design space** as a collection of visual encodings (V) that are produced by experts in some domain. We define a **domain prevalence design space** (D) as one that both captures the full scope of visual encodings used by some definable set of experts (within reason) and includes an estimate for prevalence of these visual encoding strategies in the domain. We use this quantitative prevalence information in our algorithm to define relevance scores for different visual encodings. We describe an instance of a quantitative prevalence visualization design space in Section 7.2; the method for developing it was presented in prior work [24]. Our definition of a domain prevalence design space, which we will frequently abbreviate below as simply a **design space**, has two important properties that our algorithm exploits:

The first property is that the design space is generated from a corpus of documents that link to some analysis. A collection of images from internet sources without additional contextual information would *not* have this property because there are no guarantees that we could assess the datatypes used in the analysis. Our design space is constructed from a corpus of research articles precisely because we could establish a link between data and visual encodings.

The second property is that the design space is quantitatively representative of visualization strategies that are used by domain experts. Both SetVis [1] and TreeVis [101] are examples of visualization design spaces that do *not* have this quantitative property. They present visualization design strategies for a specific types of encodings and only single instances of these strategies.

While the visualization strategies of domain experts may not conform to all of the guidelines and best practices currently articulated by the visualization

research community, a collection of domain expert visualizations reflects how they attempt to present their own data and incorporates their knowledge of the phenomena they investigate. We leverage this connection between data and visual encodings within the design space to support the generation of domain-aware, and thus relevant, visual encoding recommendations.

7.4.2 Data Model

We define two broad categories of heterogeneous datatypes: tabular and non-tabular. The non-tabular data category broadly captures many different datatypes; for our specific application context in genomic epidemiology, non-tabular datatypes include spatial, tree, network, image, genomic, and temporal. Our algorithm takes into consideration both the data category (tabular vs. non-tabular) and specific datatype when recommending different visualization encodings.

We describe a process for **data harmonization** that integrates and links different data sources for visualization. Data harmonization relies on breaking heterogeneous data sources into atomized attribute fields for analysis, a step that we call *exploding the data*. Attribute fields can be further classified as numeric or non-numeric. For tabular data, fields are simply each of the individual columns. For non-tabular data, we assume that a data source has at least one field that corresponds to a unique identifier, and that there may be associated tabular data that contains additional attribute fields. *For example, a data source for a tree datatype is a flat file that contains the tree structure and the ids of its leaf nodes. This data source may also have an associated tabular data file where one column contains the leaf node ids in addition to other attribute fields.*

As we will describe in detail in Section 7.5, both the datatype and attribute fields of individual sources can be used to create a visual encoding. We can also reliably find linkages between different data sources by conducting a simple set-similarity analysis of categories between non-numeric fields

of two data sources. The different data sources, their attribute fields, and the derived linkages between the attribute fields of different sources can be collectively modeled as an entity graph (Figure 7.5). Nodes in the graph represent individual data sources and their associated fields. Edges also connect data sources and their attribute fields and also represent the linkages between fields (and thus by extension link between data sources). Importantly, modeling data as a graph allows us to identify linkages between two datasets that exist only due to a third intermediary. We analyze the entity graph to automatically generate specifications for visual encodings.

7.4.3 Visualization Specification

A visualization specification can describe either a single visualization encoding (V) or a specific combination of encodings (C_V)

For a single visual encoding, we follow the same formalisms for a visualization specification as others have [6] [68] [76], but with some minor modifications. Briefly, prior work specifies some chart type with a set of visual encoding marks and channels that map to some attribute field. *For example, to create a scatter chart, numeric attribute fields from tabular data can map to positional encoding channels for point marks.* Some specifications can also impose constraints, *for example, attribute fields encoding colour channels must be non-numeric and have fewer than 12 unique categories.* To support heterogeneous data, we also include a datatype attribute for each visualization specification, *for example, a phylogenetic tree visual encoding requires a specific tree datatype.*

The specification for a combination of visualization encodings includes the names of individual chart types to combine, the type of combination, and if applicable the field on which to link the charts. We have previously reported on specifications for combinations [27], and briefly summarize our prior work here. We support four types of combinations: small multiples, spatially aligned, colour aligned, and unaligned. The first three combination types

produce coordinated static combinations of multiple chart types, while the latter produces an uncoordinated combination.

7.5 General Algorithm

We describe a general algorithm that given some domain-specific visualization design space (D) and heterogeneous collection of data (H) as input will attempt to automatically recommend visual encodings (V) and, if applicable, combinations of encodings (C_v) that are relevant to a stakeholder.

We first describe how we analyze a domain prevalence design space (with the properties described in Section 7.4.1) to map between different datatypes and visual encodings, and how we obtain rankings of different visual encodings that are used to constrain the possible set of relevant visualizations presented to the user. We next describe the process of data harmonization by “exploding” data fields out of data sources to generate an entity graph. Finally, we demonstrate how we analyze and rank paths within the entity graph to automatically create specifications of visual encodings and their combinations that are then presented to stakeholders for viewing and assessment.

7.5.1 Mapping From Datatypes to Visual Encodings with a Design Space

Our algorithm assumes there exists a design space that corresponds to data within an expert’s domain, whether created with the specific method that we proposed in prior work [24] or some other approach that guarantees the requirements are met. We analyze this design space to identify correspondences between heterogeneous datatypes and visual encodings and use this analysis result in our algorithm to identify a set of candidate visual encodings for each data source.

Relevance-Ranking Visual Encodings

We can generate a relevance score for visual encodings by taking into account their prevalence in a design space and also additional information about the creation context of each encoding. Here, we use the year that the encoding was created and place a higher weight on those visual encodings created more recently. Thus, we can calculate the relevance score for all unique chart types within a design space (V_D) by the total weighted sum of occurrences for each unique chart type (V_i) by year (Y), where N is the total number of visual encodings corresponding to a specific year (Y). Formally,

$$Rel(V_i) = \sum_y \sum_{n=1}^N V_{i,n} * w_y \quad (7.1)$$

where i is the total number of unique chart types in D and w_y is a weighting factor that penalizes older visual encodings. The raw relevance score (Rel) is then rescaled for computability, interpretability, and consistency across different design spaces:

$$rescale(V_i) = \frac{V_i}{\max(Rel(D_V))} * 10 \quad (7.2)$$

where V_i is a single visual encoding and D_V is the visual encoding with the maximum relevance score in the design space D . This scaled relevance score ranges between 1 (least common) and 10 (most common). It is designed to produce non-linear results to emphasize the relative importance of different types of visual encodings. *For example, a phylogenetic tree is the most used visual encoding in genomic epidemiology and the next most common visual encoding is bar chart. Instead of giving a phylogenetic tree a rescaled relevance score of 10 (maximum) and bar chart a score of 9, our rescaled rank score produces scores of 10 and 4 respectively. These scores emphasizes the relative importance of a phylogenetic tree compared to all other encodings.*

7.5.2 Data Harmonization and Entity Graph Generation

Data harmonization is the process of integrating heterogeneous datatypes and identifying linkages through their multidimensional field attributes. To simplify computational procedures across different heterogeneous datatypes we have developed a common data structure (CDS). For each input dataset, our data structure stores a unique identifier, the source of the data on disk, the type of data, and finally the data itself. We also store non-tabular data and its associated data within the same CDS. We compute over the complete set of CDS in order to “explode” attribute fields from individual data sources (Figure 7.5). Exploded fields are categorized as numeric or non-numeric, and for non-numeric fields we further quantify and store the number of unique categories. We generate an internal metadata object that keeps track of all the different data sources, along with their datatypes, attribute fields, and field classifications.

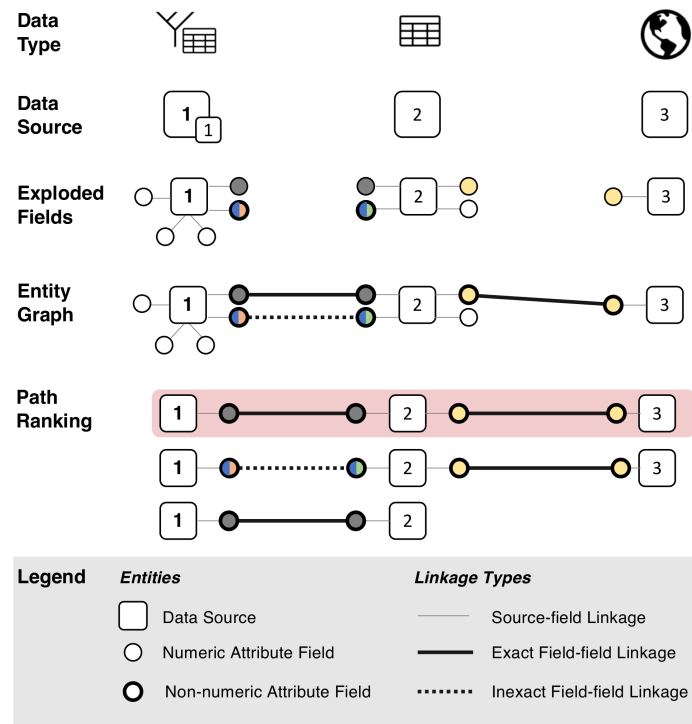


Figure 7.5: Data harmonization and entity graph generation schematic algorithm overview Data source #1 is a tree datatype with associated tabular data, data source #2 is a tabular datatype, and the final data source #3 is a spatial datatype. The attribute fields within these data sources are “exploded” and classified into numeric and non-numeric field types. For data source #1, attribute fields are exploded from the tree data and associated data, for data source #2 the attribute fields are the tabular data column, and for data source #3, attribute fields are ids associated with spatial polygons. We use the Jaccard index to compute the similarity of categories between two pairs of non-numeric attribute fields and establish exact and in-exact linkages between data sources through their attribute fields. The data sources, their attribute fields, and the linkages between their fields are used to generate an entity graph. We enumerate paths of the entity graph that link pairs of data sources and rank paths according to their link strength, diversity, and total relevance. We generate specification for visual encodings from each of these paths, beginning with the most highly ranked path (highlighted in red).

Once fields are “exploded” from their data sources, we consider the unique categories that occur within each non-numeric field, and compute the set

similarity for the categories in common between a pair of such fields. For all possible pairs of non-numeric fields from different data sources, we calculate the set similarity between fields A and B using the Jaccard Index, a normalized score between 0 and 1:

$$J(A,B) = \frac{A \cap B}{A \cup B} \quad (7.3)$$

A Jaccard index of 1 indicates an exact match between all unique categories in the two fields, 0 indicates no matching categories, and an inexact match between 0 and 1 indicates some but not all category values in common. When pairs of fields between different data sources have a Jaccard index greater than 0 there is some linkage between these data sources. We can use this linkage information to derive an entity graph of data sources and their associated attribute fields (Figure 7.5). As we show in Section 7.6, when generating the entity graph it is also possible to specify the minimal Jaccard index value required to designate a linkage between two non-numeric fields.

Our algorithm analyzes paths within the entity graph in order to come up with specifications for visual encodings and, if applicable, combinations. The path analysis process involves two steps. First all paths linking all possible paired combinations of data sources are enumerated and ranked. We describe this ranking process in the subsequent Section 7.5.3. Second, within each individual path we identify highly connected nodes (fields that link two or more datasets) and use these nodes to seed specifications; we describe this process in Section Section 7.5.4.

7.5.3 Ranking Paths Within the Entity Graph

We enumerate paths between pairs of data sources and rank the resulting paths according to strength of the connections between data sources (**link strength**), the diversity of visual encodings that can be generated from the

different data sources (**diversity**), and finally the cumulative relevance of those visual encodings (**total relevance**). Depending upon the degree of connectivity in the entity graph, paths may not exist between some pairs of data sources; in the worst case scenario, none of the data sources will be connected and there will be no paths to rank. Where there exist two or more disconnected components with the entity graph, paths are ranked, analyzed, and used to recommend visual encodings for each individual component.

The **link strength** of a path is the normalized sum of edge weights (e_i):

$$strength = \frac{\sum_{i=1}^N e_i}{N} \quad (7.4)$$

where N is the total number of edges on that path, e_i is the Jaccard index, and where $i \in \{1 \dots N\}$. The total normalized link strength value varies from 0 up to 1, where a strength of 1 means that a path is composed entirely of exact matches (edge weights of 1).

Diversity measures the variability of the visual encodings that can be produced from the data sources along the path. In Section 7.5.4 we present a mapping of different data sources to specific visual encodings. We count the number of unique instances of a visual encoding. Diversity scores range from 1 (the minimum relevance score) for a single visual encoding from a single data source, up to a maximum value that depends on the input data.

Finally, we compute the **total relevance** of a path by summing the relevance scores of the unique visual encodings that can be produced from the different data sources along the path. For tabular data, we take the relevance score of the highest-ranking visual encoding that could be produced. The relevance score ranges from 1 for a single data source, up to $10 \times T$, where T is the total number of data sources along the path.

The link strength, diversity, and total relevance score for each path is computed. Higher ranks are given to paths with link strengths closer to the

maximum value of one, with greater diversity, and that could produce highly relevant chart types. To produce a final quantitative summary rank score, we rank all paths from 1 (highest) to P (lowest, where P is the total number of paths) for each of our three criteria (link strength, diversity, and relevance) and then sum the results. The final rank score can range from 3 (highest link strength, diversity, and relevance) to $P \times 3$ (lowest). Paths are subsequently analyzed in order of rank.

7.5.4 Generating Specifications

A) Scatter Chart Template

```
chart_spec(chart_type = "scatter",
  data = data_obj(NA, "table"),
  x=var_obj(NA, "quant|qual", dataSource=NA, TRUE),
  y=var_obj(NA, "quant|qual", dataSource=NA, TRUE),
  color = var_obj(NA, "qual-12", dataSource=NA, FALSE),
  shape = var_obj(NA, "qual-6", dataSource=NA, FALSE))
```

B) Phylogenetic Tree Chart Template

```
chart_spec(chart_type = "phylogenetic tree",
  data = data_obj(NA, "phyloTree"),
  metadata = data_obj(NA, "table"),
  color = var_obj(NA, "qual-12", dataSource=NA, FALSE),
  shape = var_obj(NA, "qual-6", dataSource=NA, FALSE))
```

Figure 7.6: Internal Templates. The initial partial specifications in the internal templates, with empty slots denoted by *NA*, are mapped to specific data and fields during the visualization recommendation process to produce a full specification. Two specific examples used within GEViTRec: (a) Scatter chart template (b) Phylogenetic tree template.

Beginning with the highest ranked path, we produce visualization specifications according to a possible set of visual encodings as established from the data sources. Paths are analyzed for highly connected nodes that link two or more datasets and the fields pertaining to those nodes are used to seed an initial specification for all possible visual encodings. We now describe how attribute fields from different data sources are assigned to visual encodings

within a specification.

For each chart type, we have created an internal visual encoding template with initial set of partial specifications that is transformed into a full specification at run-time by our algorithm. Each template has encoding slots that are initially empty, specifies constraints on the properties of the field type that can be mapped to an encoding slot, and indicates whether a field is required or not to generate the chart. We express encoding constraints in a simple manner, indicating whether an encoding of some type requires a numeric or non-numeric field and for non-numeric fields also indicating the limitations on the number of unique categorical elements a field may contain. *For example, a bar chart has three encoding slots: x-position, y-position, and colour. The y-position constraint is that the mapped field must be numeric. The colour constraint is for a non-numeric field with fewer than 12 categories. Positional encoding slots must be assigned to some field in order to generate the bar chart, but colour encodings need not be assigned.* Figure 7.6 illustrates a specification template for two other chart types, scatter charts and phylogenetic trees.

Our template approach is easy to debug and clearly lays out the mappings between visual encodings and datatypes. Because these templates are provided within the system, the user does not have to provide any initial specifications at all. We are able to provide these templates within the system because the scope of chart types is known in advance; in contrast, previous mixed-initiative systems leave that burden with the user. Our templates only serve to provide constraints on the mapping from data to visual encodings. The intent of previously proposed query languages, such as VizQL [50] and CompassQL [127], is to also indicate what methods to use for choosing, ranking, and grouping recommendations; in contrast, we use the entity graph to perform these computations.

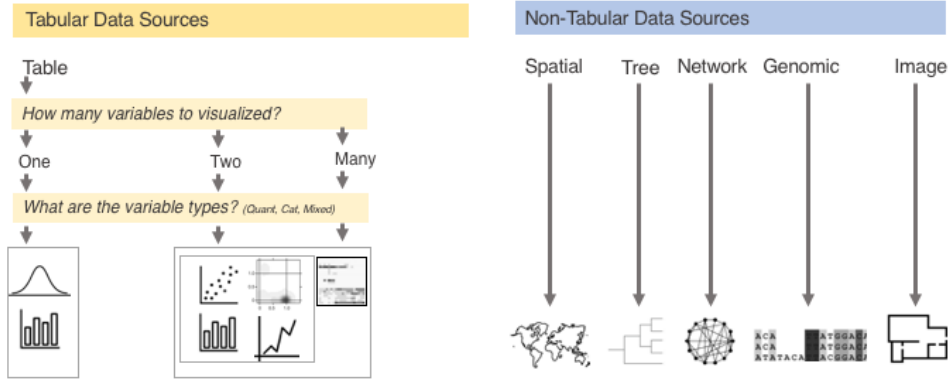


Figure 7.7: Mapping from datatypes to chart types. The above examples demonstrate how tabular and non-tabular data are mapped to different chart types. The mappings shown here are informed by a domain prevalence design space generated for microbial genomic epidemiology [24]. Our technique assumes that non-tabular data sources contain encodings for specific marks and their associated positional channels. Conversely, tabular data must be assigned to different marks and channels depending upon the total number of fields to be visualized and the type of field (numeric or non-numeric).

Our algorithm attempts to generate as many visual encodings as possible given the data sources along a path in the entity graph. In the implementation we have hard programmatic constraints on the total number of paths and visual encodings investigated, always prioritizing highly ranked paths and visual encodings. Data sources and attribute fields are mapped to data and encoding slots, respectively, within the internal visual encoding template. As previously indicated, fields that connect two or more datasets (highly connected nodes) have the highest priority for mapping to encoding slots. Our algorithm also allows a user to specify the names of fields that should be in the results and these are also preferentially mapped.

Tabular and non-tabular data are treated differently when mapping to specific encoding slots. We assume that non-tabular data has pre-specified marks and positional channels that produce specific visual encodings. If there exists associated data for the non-tabular datatype, then the associated data supplies fields that can be used to encode non-positional information. *For*

example, tree data data supplies the positions for different line marks and associated data encodes additional information through line colour, size, or style channels. Conversely, tabular data has no pre-specified marks or channels and can produce one or more statistical visual encodings (bar, scatter, histogram, etc.) depending upon available fields. Figure 7.7 presents a concrete mapping of data to visual encodings that was informed by a visualization design space for microbial genomic epidemiology.

If there are no highly connected nodes or no user defined fields to seed the encodings, then fields are selected at random to populate visual encoding templates. Similarly, if all high priority fields are assigned and there remain unassigned encoding slots, lower priority fields are randomly assigned to encoding slots.

7.5.5 Composing Views for Display

As a final step, visual encoding specifications with all required encoding slots assigned to an attribute field are rendered for display to the stakeholders. We generate a single view of coordinated static chart combinations per path. The implementation describes how these views are composed and how we again use the relevance of visual encodings to create a view with a manageable number of visual encodings.

7.6 Implementation of GEViTRec

We now present GEViTRec, which is an instance of the general algorithm described in Section 7.5 when applied to a domain-specific problem within public health microbial genomic epidemiology. We have implemented GEViTRec in the R programming language as a package and have publicly released it at <https://github.com/amcrisan/GEViTRec>. The repository also contains all analysis code for the results presented in Section 7.7.

GEViTRec uses the previously described GEViT domain prevalence design space (Section 7.2) to identify connections between visual encodings and datatypes and generate visualization recommendations. Our implementation supports tabular, genomic, spatial, network, tree, and genomic datatypes. Visual encoding specifications generated by GEViTRec are rendered by minCombinR, a system that we developed to support a minimal specification syntax for chart types and combinations [27]. The minCombinR systems supports the generation of 18 unique chart types and four combinations (small multiples, spatially aligned, colour aligned, and unaligned). Please refer to the minCombinR publication for more details.

Stakeholders can run GEViTRec in their R environments and we have developed a set of functions to help them load heterogeneous data, perform a data harmonization, generate specifications, and finally render and display views of visual encodings (Figure 7.8).

`input_data` is a common interface for loading different datatypes into R. This function requires as input the location of the data source on disk and the datatype. If applicable, for non-tabular datatypes associated data can also be loaded. We have developed a series of datatype specific functions that will load a dataset into the R environment and store it in a common data structure (Section 7.5.2).

`data_harmonization` takes a collection of datasets and “explodes” attribute fields from different data sources, finds linkages, and creates the entity graph. The `view_entity_graph` command will display it. Stakeholders can also view a metadata table for the different data sources and their attribute fields.

`get_spec_list` performs the computations on the entity graph. It ranks paths and processes them in order of rank (highest to lowest) to generate specifications for visual encodings.

`plot_view` takes a list of specifications, renders the visual encodings, and

arranges them for display. In instances where an entity graph contains multiple disconnected components, our current implementation limits the total number of views per component to ten; that is, we assemble views from up to ten paths within a single component. Our implementation also limits the number of chart types per coordinated static combination to five, which are selected based upon their relevance scores per datatype. However, these limitations are modifiable.

7.7 Results

We demonstrate the capabilities of GEViTRec with publicly available datasets from the 2013-2016 Ebola outbreak [34], which are included along with an R notebook of the analysis below in the GEViTRec code repository. These data include a phylogenetic tree for roughly 1610 Ebola virus genomic samples (each sample is unique to one person), spatial data of the affected nations, and tabular data with additional information for each sample. The primary reason for using these data is that they are publicly available and there exists considerable scientific research on this outbreak that allows us to objectively assess the quality of GEViTRec’s results. We were also motivated by the challenges present in the domain of genomic epidemiology, which in recent years saw both Ebola and Zika outbreaks. We heed the call to arms for better data sharing and analysis tools [42] in exploring how an automated visualization recommender could have presented these data to researchers.

GEViTRec code

```
# Analyze different
# data types automatically
harmon_obj<-data_harmonization(tab_dat,
tree_dat,genomic_dat,all_spatial)

# Create specifications
component_specs<-get_spec_list(harmon_obj)

#plot the result one view at a time
plot_view(component_specs,view_num=1)
```

Combination #1

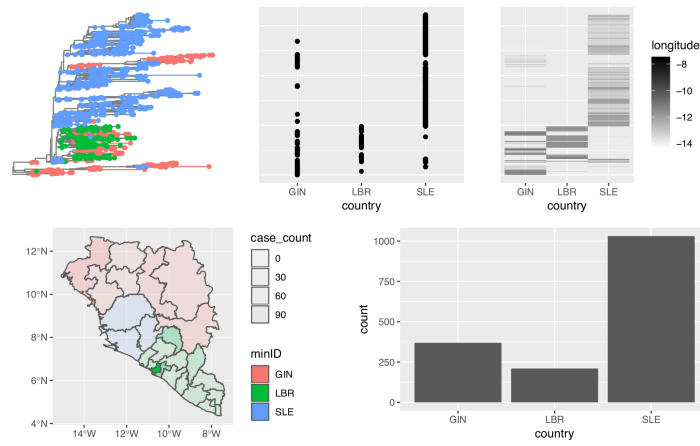
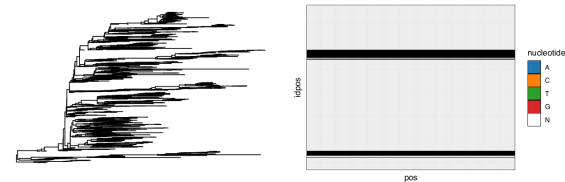


Figure 7.8: GEViTRec Results. Left: GEViTRec code to harmonize heterogeneous data into an entity graph, generate specifications based on high-ranking paths through it, and display coordinated combinations of static charts. Right: The highest-ranked combination, with a top row containing charts that have been spatially aligned along a common horizontal axis, and a bottom row coordinated through colour alignment and shared attribute fields. These automatic recommendations support fast data reconnaissance.

Entity Graph



Combination #4



Combination #5

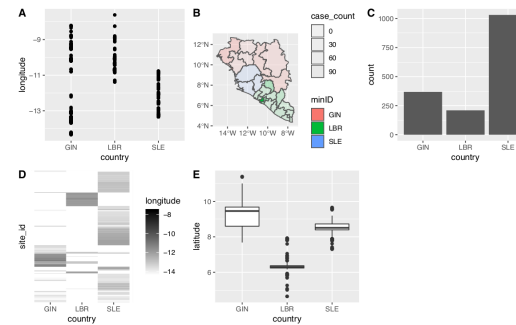


Figure 7.9: GEViTRec results with Ebola outbreak data. We load publicly available data including tabular, genomic, and spatial data of about 1610 viral genome samples, one from each affected person. On the left is the entity graph GEViTRec created in its data harmonization step, annotated in red to show the attribute names we use in our discussion. On the right are two additional chart combinations; the highest-ranked one appears in Figure 7.8. Combination #4 represents a path where a phylogenetic tree is automatically spatially aligned with genomic data. Combination #5 shows a path that links tabular and spatial data with a combination of five charts.

Without any input from the user, GEViTRec automatically generates coordinated combinations of charts, and where possible, has aligned charts according to shared positional axes and colour. Figure 7.8 shows the highest ranked combination, alongside the code to generate it. These three lines of simple code harmonize data, generate the specifications, and display the views; notably, the user does not specify anything about the mapping between the data and the visual encoding. Figure 7.9 shows the resulting automatically generated entity graph and two more of the coordinated combinations, ranked 4 and 5; all of the top five views are in Appendix E. We can see that GEViTRec has identified that `country` and `site_id` (the unique identifier for each viral sample) are the high-degree nodes in the entity graph. We can tell that these two attributes were the seeds for partial specifications because they are present in all applicable charts: `country` appears in all, encoded by position with some and colour with others. Many of the generated views contain a phylogenetic tree, the most relevant chart type for this domain. Our prior research [24] noted the tendency of stakeholders to show tree data with accompanying tables rather than appropriately coordinated with other charts; GEViTRec has overcome this limitation by facilitating such coordination automatically and without detailed input from the user.

Beyond the two attributes used to seed specifications, GEViTRec chooses the remaining attributes to visualize at random, which can produce some interesting results. For example, the latitudes and longitudes for where different viral samples were collected are treated as generic numeric variables, which results in charts of either latitude or longitude against the country. Generally this information is shown on geographic maps instead of the heatmap GEViTRec has chosen, but it would require additional information, either from the user or hard coded into the system, to know these are privileged attributes containing spatial information. It would be fruitful future work for us to automatically mine some of this attribute-level data from our existing domain prevalence design space and automatically incorporate this information.

It took approximately 35 seconds to produce these results including loading

data, harmonizing it, generating specifications, and displaying the visualizations. The findings we show Figure 7.9 are produced on a MacBook Pro (Quad-core Intel Core i5 CPU, 16 GB of RAM, although the full usage of these compute resources were not necessary).

It is instructive to compare the automatically created GEViTRec output to the current state of the art in this community, the Nexstrain and Microreact applications, both of which visualize the ebola virus data. Both feature a phylogenetic tree and a map; our highest-ranked path also has these two chart types, in addition to some others. Nextstrain features multiple interactive views that are coordinated through brushing and animation, but sits atop a highly controlled data analysis pipeline and is not easily adaptable to any other datasets. Microreact is more flexible to different datasets, which can be loaded as CSV files, but requires the user to manually specify colour coordination by explicitly adding paired colour columns to their tabular datasets. Neither can show any other visual encodings. The goal of GEViTRec is not to create polished tools for immediate public dissemination, but rather to facilitate very fast exploration of a data landscape and provide initial suggestions for visual encodings that may be helpful. These results show its utility in providing quick glances of the data in support of data reconnaissance.

7.8 Related Work

We situate GEViTRec within existing research in visualization recommendation systems categorized into rule-based, ontology, and machine learning approaches. We also compare our R-based implementation stack in detail to the JavaScript-based work of Heer and colleagues, discussing the similarities and differences of our design contexts and decisions.

7.8.1 Rule-Based Approaches

Previous rule-based approaches assign fields from tabular data to visual encoding marks and channels according to the field type (nominal, ordinal, quantitative). They further rank the resulting visual encodings based upon perceptual effectiveness scores that have been traditionally “hand-assigned” by experts [69] and more recently are learned from crowd-sourced studies [76]. First formalized in Mackinalys APT system [68], rule-based approaches are the most common strategy for visualization recommendation, with two notable implementations as Tableaus Show Me [69] module and the Compass and CompassQL recommender engine within the Voyager I [128] and II [129] systems, respectively. These systems are seeded with partial specifications supplied by users and then recommend additional ranked alternative visualizations according to perceptual effectiveness scores. Apart from stand-alone systems, rule-based approaches can also be implemented as decision trees, such as Data-to-Viz (www.datatoviz.com) or Chart Chooser (<http://experception.net/>), that guide users toward a specific visual encodings. Rule-based systems are powerful and transparent methods of visualization recommendations and are at the heart of many more advanced solutions.

GEViTRec uses similar rule-based approaches to assign both datatypes and fields to specific visual encodings. However, our approach explores broader types of data, visual encodings, and combinations of encodings than previous rule-based systems. Furthermore, GEViTRec uses relevance in lieu of perceptual effectiveness to rank visual encodings, a choice necessitated by the lack of perceptual studies and expert consensus of visual encodings that exist in our design space.

7.8.2 Ontology-Based Approaches

Ontology-based approaches use existing domain knowledge to recommend useful visualizations. Existing approaches such as SemViz [45] and Cammarano’s schema matching technique [14] are primarily developed for semantic web applications, where rich knowledge ontologies exist. As with GEViTRec, ontologies inject domain-aware prior knowledge to generate more informed visual encoding recommendations, while minimizing the burden to the user of providing initial specifications. The strategies used by GEViTRec rank to identify and rank paths that link data sources are similar to Cammarano’s technique, albeit for different input data structures. While there do exist some ontologies that GEViTRec could in theory use, such as GenEpiO ([46], a genomic epidemiology ontology), the use of ontologies in visualization recommendation may apply only in specific cases: they are unlikely to result in a general solution because ontologies are resource intensive to generate and maintain.

7.8.3 Machine Learning

Machine learning techniques can be used to automatically learn visual encoding recommendations based upon prior knowledge. Draco extends rule-based approaches by using the results from graphical perception studies to learn the effectiveness ranks of different visual encodings [76]. Prior research demonstrates the viability of crowd-sourcing graphical perception studies, which presents an opportunity to scale data collection procedures [52]. A different approach is presented by the Data2Vis system, which attempts to translate data to visualization encodings by learning from paired examples of data and visualization specifications [31]. Although this approach may scale, it is again unclear how robustly it will extend to non-tabular sources of data and complex chart types and combinations; we note the potential of biasing the results toward certain solutions. Overall, machine learning solutions to visualization recommendation are relatively new and there exist many

interesting and open challenges. The approaches taken by GEViTRec, Draco, and Data2Vis are not mutually exclusive and can be considered together as important components for future recommender systems.

7.8.4 Stack Comparisons

It is instructive to situate GEViTRec with respect to the previous work in two software stacks: the R-based stack it lies within (RStack), and the JavaScript-based stack (JSStack) developed by Heer and colleagues that includes the Compass [127], Draco [76], and Voyager [128] [129] systems. Both stacks have a base of visualization grammars implemented within different charting libraries (ggplot2 for RStack; D3 and Vega for JSStack), and higher-order charting libraries at the next level (minCombinR for RStack; Vega-Lite for JSStack). In both, the recommender systems (GEViTRec for RStack; Compass and Draco for JSStack) sit one level higher, and emit specifications for visual encodings that are rendered by these libraries. The JSStack also features faceted browsing with the Voyager I and II systems, while our RStack currently uses minCombinR’s capabilities to render coordinated combinations of static visual encodings (charts) that can be displayed in any R environment such as analysis notebooks or output to a graphics device. A custom R-based interactive browser could be created as future work, but it is not necessary to build a new interface to view our recommender output.

The most important difference between our approaches is that GEViTRec supports a much wider range of visual encodings and datatypes than any of the current JSStack recommenders. There is no technical reason why the algorithm we present in Section 7.5 could not have been implemented in JavaScript, but we chose an R-based implementation so that our recommender could be easily embedded with the analysis procedures of our stakeholder community. The Compass and Draco recommender systems in JSStack were developed for tabular data and it is not obvious how and whether their architectures would extend to a wider set of datatypes; we do

note that there is no low-level technical barrier to such future work, since D3 and Vega/Vega-Lite are highly expressive charting libraries.

Another important difference is that GEViTRec does not require any initial specifications or prompts from the user beyond simply importing datasets: the data harmonization and generation of encoding combinations is fully automatic. In contrast, both JSStack recommender systems require at least partial specifications ahead of generating recommendations. One small exception is the Voyager 2 system, which is able to produce histogram or bar chart univariate summaries for dataset attributes. However, Voyager does require additional partial specifications to generate multivariate charts and to recommend alternative charts. Voyager’s output is also presently limited to the generation of singular charts, with minor variations, whereas each GEViTRec specification is a coordinated combination of multiple encodings.

Finally, an interesting consideration is the degree of architectural control across stack layers. Each system in the JSStack is a modular piece, but was developed by the same research group over a multi-year period. In the RStack, we developed both the GEViTRec recommender described here and the minCombinR charting layer beneath it, but these lie atop a software ecosystem designed and built by many others.

7.9 Discussion and Future Work

We consider the generalizeability of our work, and the tradeoffs between relevance and perceptual effectiveness for ranking.

7.9.1 Generalizability

Our contributions are intended to generalize to other domain contexts, but we have not yet fully validated this claim. There are several rate-limiting obstacles to such a validation. One bottleneck is the effort involved in generating

domain prevalence design spaces. While our previously proposed method for doing requires a mix of automatic computation and human effort [24], a robust method that is fully automatic would address that problem. A more prosaic problem where automation would be more difficult is the amount of tedious and time-consuming work required to integrate domain-specific software packages that visually encode new datatypes. Alleviating these bottlenecks and developing more instantiations in other domains will be an exciting area of future work. In order to encourage others to build on these exploratory ideas, we have clearly outlined our assumptions in Section 7.4 and Section 7.5 and have made all of the code and source materials for our end-to-end implementation of GEViTRec publicly available.

7.9.2 Is Relevance Relevant?

Our design of a domain-aware visualization recommender prioritizes the relevance of different chart types, which we have defined by examining commonly used visualization strategies of domain experts, in contrast to the many existing recommender systems use graphical perceptual effectiveness to rank visual encodings. We raise the concern that relying on perceptual effectiveness as the sole ranking mechanism may not be sustainable at scale because of the large number of studies that would be necessary even to assess single charts, in light of the full range of visual encodings possible for a heterogeneous array of datatypes. Worse yet is the combinatorial explosion for the experiments required to understand combined or linked visual encodings, because these combinations introduce many perceptual questions that are challenging to isolate in a single experiment.

Nevertheless, one could argue that domain experts are not visualization experts and thus may visualize their data inappropriately or without a full awareness of the visualization design space. Although we agree that individual domain experts may not be fully aware of how to visualize their data, our finding is that the collective strategies of a large group of experts can reveal

a complex combinatorial design space. When we use this domain prevalence design space in a recommender systems, we can generate visual encodings that individual experts would not themselves consider.

Importantly, using relevant visualizations does not preclude passing judgments of efficacy. It would be fruitful future work to examine the tradeoffs between perceptual effectiveness and relevance. If perceptual experiments provide adequate coverage for any specific design space, then it would be possible to penalize relevant visualizations that are not perceptually effective. As a coarse measure, our prior GEViT study tagged some visualizations as being “good” or “missed opportunity” based upon our expert judgement as visualization researchers. We did not use these classifications in the current implementation of our recommender algorithm because we did not have enough tagged examples for adequate coverage of the space, but as we scale the exploration of visualization design spaces we can generate more of this labelled data for use by recommenders.

7.10 Conclusion

Heterogeneous and multidimensional data are already the norm in many domains and stakeholders are increasingly expected to use these complex data to derive informed insights. In our own collaborations we saw that stakeholders are struggling to understand their landscape of heterogeneous data and we find evidence that others in the visualization research community have encountered similar difficulties in their own collaborations. We contribute a framework for data reconnaissance and task wrangling that reifies these difficulties into a concrete vocabulary and processes that visualization researchers can use in their future work. Our framework identifies four phases (acquire, view, assess, pursue) that are repeated over time to gain a better understanding of the data landscape – both what is available and what is still required – and to develop a crisper notion of the tasks these evolving data can support. We propose that the process of exploring an unfamiliar data

landscape requires specific solutions and have developed a general algorithm that allows stakeholders to view relevant and automatically generated data visualizations that are informed by the data alone, through a harmonization process that generates an entity graph and finds top-ranked paths through it that produce coordinated combinations of chart types. We present a proof of concept of our algorithm through the implementation of GEViTRec, a domain-aware visualization recommendation system that supports many datatypes beyond tabular data, including genomic, spatial, temporal, and network data. It produces four types of coordinated combinations of visual encodings: spatially aligned, colour aligned, small multiples, and unaligned. In contrast to previous systems that suggest different ways to drawing a single chart, our emphasis is on generating many ways to draw many different kinds of data that are automatically coordinated. Our work adds to the growing research in visualization recommendation systems by exploring the application of such systems to more diverse types of data sources and visual encodings.

Chapter 8

Reflections and Conclusion

The greatest value of a picture is when it forces us to notice what we never expected to see — John W. Tukey

Technological change is enabling health care systems to collect and analyze an unprecedented amount and variety of data. If harnessed effectively, these data can be used to transform public health policy making, just as John Snow's seminal work did two centuries earlier. Snow's work drew on a combination of statistical analysis and data visualization to identify and then showcase the probable source of London's Cholera outbreak. While statistical methods have continued to develop along with the field of epidemiology, data visualization tools and practices have languished. The dissociation between statistical and visual analysis has become more pronounced as new sources of data need to be integrated, analyzed, and communicated to diverse groups of stakeholders who include, clinicians, nurses, researchers, policy makers, and the general public.

In this dissertation I have sought to **understand** the relationship between new and existing sources of data with the diagnostic, treatment, and surveillance tasks carried out by stakeholders in public health genomic epidemiology. I focused on a targeted set of TB stakeholders who were interested in integrating pathogen genomic data into their existing tasks but were uncertain of how to integrate genomic data. This understanding enabled me to design and assess ways of communicating these new types of data in an interpretable manner to

a diverse group of stakeholders and to identify constraints restricting access to data. Having a concrete idea of stakeholders data, tasks, constraints, and unmet needs I then sought to **explore and characterize** existing strategies for visualizing the heterogeneous and multidimensional data sources that comprise modern genomic epidemiology (genEpi) studies. I developed a systematic method to review data visualizations and created a typology capable of describing and enumerating a visualization design space. Finally, I merged the collective knowledge of the genEpi research community with the specific needs, data, and tasks of a select stakeholder group to **design and implement** several tools (Adjutant, the GEViT Gallery, minCombinR, and GEViTRec) that improve the creation of data visualizations for genomic epidemiological surveillance tasks and that integrate with a broad ecosystem of analytic methods.

I have taken an interdisciplinary research approach that borrows methods and techniques from information visualization, human computer interaction, machine learning, and statistics. I have innovatively woven this approach together to develop novel technical and domain specific contributions that are described throughout the chapters of this dissertation. The technical contributions may be applied to other domains beyond genomic epidemiology, but it will require further study before such a claim can be thoroughly validated. The domain specific contributions produce research findings and artifacts that others in public health, biological visualization, or information visualization communities can use to develop tools for the genEpi community. It is important to appreciate that genEpi is a critical component of how disease outbreaks will be monitored, prevented, and controlled in the future [34, 35, 42]. As climate change and other sources of environmental disruption will contribute to the emergence and spread of more disease outbreaks, the domain specific contributions I present here form useful source materials for others seeking to develop the analytic and visualization tools needed combat the disease outbreaks of the future.

Reflecting on my contributions, I begin by considering each project chapter

individually and for those already in press describe its post-publication reception. Next, I comment on my overall interdisciplinary research approach. Finally, I discuss the limitations and future trajectory of my research.

8.1 Reflections on Research Projects and Contributions

8.1.1 Regulatory and Organizational Constraints

In Chapter 2, I defined regulatory and organizational constraints and through a case study demonstrated the impact of these constraints on visualization design and evaluation. To address the impact of these constraints I modified a widely used Design Study Methodology (DSM) using methods from agile software development and statistical analysis. I also provided a set of six recommendations that visualization researchers and practitioners could incorporate into their process. The technical contributions of this work were the modifications to the DSM that could be applied to other research projects. The domain specific contributions were artifacts specific in the form of a resolved stakeholders “power-interest” matrix and case study specific to public health genomic epidemiology.

The research in Chapter 2 was highly influential in my subsequent projects because it laid out the foundation of my research environment. It influenced the approaches I took to create multiple sources of knowledge, from specific stakeholder groups as in Chapter 3 and [25], to broader community strategies for data visualization as shown in Chapter 5 and [24]. The findings from the analysis of these different knowledge sources would go on to influence the design and implementation of different visualization tools. Importantly, actively incorporating regulatory and organizational constraints into my doctoral research enabled me to circumvent the restrictions imposed by these constraints and to develop usable tools.

Since the publication of this manuscript, regulatory attitudes toward data and

especially data privacy have changed significantly and imposed constraints on fields beyond healthcare. Most notably, the European Union General Data Protection Regulation (GDPR) came into effect in May 2018 and began to impose constraints on both commercial and research use of data globally. At the time of this writing, there continues to be a growing and active discussion on how data should be analyzed and, by extension, visualized. In light of these more recent developments, the content of Chapter 2, which was published in 2016, was a foreshadowing of how visualization researchers and practitioners should position themselves to address growing constraints on data access and use.

8.1.2 Evidence Based Design

In Chapter 3, I presented a multi-phased mixed approach that integrated with a DSM to collect data using both qualitative and quantitative methods. The pretext of this project was a collaboration with the COMPASS-TB team of Public Health England to redesign a clinical report communicating the results for mycobacterium whole genome sequencing, but my goal was also to collect detailed information on the connection between stakeholders data and tasks. In carrying out the report redesign research, I along with my collaborators assessed design preferences to represent the information on the clinical report. These design preferences included wording choice, visual emphasis, layout, and data visualization representation. Our approach compared against “control” designs from the existing clinical report. We also assessed user preferences from presentations of individual “isolated components” as well as whole clinical reports. Our findings showed that isolated components were better at eliciting concrete preferences compared to whole reports, which has implications for how design alternatives should be assessed more generally. The technical contributions of this work was the novel mixed methods approach to visualization design and evaluation and the set of experimental and design guidelines. The domain contributions pertained to the links between specific data types and genEpi tasks as well

as the final re-designed mycobacterium WGS clinical report. One limitation of the clinical report contribution is its specificity to tuberculosis data. Compared to other organisms, TB has a large and comparatively less complex genomic structure, while other pathogens have complex genomic structures that include plasmids and mobile elements that confer mechanisms of drug resistance that are difficult to detect with certainty and thus report. While I believe that the mixed method approach we present is capable of identifying and designing for such scenarios, the current implementation of our WGS clinical report does not directly address challenges arising from more complex genomic structures.

The findings from the research in Chapter 3 provided some of the first insights into the modern genomic epidemiology data landscape. Surprisingly, it revealed that there existed little consensus towards the data that should be used for genomic epidemiology surveillance tasks. This project also showed how the confidence of stakeholders to incorporate and interpret genomic data with existing data sources was variable; some stakeholders were much more confident than others toward their ability to use genomic data within their current set of tasks. The ambiguity surrounding genEpi data and tasks motivated me to gather a broader community perspective that would become the GEViT research (Chapter 5). In particular, I reasoned that a research community trying to understand its own data might not be well suited to articulate their preferences toward the visualization of data much more complex than that shown on the mycobacterium clinical report. This insight led to the development of the data reconnaissance and task wrangling framework presented in Chapter 7.

The methodology and resulting clinical report presented in Chapter 3 has been well received by the wider research communities. It was among the most widely viewed manuscripts on PeerJ (its publication venue) in 2018. I presented this work at the 2018 American Society for Microbiology Next Generation Sequencing Conference and my co-authors have also presented this research in other venues. I also presented a poster of this research at

the 2017 IEEE VIS conference, showcasing how domain specific examples can be used to translate infovis research knowledge. The re-designed clinical report was made available online as a LaTeX template and could be programatically filled in. Since its release, and as of this writing, the report template has been viewed over 3000 times and has been incorporated by PHE, ReSeqTB, and PathogenWatch analytic pipelines. Anecdotally, the community consensus has also been that this research tackles the important, but frequently overlooked, topic of effectively communicating results from genomic analyses.

8.1.3 Adjutant

The Adjutant tool presented in Chapter 4 was a serendipitous addition to dissertation research that evolved from general interest in the text mining approach I developed for the systematic visualization review methodology presented in Chapter 5. Adjutant is R-based and was released as a open source tool. Since it was released online, Adjutant has developed an active user community, especially in under resourced settings that are not able to afford more expensive text mining tools. I have continued to maintain Adjutant over the course of my doctoral dissertation thanks to feedback from its community. Adjutant’s primary contributions were technical: the development of rapid unsupervised topic clustering method.

Ultimately, Adjutant was also influential in the implementation of minCombinR and GEViTRec because it established a practical model for development and deployment of a maintainable R-based open source tool.

8.1.4 GEViT and the GEViT Gallery

In Chapter 5 I presented a systematic method for surveying data visualizations and that can be used to generate a typology to describe and enumerate a visualization design space. I also applied this method to the genEpi

research literature and developed a genomic epidemiology visualization typology (GEViT) that describes *how* data were visualized for different creation contexts. GEViT contains three taxonomies that describe chart types, combinations, and enhancements. The Adjutant system was born out of the need to rapidly and effectively summarize the structure of a document corpus comprising the genEpi research literature. The GEViT gallery was developed to provide stakeholders with an interactive means to browse a visualization design space and explore different visual alternatives. The primary technical contribution of this work is the systematic visualization review methodology. I believe this method can generalize very well to other application domains beyond genomic epidemiology, but this claim has not yet been validated. The domain specific contributions were GEViT and the GEViT Gallery, which summarize the visualization strategies of the genEpi research community.

The research in Chapter 5 was important for connecting data to specific visual representations and identifying unmet visualization needs. The analysis of the visualization design space reveals that only a minority of publications contained more sophisticated visual designs while the majority defaulted to using primarily phylogenetic trees. The extensive use of trees contrasted with our findings from Chapter 3 that showed many public health stakeholders did not find trees useful and were unsure how to clinically interpret them. I was also surprised by the amount of text that accompanied the visualizations that were analyzed and I speculated that this overuse of text contributed to the interpretability difficulties of trees. Closely reviewing the visualizations from this project, I also noticed artifacts of post processing in different visualizations, especially those that used less text and more combinations of different chart types to show different aspects of the data. I hypothesized that stakeholders did not have enough support to visualize the various types of charts and their combinations and that less programmatically sophisticated stakeholders suffered the most because they could not develop their own solutions as others could. This finding motivated the

development of minCombinR, a toolkit that supported generation of different chart types and their combinations through a minimal specification syntax. Importantly, minCombinR is the first toolkit that supports the visualization design space revealed by GEViT. The GEViTRec tool uses the data from GEViT to automatically generate data visualizations. The motivation for developing GEViTRec was also to support stakeholders exploration of the visualization design space using their own data.

GEViT and the GEViT Gallery have been well received by the public health genomic epidemiology and biological visualization research community. The GEViT gallery receives on average 47 monthly users and has a total viewership of 662 unique users from countries all over the world. The majority (approximately 77%) of users of the GEViT gallery find the site via a direct link, whereas the rest of traffic is derived from organic searches (15%) and social media (8%). I have presented this research at several venues including the 2017 Applied Bioninformatics and Public Health Microbiology Conference, 2018 American Society for Microbiology Next Generation Sequencing Conference, and the 2019 Visual and Automated Disease Analytics Summer School. I have also been invited to present this research as part of the Harvard Department of Biomedical Informatics Data Insights Seminar Series and at the 2018 Biological Visualization Dagshtul Seminar. Finally, I also presented this research as a poster at the 2018 IEEE Vis Conference and at the 2018 Canadian Student Health Research Forum, where I won a CIHR Gold distinction for the research and its presentation. While the GEViT work continues to mature, the overall response has been that I have tackled a complex problem with a creative and innovative approach. I have been frequently asked to apply this method to other application domains outside of genEpi and I have a number of directions to take this research in that I will describe in the future work.

8.1.5 minCombinR

In Chapter 6, I presented the minCombinR toolkit. I developed minCombinR to address unmet needs of genomic epidemiology stakeholders to generate data visualizations that were needed to support their analysis of heterogeneous data sources. The primary technical contribution for minCombinR was an software architectural framework that supports a consistent declarative syntax for generating data visualizations. Through only three commands, `specify_single`, `specify_combination`, and `plot` stakeholders could expressively generate a variety of data visualizations. This architectural framework implemented a gradual binding technique that enabled minCombinR to add to or modify an initial user provided visualization specification. Gradual binding allows minCombinR to harmonization multiple different chart types so that they can be more cohesively combined. While minCombinR can be applied more generally, it is primarily a domain specific contribution because it is optimized to support genEpi data and visualizations. However, I do demonstrate in Appendix D that minCombinR can used for datasets from other domains.

The minCombinR toolkit also serves as a bridge between the prior GEViT research and the subsequent GEViTRec automated visualization algorithm. It is too early to assess the broader impact of minCombinR, as it has thus far primarily been presented in conjunction with GEViTRec.

8.1.6 GEViTREC

Finally, the research Chapter 7 presents the culmination of these various research threads throughout this dissertation. The GEViTRec implementation combines domain specific data visualization knowledge, with the expressivity of the minCombinR toolkit, and the versatility of recommendation algorithm. GEViTRec is an innovative approach to help stakeholders visualize their own data with the assistance of an algorithm. The algorithm-

mic technique unpinning GEViTRec represents multiple heterogeneous data sources as a graph and uses this graph structure to explore and rank potential ways to visualize the data. This research also introduces a novel method for ranking data visualization: *relevance*. Relevance ranking is informed by the domain prevalence visualization design space described in Chapter 5 and is an alternative to the widely used graphical perception based ranking. GEViTRec is the first automated data visualization generating algorithm that supports non-tabular data, a complex diversity of chart types beyond common statistical charts (scatter charts, bar charts, *etc.*), and that only requires the datasets of interest as input. The development of GEViTRec is motivated by the larger challenges of exploring unknown data landscapes that accompany modern genEpi investigations. I identified data reconnaissance and task wrangling as two co-ordinated processes that are carried out over iterative phases of stakeholders acquiring data, quickly viewing these data, making an assessment, and then if needed pursuing additional data sources. GEViTRec attempts to speed up this iterative multi-phased process by lowering the burden to view data and make a quick assessment. The technical contributions of this work are the recommendation algorithm as well as the data reconnaissance and task wrangling framework. The resulting GEViTRec implementation is a domain specific contribution.

I presented GEViTRec at the 2019 Applied Bioinformatics and Public Health Microbiology conference as part of the bioinformatics showcase. The response was very positive. Conference attendees were eager to learn about GEViTRec and its development. I demonstrated how GEViTRec worked within the wider R ecosystem, including analysis methods, and some attendees asked to change the demonstrate code so that they explore how GEViTRec worked. GEViTRec was nominated by attendees, and awarded, the best showcase application. Furthermore, the Data Reconnaissance and Task Wrangling framework I developed to motivate the development of GEViTRec is also gaining some traction. In discussion with multiple groups, the notion of data reconnaissance in particular is evocative and a number of

groups have indicated that it succinctly describes their current processes and challenges. It remains early to assess the impact of GEViTRec as well as the data reconnaissance and task wrangling framework, but the initial response is promising.

8.2 Reflecting on the Merits and Challenges of Interdisciplinary Research

There are many merits to an interdisciplinary approach to research, but it can also introduce many challenges and there exist few resources to instruct one of how to best proceed. I have included this reflection in my dissertation for future researchers that consider taking a similar approach.

Overall, taking an interdisciplinary research approach has largely benefited my dissertation projects. The training I received prior to my doctoral research emphasized quantitative methods founded in statistical, computational, and epidemiological disciplines. My doctoral studies introduced me to qualitative and especially mixed methods research approaches drawn from infovis design study methodologies and Human Computer Interaction (HCI) user research. Throughout this dissertation research, I borrowed techniques from across these disciplines in order to gain a broader and richer perspective on the role of data visualization in public health genomic epidemiology. Refining the integration of these methods over time helped me to identify fruitful research contributions that enabled my research to have tangible impacts in a relatively short period of time. Beyond these tangible impacts, interdisciplinary research was also intellectually satisfying. It was interesting to mentally wrestle with the different philosophies and research approaches taken by different disciplines, especially when these disciplines are at different stages of maturity. Information visualization and human computer interactions are quite young and in my perspective they are still developing their methodologies. Epidemiology and statistics have much longer history with more established methods but are not necessarily well equipped

to address many of the challenges I encountered within different research projects. Working within disciplines of different maturity afforded me the opportunity to reflect on the process and history of science, something that I had previously taken for granted.

There were also many challenges that attended this interdisciplinary research approach. The primary challenge was that I had to learn to speak two or more research languages and adaptively translate my findings into different, audience dependent, languages. This translation effort was an error prone process that introduced a considerable overhead to my doctoral research. Even as I improved upon ability to translate my research across different audiences and disciplines, it was still challenging to identify that appropriate target audience to begin with. For chapters 3 to 5, inclusive, I made the decision to target a bioinformatics, microbial genomics, and public health audience because the research discussed domain specific needs, data, and tasks. Having my work accepted and published by that audience was also an indirect validation of my research findings. Furthermore, I used those publications to introduce and translate infovis and HCI methods to this target audience. In Chapters 6 and 7 the primary contributions were novel framework for a visualization toolkit and recommendation algorithm, respectively. I determined that an infovis audience would find these contributions interesting and would be more capable of assess the rigour of my contributions. Despite these decisions, it would have easily been possible to divide the same work differently when it was presented to these audiences.

While I perceived it beneficial to have a wide and diverse audience for my research output, in reality there were setbacks to this strategy. The different communities do not necessarily read each other's literature, which served to diffuse the messages of my research rather than reinforce the findings I had developed over successive projects. Early on I had decided to use conferences and workshops as venues to cross pollinate my published research between different communities, but again this strategy was not as effective as I would have liked. Interestingly, at the outset of my doctoral

research I had anticipated that my own limitations to synthesize and integrate methods from different disciplines would have been the rate limiting factor. In fact, this effort of integrating techniques between disciplines was not as time consuming as managing my message to different audiences.

In spite of these challenges, my attempts to translate and integrated knowledge from multiple different disciplines have paid off. I have been invited to lecture on the intersection of data visualization, data science, and public health by a number of organizations, including the BC Centre for Disease Control, the American Public Health Association, Population Data British Columbia, the Canadian Bioinformatics Workshops, the Canadian National Microbiology Laboratory, and the Visual and Automated Disease Analytics Summer School. The response to these lectures has been quite positive.

8.3 Overall Limitations and Future Work

Within the individual chapters, I have presented the limitations and future directions of each research project. Here I will comment on the overall limitations of my doctoral research and its collective future trajectory.

The primary limitation of this research is the scope of the application domain and evaluative methods of the different tools I have developed. The narrow scope of public health genomic epidemiology was intentional, but as the research progressed it became clear that the principal challenges experienced by my collaborators and the broader genEpi research community were not domain specific. The Adjutant, GEViT Gallery, minCombinR, and GEViT-Rec tools thus have components that could be readily applied to other domain applications, which is why I have separated contributions into technical and domain-specific categories. Still, these methods are tailored and optimized for genEpi and I have not conducted extensive evaluations in this dissertation to assess that these technical contributions truly are generalizable. The scale of the effort needed to robustly assess this claim of generalizability is beyond

the scope of this dissertation.

The research and contributions described in this dissertation form a foundation that provides a trajectory and initial implementations for data visualization tools that help stakeholders navigate and visualize the complex data landscapes of the future. Data that concerned this dissertation was primarily derived from instruments that measure and record human biological data, such as genomics, treatment responses, transmission and so on. However, the landscape of health data grows more complex, with social media and online search queries becoming a major way that stakeholders provide, explore, look-up, and consume health information. As these online data sources become used and potentially integrated with others, privacy concerns stemming from regulatory and organizational constraints will play an ever larger role in determining whether data can be used and for what purposes. It is necessary to continue developing visualization and analytics tools that help stakeholders integrate, analyze, and visualize these new sources of data so that they can derive actionable insights that inform population and public health policy making.

One future avenue of this research is to further explore how synthetic, simulated, or even encrypted data can be visualized. In Chapter 2 I had begun to explore how synthetic datasets could be used in lieu of real data for evaluating visualization prototypes. I showed that using this data was more acceptable to stakeholders that had concerns about data privacy. However, the approach I used in the Chapter 2 case study could be expanded by looking to the biostatistics community, which has developed a number of different strategies for using synthetic and simulated data within statistical analyses, and the machine learning and database communities that have been making recent advances in the area of differential privacy. Combined, these emerging research on privacy preserving statistical and machine learning analyses can be leveraged with data visualization tools in the future. Such advances in privacy preserving data visualizations would not adversely impact the research tools I have developed in this dissertation. In fact, I have

purposefully developed tools that are able to adaptively respond to a variety of data types and easily extend to new data. By foregoing an expectation of specific data configurations, the minCombinR and GEViTRec tools are able to incorporate a variety of synthetic and simulated datasets. Privacy preserving data visualization is a fruitful and underexplored area of data visualization research, but I anticipate it will play a much larger role in the future.

Another viable future avenue of this research is to develop a single automated pipeline that links the generating of a domain prevalence design space (Chapter 5) to the automatic generation of data visualizations from a stakeholders datasets (Chapter 7). Generating such a pipeline would also allow my research to be assessed in other domains and to test the generality of the algorithms and techniques that comprised the technical contributions of this doctoral research. Looking beyond these immediate aims, a pipeline that connects domain knowledge to stakeholder data can also be leveraged toward developing explainable machine learning models. Presently, machine learning models are debugged using descriptive statistics that summarize model performance, such a confusion matrix, accuracy, loss, etc. However, it remains up to the individual to further probe the data to understand the model's performance. I envision a much more expressive and dynamic relationship, where the model can attempt to *show* a stakeholder how it came to its decision. Awareness of the visualization exploration and communication strategies used by stakeholders can help machine learning systems derive relevant, potentially more intuitive, and responsive explanations for their decisions. Furthermore, stakeholders could interact through these visualizations to engage in a more active dialogue with the machine learning algorithm by correcting errors or introducing new information. This collaborative interplay between human domain knowledge and machine statistical knowledge, mediated through data visualization, will be essential to the future of data driven decision making. The research I present here puts forth novel approaches for achieving this collaborative relationship between data,

people, and automated systems.

8.4 Concluding Remarks

Human understanding and descriptions of natural phenomenon have undergone many major shifts since our early ancestors began to question their environment and their place in it. John Snow's seminal research represents one such shift that championed the collection, analysis, and communication of data to understand these natural phenomenon and to derive an actionable insights that lead to an effective intervention. While today Snow's approach may seem like the norm, it was novel and innovative for its time. In the two centuries that have since passed, data-driven decision making remained the purview of a select few groups armed with enough money to purchase the necessary equipment and personnel to collect, analyze, and make decisions with data. With the accelerated evolution of computing technologies, it has become possible to collect a greater amount and variety of data and to analyze these data with ever more sophisticated means. This has been especially true in public health, where data can be collected from a variety of sources including from biological samples, healthcare administrative records, geographical contextual information, social media, and even environmental sources. In this dissertation, I have described the challenges and enthusiasms that stakeholders in public health face when using these new types of data in their routine diagnostic, treatment, and surveillance tasks. I have also characterized their attempts to visually encode this data, which represents the cutting edge of how stakeholders situate and communicate their findings. Most importantly, I linked these needs and strategies to the development of software systems that are meant to help stakeholders analyze and visualize their data. As datasets continue to grow in size and complexity, the methods, techniques, and artifacts that I have presented are important for developing data visualization tools that help stakeholders find viable paths through an evolving and unfamiliar data landscape.

Bibliography

- [1] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers. Visualizing sets and set-typed data: State-of-the-art and future challenges. In *EuroVis State of the Art Report*. The Eurographics Association, 2014. ISBN -. doi:10.2312/eurovisstar.20141170. → pages 6, 88, 157
- [2] J. S. Ancker, Y. Senathrajah, R. Kukafka, and J. B. Starren. Design Features of Graphs in Health Risk Communication : a Systematic Review. *Journal of the American Medical Informatics Association*, 13(6):608–619, 2006. doi:10.1197/jamia.M2115.Introduction. → page 52
- [3] K. Andrews. Evaluation Comes in Many Guises. In *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, volume 1, pages 8–10, 2008. → pages 34, 36
- [4] S. Argimón, K. Abudahab, R. J. E. Goater, A. Fedosejev, J. Bhai, C. Glasner, E. J. Feil, M. T. G. Holden, C. A. Yeats, H. Grundmann, B. G. Spratt, and D. M. Aanensen. Microreact: Visualizing and Sharing Data for Genomic Epidemiology and Phylogeography. *Microbial Genomics*, 2, 2016. doi:10.1099/mgen.0.000093. → pages 4, 51, 87, 122
- [5] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–14, 1987. → page 113
- [6] J. Bertin and W. J. Berg. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, Wisconsin, 1st ed edition, 1983. ISBN 978-1-58948-261-6. → pages 104, 159

- [7] A. Bigelow, S. Drucker, D. Fisher, and M. Meyer. Iterating between tools to create and edit visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):481–490, 2016. doi:10.1109/TVCG.2016.2598609. → page 114
- [8] M. Bostock. Observable notebooks. <https://observablehq.com/>, 2019. Accessed: 2019-03-25. → pages 113, 120
- [9] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. doi:10.1109/TVCG.2011.185. → pages 3, 120, 124
- [10] M. Bouchet-Valat. SnowballC: Snowball stemmers based on the C Libstemmer Utf-8 library. <https://CRAN.R-project.org/package=SnowballC>, 2014. → page 278
- [11] P. Bradley, N. C. Gordon, T. M. Walker, L. Dunn, S. Heys, B. Huang, S. Earle, L. J. Pankhurst, L. Anson, M. de Cesare, P. Piazza, A. A. Votintseva, T. Golubchik, D. J. Wilson, D. H. Wyllie, R. Diel, S. Niemann, S. Feuerriegel, T. A. Kohl, N. Ismail, S. V. Omar, E. G. Smith, D. Buck, G. McVean, A. S. Walker, T. E. A. Peto, D. W. Crook, and Z. Iqbal. Rapid Antibiotic-resistance Predictions from Genome Sequence Data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nature Communications*, 6:10063, Dec 2015. doi:10.1038/ncomms10063. → page 51
- [12] M. Brehmer, S. Carpendale, B. Lee, and M. Tory. Pre-design Empiricism for Information Visualization: Scenarios, Methods, and Challenges. *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, 1:147–151, 2014. → pages 31, 36
- [13] B. Budowle, N. D. Connell, A. Bielecka-Oder, R. R. Colwell, C. R. Corbett, J. Fletcher, M. Forsman, D. R. Kadavy, A. Markotic, S. A. Morse, R. S. Murch, A. Sajantila, S. E. Schmedes, K. L. Ternus, S. D. Turner, and S. Minot. Validation of High Throughput Sequencing and Microbial Forensics Applications. *Investigative Genetics*, 5:9, 2014. doi:10.1186/2041-2223-5-9. → page 51
- [14] M. Cammarano, X. Dong, B. Chan, J. Klingner, J. Talbot, A. Halevy, and P. Hanrahan. Visualization of heterogeneous data. *IEEE*

Transactions on Visualization and Computer Graphics, 13:
1200–1207, 2007. URL
<http://vis.stanford.edu/papers/visualization-heterogeneous-data>. →
page 177

- [15] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-Based Clustering Based on Hierarchical Density Estimates. *Advances in Knowledge Discovery and Data Mining*, -:160–172, 2013. doi:10.1007/978-3-642-37456-2_14. → pages 84, 92, 278
- [16] N. Cao and W. Cui. Overview of Text Visualization. In *Introduction to Text Visualization*, pages 11–40. Atlantis Press, Paris, France, 2016. doi:10.2991/978-94-6239-186-4_2. → page 84
- [17] S. Carpendale. Evaluating Information Visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, pages 19–45. Springer, Berlin, Heidelberg, 2008. ISBN 978-3-540-70956-5. doi:10.1007/978-3-540-70956-5_2. URL https://doi.org/10.1007/978-3-540-70956-5_2. → pages 94, 317
- [18] L. N. Carroll, A. P. Au, L. T. Detwiler, T.-C. Fu, I. S. Painter, and N. F. Abernethy. Visualization and Analytics Tools for Infectious Disease Epidemiology: A Systematic Review. *Journal of Biomedical Informatics*, 51:287–298, apr 2014. doi:10.1016/j.jbi.2014.04.006. → pages 2, 5, 27, 51
- [19] T. W. Chang, Kinshuk, N. S. Chen, and P. T. Yu. The Effects of Presentation Method and Information Density on Visual Search Ability and Working Memory Load. *Computers and Education*, 58 (2):721–731, 2012. doi:10.1016/j.compedu.2011.09.022. → page 79
- [20] K. Charmaz. *Constructing Grounded Theory*. Sage Publications, London; Thousand Oaks, Calif, 2006. ISBN 978-0-7619-7352-2. → pages 18, 91, 94, 317
- [21] T. Cohen, G. K. Whitfield, R. W. Schvaneveldt, K. Mukund, and T. Rindflesch. EpiphaNet: An Interactive Tool to Support Biomedical Discoveries. *Journal of Biomedical Discovery and Collaboration*, 5: 21–49, sep 2010. ISSN 1747-5333. doi:10.5210/%2Fdisco.v5i0.3090. URL <https://bit.ly/2JqjDg4>. → page 82

- [22] J. W. Creswell. *Research Design: Qualitative, Quantitative, and Mixed Methods approaches*. Sage Publications, Thousand Oaks, CA, 2014. ISBN 9781452274614. doi:10.1007/s13398-014-0173-7.2. → pages 14, 56, 57, 317
- [23] A. Crisan, J. L. Gardy, and T. Munzner. On regulatory and organizational constraints in visualization design and evaluation. *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, 1:1–9, 2016. doi:10.1145/2993901.2993911. → pages vi, 2, 12, 24, 150
- [24] A. Crisan, J. L. Gardy, and T. Munzner. A Systematic Method for Surveying Data Visualizations and a Resulting Genomic Epidemiology Visualization Typology: GEViT. *Bioinformatics*, 35(10):1668–1676, 09 2018. doi:10.1093/bioinformatics/bty832. → pages viii, 17, 86, 114, 116, 151, 157, 160, 168, 174, 180, 185
- [25] A. Crisan, G. McKee, T. Munzner, and J. L. Gardy. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ*, 6:e4218, Jan. 2018. doi:10.7717/peerj.4218. → pages vi, 2, 14, 49, 108, 148, 185
- [26] A. Crisan, T. Munzner, and J. L. Gardy. Adjutant: an R-based Tool to Support Topic Discovery for Systematic and Literature Reviews. *Bioinformatics*, 35(6):1070–1072, 08 2018. doi:10.1093/bioinformatics/bty722. → pages vii, 16, 81, 91, 92
- [27] A. Crisan, S. Fisher, S. Kasica, J. L. Gardy, and T. Munzner. minCombinR: Coordinating chart combinations with minimal specifications. *Under Review*, 09 2019. → pages viii, 112, 159, 170
- [28] A. Crisan, J. L. Gardy, and T. Munzner. GEViTRec: Domain-aware visualization recommendation for data reconnaissance and harmonization. *Under Review*, 09 2019. → pages ix, 146
- [29] T. M. Daniel. The History of Tuberculosis. *Respiratory Medicine*, 100(11):1862–1870, 2006. doi:10.1016/j.rmed.2006.08.006. → page 42
- [30] B. K. Dhillon, M. R. Laird, J. A. Shay, G. L. Winsor, R. Lo, F. Nizam, S. K. Pereira, N. Wagglechner, A. G. McArthur, M. G. Langille, and F. S. Brinkman. Islandviewer 3: More flexible, interactive genomic

island discovery, visualization and analysis. *Nucleic Acids Research*, 43(W1):W104–W108, Mar 2015. ISSN 0305-1048.
doi:10.1093/nar/gkv401. → page 4

- [31] V. Dibia and Ç. Demiralp. Data2Vis: Automatic generation of data visualizations using sequence to sequence recurrent neural networks. *CoRR*, abs/1804.03126, 2018. URL <http://arxiv.org/abs/1804.03126>. → page 177
- [32] T. Driscoll, J. L. Gabbard, C. Mao, O. Dalay, M. Shukla, C. C. Freifeld, A. G. Hoen, J. S. Brownstein, and B. W. Sobral. Integration and Visualization of Host-pathogen Data Related to Infectious Diseases. *Bioinformatics*, 27(16):2279–87, Aug 2011. ISSN 1367-4811. doi:10.1093/bioinformatics/btr391. → page 51
- [33] G. Dudas. Baltic. <https://github.com/evogytis/baltic>, 2019. Accessed: 2019-03-25. → page 122
- [34] G. Dudas, L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D’Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah, J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. T. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vand, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wölfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Günther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keita, P. Kellam, M. P. G. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Ströher, I. Wurie, M. A. Suchard, P. Lemey, and A. Rambaut. Virus Genomes Reveal Factors that Spread and Sustained the Ebola Epidemic. *Nature*, 544: 309, Apr 2017. doi:10.1038/nature22040. → pages 140, 171, 184

- [35] N. R. Faria, E. C. Sabino, M. R. T. Nunes, L. C. J. Alcantara, N. J. Loman, and O. G. Pybus. Mobile real-time surveillance of Zika virus in Brazil. *Genome Medicine*, 8(1):97, Sept. 2016. doi:10.1186/s13073-016-0356-2. → pages 87, 184
- [36] B. Fitzgerald, K. J. Stol, R. O’Sullivan, and D. O’Brien. Scaling Agile Methods to Regulated Environments: An Industry Case Study. *International Conference on Software Engineering*, 1:863–872, 2013. doi:10.1109/ICSE.2013.6606635. → page 29
- [37] B. Flyvbjerg. Five Misunderstandings About Case-Study Research. *Qualitative Inquiry*, 12(2):219–245, Apr 2006. doi:10.1177/1077800405284363. → page 40
- [38] R. Foster. Don’t Go Chasing Waterfalls: A More Agile Healthcare.gov, 2013. URL <http://www.newyorker.com/tech/elements/dont-go-chasing-waterfalls-a-more-agile-healthcare-gov>. → page 29
- [39] B. Fry and C. Reas. Processing.js. <http://processingjs.org/>, 2019. Accessed: 2019-03-25. → page 120
- [40] Y. Fukui, K. Aoki, S. Okuma, T. Sato, Y. Ishii, and K. Tateda. Metagenomic Analysis for Detecting Pathogens in Culture-negative Infective Endocarditis. *Journal of Infection and Chemotherapy*, 21(12):882–884, 2015. doi:10.1016/j.jiac.2015.08.007. → page 50
- [41] D. Furniss, A. Blandford, and P. Curzon. Confessions from a grounded theory phd: Experiences and lessons learnt. In *Conference on Human Factors in Computing Systems*, CHI ’11, pages 113–122, 2011. ISBN 978-1-4503-0228-9. doi:10.1145/1978942.1978960. → page 317
- [42] J. L. Gardy and N. J. Loman. Towards a Genomics-informed, Real-time, Global Pathogen Surveillance system. *Nature Reviews Genetics*, 19(1):9–20, 2018. ISSN 14710064. doi:10.1038/nrg.2017.88. → pages 2, 171, 184
- [43] A. S. Gargis, L. Kalman, and I. M. Lubin. Assuring the Quality of Next-generation Sequencing in Clinical Microbiology and Public Health Laboratories. *Journal of Clinical Microbiology*, 54(12):

2857–2865, Dec 2016. ISSN 1098-660X.
doi:10.1128/JCM.00949-16. → page 51

- [44] S. Ghani, B. C. Kwon, S. Lee, J. S. Yi, and N. Elmqvist. Visual analytics for multimodal social network analysis: A design study with social scientists. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2032–2041, Dec 2013. ISSN 1077-2626. doi:10.1109/TVCG.2013.223. → page 152
- [45] O. Gilson, N. Silva, P. W. Grant, and M. Chen. From web data to visualization via ontology mapping. In *Proc. Joint Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis’08, pages 959–966, 2008. doi:10.1111/j.1467-8659.2008.01230.x. → page 177
- [46] E. Griffiths, D. Dooley, M. Graham, G. Van Domselaar, F. S. L. Brinkman, and W. W. L. Hsiao. Context is everything: Harmonization of critical food microbiology descriptors and metadata for improved food safety and surveillance. *Frontiers in Microbiology*, 8:1068, 2017. ISSN 1664-302X. doi:10.3389/fmicb.2017.01068. URL <https://www.frontiersin.org/article/10.3389/fmicb.2017.01068>. → pages 106, 177
- [47] G. Grolemond and H. Wickham. A cognitive interpretation of data analysis: A cognitive interpretation of data analysis. *International Statistical Review*, 82(2):184–204, Aug 2014. doi:10.1111/insr.12028. → page 87
- [48] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 34(23): 4121–4123, Dec 2018. doi:10.1093/bioinformatics/bty407. → pages 4, 87, 122
- [49] M. Hahsler and M. Piekenbrock. DbSCAN: Density based clustering of applications with noise (DBSCAN) and related algorithms. <https://CRAN.R-project.org/package=dbscan>, 2017. → page 278
- [50] P. Hanrahan. VizQL: A language for query, analysis and visualization. In *ACM SIGMOD International Conference on Management of Data*, SIGMOD ’06, pages 721–721, New York, New York, 2006. ACM. ISBN 1-59593-434-0. doi:10.1145/1142473.1142560. → page 167

- [51] Harvard T.H. Chan School of Public Health. Distinctions Between Public Health and Clinical Medicine, 2016. URL <http://www.hsph.harvard.edu/about/public-health-medicine/>. Accessed: 2016-03-21. → page 40
- [52] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Conference on Human Factors in Computing Systems*, CHI '10, pages 203–212, 2010. ISBN 978-1-60558-929-9. doi:10.1145/1753326.1753357. → page 177
- [53] A. Z. Hettinger, E. M. Roth, and A. M. Bisantz. Cognitive Engineering and Health Informatics: Applications and Intersections. *Journal of Biomedical Informatics*, 67:21–33, Mar 2017. ISSN 1532-0480. doi:10.1016/j.jbi.2017.01.010. → page 52
- [54] J. Horsky, G. D. Schiff, D. Johnston, L. Mercincavage, D. Bell, and B. Middleton. Interface Design Principles for Usable Decision Support: a Targeted Review of Best Practices for Clinical Prescribing Interventions. *Journal of Biomedical Informatics*, 45(6):1202–16, Dec 2012. doi:10.1016/j.jbi.2012.09.002. → page 52
- [55] J. Huerta-Cepas, F. Serra, and P. Bork. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, 33(6):1635–1638, June 2016. ISSN 0737-4038, 1537-1719. doi:10.1093/molbev/msw046. → pages 4, 87
- [56] J. Hunter, D. Dale, E. Firing, and M. Droettboom. Matplotlib. <https://matplotlib.org/>, 2019. Accessed: 2019-03-25. → pages 3, 121
- [57] J. A. Jacko, editor. *The Human-computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. Human Factors and Ergonomics. CRC Press, Boca Raton, Florida, 3rd ed edition, 2012. ISBN 978-1-4398-2943-1. → pages 5, 94, 316
- [58] A. Kerren, K. Kucher, Y.-F. Li, and F. Schreiber. BioVis Explorer: A visual guide for biological data visualization techniques. *PLOS ONE*, 12(11):1–14, 2017. doi:10.1371/journal.pone.0187341. → pages 6, 88
- [59] S. Kovalchik. RISmed: Download content from ncbi databases. <https://CRAN.R-project.org/package=RISmed>, 2016. → page 277

- [60] J. H. Krijthe. Rtsne: T-distributed stochastic neighbor embedding using a barnes-hut. <https://github.com/jkrijthe/Rtsne>, 2015. → pages 92, 278
- [61] H. Lam. A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1149–1156, 2008. doi:10.1109/TVCG.2008.109. → page 113
- [62] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9): 1520–1536, 2012. doi:10.1109/TVCG.2011.279. → pages 25, 31
- [63] C. Larman and V. Basili. Iterative and Incremental Developments: A Brief History. *Computer*, 36(6):47–56, 2003. doi:10.1109/MC.2003.1204375. → page 29
- [64] K. O. Leslie and J. Rosai. Standardization of the Surgical Pathology Report: Formats, Templates, and Synoptic Reports. *Seminars in Diagnostic Pathology*, 11(4):253–7, Nov 1994. ISSN 0740-2570. → page 51
- [65] L. Liebenberg, N. Didkowsky, and M. Ungar. Analysing Image-based Data using Grounded Theory: the Negotiating Resilience Project. *Visual Studies*, 27(1):59–74, 2012. doi:10.1080/1472586X.2012.642958. → page 317
- [66] D. Lloyd and J. Dykes. Human-centered Approaches in Geovisualization Design: Investigating Multiple Methods through a Long-term Case Study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2498–2507, 2011. doi:10.1109/TVCG.2011.209. → pages 30, 37, 48, 57, 154, 155
- [67] N. J. Loman, C. Constantinidou, M. Christner, H. Rohde, J. Z.-M. Chan, J. Quick, J. C. Weir, C. Quince, G. P. Smith, J. R. Betley, M. Aepfelbacher, and M. J. Pallen. A Culture-independent Sequence-based Metagenomics Approach to the Investigation of an Outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *Journal of the American Medical Association*, 309(14):1502, 2013. doi:10.1001/jama.2013.3231. → page 50

- [68] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions of Graphics*, 5(2):110–141, Apr 1986. ISSN 0730-0301. doi:10.1145/22949.22950. → pages 147, 159, 176
- [69] J. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic Presentation for Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, Nov. 2007. doi:10.1109/TVCG.2007.70594. → pages 21, 110, 147, 176
- [70] J. A. Maxwell. *Qualitative Research Design: An Interactive Approach*, volume 41. SAGE Publications, Thousand Oaks, California, 2013. ISBN 0761926070. → page 317
- [71] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: Interactive visual exploration of system management time-series data. In *Conference on Human Factors in Computing Systems*, pages 1483–1492, New York, NY, USA, 2008. ACM. doi:10.1145/1357054.1357286. → page 38
- [72] I. Meirelles. *Design for Information: an Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport, Beverly, Mass, 2013. ISBN 978-1-59253-806-5. → page 104
- [73] M. Meyer, M. Sedlmair, and T. Munzner. The four-level nested model revisited: Blocks and guidelines. *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pages 11:1–11:6, 2012. doi:10.1145/2442576.2442587. → page 30
- [74] M. Meyer, M. Sedlmair, P. S. Quinan, and T. Munzner. The nested blocks and guidelines model. *Information Visualization*, 14(3): 234–249, 2015. doi:10.1177/1473871613510429. → page 30
- [75] P. Moore and C. Fitz. Using Gestalt Theory to Teach Document Design and Graphics. *Technical Communication Quarterly*, 2(4): 389–410, 1993. doi:10.1080/10572259309364549. → page 79
- [76] D. Moritz, C. Wang, G. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in Draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, Jan 2019.

ISSN 1077-2626. doi:10.1109/TVCG.2018.2865240. → pages 21, 110, 147, 159, 176, 177, 178

- [77] M. Muller, S. Guha, E. P. S. Baumer, D. Mimno, and N. S. Shami. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. *Proc. GROUP*, pages 0–6, 2016. doi:10.1145/2957276.2957280. → page 316
- [78] T. Munzner. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6): 921–928, Nov. 2009. doi:10.1109/TVCG.2009.111. → pages 30, 90
- [79] T. Munzner. *Visualization Analysis and Design*. A.K. Peters Visualization Series. CRC Press, Boca Raton, Florida, 2015. ISBN 978-1-4665-0891-0. → pages 88, 94, 104, 109, 317
- [80] V. Nikolayevskyy, K. Kranzer, S. Niemann, and F. Drobniowski. Whole Genome Sequencing of Mycobacterium tuberculosis for Detection of Recent Transmission and Tracing Outbreaks: A Systematic Review. *Tuberculosis*, 98:77–85, May 2016. doi:10.1016/j.tube.2016.02.009. → page 51
- [81] C. North. Toward Measuring Visualization Insight. *IEEE Transactions on Visualization and Computer Graphics*, 26(3):6–9, 2006. doi:10.1109/MCG.2006.70. → page 28
- [82] E. Nygren, J. C. Wyatt, and P. Wright. Helping Clinicians to Find Data and Avoid Delays. *Lancet*, 352(9138):1462–1466, 1998. doi:10.1016/S0140-6736(97)08307-4. → page 52
- [83] A. O’Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches. *Systematic Reviews*, 4(1), 2015. doi:10.1186/2046-4053-4-5. → page 82
- [84] J. Ooms. The jsonlite package: A practical and consistent mapping between json data and r objects. <https://arxiv.org/abs/1403.2805>, 2014. → page 277
- [85] L. J. Pankhurst, C. Del Ojo Elias, A. A. Votintseva, T. M. Walker, K. Cole, J. Davies, J. M. Fermont, D. M. Gascoyne-Binzi, T. A. Kohl,

- C. Kong, N. Lemaitre, S. Niemann, J. Paul, T. R. Rogers, E. Roycroft, E. G. Smith, P. Supply, P. Tang, M. H. Wilcox, S. Wordsworth, D. Wyllie, L. Xu, D. W. Crook, and COMPASS-TB Study Group. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *The Lancet Respiratory Medicine*, 4(1):49–58, Jan. 2016. doi:10.1016/S2213-2600(15)00466-X. → pages 51, 52, 87
- [86] M. Paprocki and B. Van de Ven. Bokeh. <https://bokeh.pydata.org>, 2019. Accessed: 2019-03-25. → page 121
- [87] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35: 526–528, 2018. → pages 4, 122
- [88] D. H. Parks, T. Mankowski, S. Zangooei, M. S. Porter, D. G. Armanini, D. J. Baird, M. G. I. Langille, and R. G. Beiko. GenGIS 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and an Extensible Plugin Framework. *PLOS ONE*, 8(7):1–10, 2013. doi:10.1371/journal.pone.0069885. → pages 4, 87
- [89] PHE. Tuberculosis in England: 2016, 2016. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/581238/TB_Annual_Report_2016_GTW2309_errata_v1.2.pdf. → page 53
- [90] Plotly. Plot.ly. <https://plot.ly/>, 2019. Accessed: 2019-03-25. → page 120
- [91] Z. Qu and J. Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 24(1): 468–477, 2018. doi:10.1109/TVCG.2017.2744198. → page 117
- [92] J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, N. Ouédraogo, B. Afrough, A. Bah, J. H. J. Baum, B. Becker-Ziaja, J. P. Boettcher, M. Cabeza-Cabrero, Á. Camino-Sánchez, L. L. Carter, J. Doerrbecker, T. Enkirch, I. García-Dorival, N. Hetzelt, J. Hinzmann, T. Holm, L. E. Kafetzopoulou, M. Koropogui,

A. Kosgey, E. Kuisma, C. H. Logue, A. Mazzearelli, S. Meisel, M. Mertens, J. Michel, D. Ngabo, K. Nitzsche, E. Pallasch, L. V. Patrono, J. Portmann, J. G. Repits, N. Y. Rickett, A. Sachse, K. Singethan, I. Vitoriano, R. L. Yemanaberhan, E. G. Zekeng, T. Racine, A. Bello, A. A. Sall, O. Faye, O. Faye, N. Magassouba, C. V. Williams, V. Amburgey, L. Winona, E. Davis, J. Gerlach, F. Washington, V. Monteil, M. Jourdain, M. Bererd, A. Camara, H. Somlare, A. Camara, M. Gerard, G. Bado, B. Baillet, D. Delaune, K. Y. Nebie, A. Diarra, Y. Savane, R. B. Pallawo, G. J. Gutierrez, N. Milhano, I. Roger, C. J. Williams, F. Yattara, K. Lewandowski, J. Taylor, P. Rachwal, D. J. Turner, G. Pollakis, J. A. Hiscox, D. A. Matthews, M. K. O'Shea, A. M. Johnston, D. Wilson, E. Hutley, E. Smit, A. Di Caro, R. Wölfel, K. Stoecker, E. Fleischmann, M. Gabriel, S. A. Weller, L. Koivogui, B. Diallo, S. Keïta, A. Rambaut, P. Formenty, S. Günther, and M. W. Carroll. Real-time, Portable Genome Sequencing for Ebola Surveillance. *Nature*, 530 (7589):228–32, 2016. doi:10.1038/nature16996. → page 87

- [93] P. S. Quinan and M. Meyer. Visually Comparing Weather Features in Forecasts. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):389–398, 2016. doi:10.1109/TVCG.2015.2467754. → page 25
- [94] J. Rathbone, T. Hoffmann, and P. Glasziou. Faster Title and Abstract Screening? Evaluating Abstrackr, a Semi-automated Online Screening Program for Systematic Reviewers. *Systematic Reviews*, 4 (1), 2015. ISSN 20464053. doi:10.1186/s13643-015-0067-6. → page 82
- [95] S. A. Renshaw, M. Mena-Allauca, M. Touriz, A. Renshaw, and E. W. Gould. The Impact of Template Format on the Completeness of Surgical Pathology Reports. *Archives of Pathology & Laboratory Medicine*, 138(1):121–4, Jan 2014. doi:10.5858/arpa.2012-0733-OA. → page 52
- [96] T. C. Rindfleisch and T. C. Rindfleisch. Privacy, Information Technology, and Health Care. *Communications of the ACM*, 40(8): 92–100, 1997. doi:10.1145/257874.257896. → pages 28, 38, 41
- [97] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative Genomics Viewer.

Nature Biotechnology, 29(1):24–26, 2011. ISSN 1087-0156.
doi:10.1038/nbt.1754. → page 4

- [98] C. Safran, M. Bloomrosen, W. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang, and D. E. Detmer. Toward a National Framework for the Secondary use of Health Data: an American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, 14(1):1–9, 2007.
doi:10.1197/jamia.M2273.Introduction. → page 41
- [99] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):659–668, Jan 2016. ISSN 1077-2626.
doi:10.1109/TVCG.2015.2467091. → page 120
- [100] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2017.
doi:10.1109/TVCG.2016.2599030. → pages 3, 113, 120, 124
- [101] H.-J. Schulz. Treevis.net: A Tree Visualization Reference. *IEEE Computer Graphics and Applications*, 31(6):11–15, Nov 2011.
doi:10.1109/MCG.2011.103. → pages 6, 88, 122, 157
- [102] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz. Information Visualization Evaluation in Large Companies : Challenges, Experiences, and Recommendations. *Information Visualization Journal*, 10(3):248—266, 2011. doi:10.1177/1473871611413099.
→ pages 27, 31
- [103] M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12): 2431–2440, dec 2012. doi:10.1109/TVCG.2012.213. → pages 5, 12, 30, 32, 35, 36, 53, 54, 148, 152, 154, 155
- [104] A. K. Shah and D. M. Oppenheimer. Heuristics Made Easy: an Effort-reduction Framework. *Psychology Bullitin*, 134(2):207–222, 2008. doi:10.1037/0033-2909.134.2.207. → page 79

- [105] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome research*, 13(11):2498–504, 2003. doi:10.1101/gr.1239303. → page 4
- [106] I. Shmelit, A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, A. O’Mara-Eves, M. P. Kelly, and J. Thomas. Pinpointing Needles in Giant Haystacks: Use of Text Mining to Reduce Impractical Screening Workload in Extremely Large Scoping Reviews. *Research Synthesis Methods*, 5(1):31–49, 2014. doi:10.1002/jrsm.1093. → page 82
- [107] B. Shneiderman. The Eyes Have It: a Task by Data Type Taxonomy for Information visualizations. *IEEE Symposium on Visual Languages*, pages 336–343, 1996. doi:10.1109/VL.1996.545307. → page 80
- [108] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pages 1–7, 2006. doi:10.1145/1168149.1168158. → pages 30, 40
- [109] J. Silge and D. Robinson. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, 1(3), 2016. doi:10.21105/joss.00037. URL <http://dx.doi.org/10.21105/joss.00037>. → pages 84, 278
- [110] H. A. Simon. *The Sciences of the Artificial(2nd ed.)*. MIT press, 1981. → page 25
- [111] H. Stitz, S. Gratzl, H. Piringer, T. Zichner, and M. Streit. KnowledgePearls: Provenance-based visualization retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 25(1): 120–130, 2018. doi:10.1109/TVCG.2018.2865024. → page 114
- [112] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1): 52–65, 2002. doi:10.1145/1400214.1400234. → page 113

- [113] M. Streit, H.-J. Schulz, A. Lex, D. Schmalstieg, and H. Schumann. Model-driven design for the visual analysis of heterogeneous data. *IEEE Transactions on Visualization and Computer Graphics*, 18(6): 998–1010, 2012. doi:10.1109/TVCG.2011.108. → page 122
- [114] P. N. Valenstein. Formatting Pathology Reports: Applying Four Design Principles to Improve Communication and Patient Safety. *Archives of Pathology & Laboratory Medicine*, 132(1):84–94, Jan 2008. doi:10.1043/1543-2165(2008)132[84:FPRAFD]2.0.CO;2. → page 52
- [115] L. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15:3221–3245, 2014. URL dl.acm.org/citation.cfm?id=2697068. → pages 84, 92, 278
- [116] C. J. Van Rijsbergen, S. E. Robertson, and M. F. Porter. *New Models in Probabilistic Information Retrieval*. British Library research & development report. Computer Laboratory, University of Cambridge, 1980. URL <https://books.google.com/books?id=WDZ3bwAACAAJ>. → page 278
- [117] J. VanderPlas, B. E. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, A. Satyanarayan, E. Lees, I. Timofeev, B. Welsh, and S. Sievert. Altair: Interactive statistical visualizations for python. *The Journal of Open Source Software*, 3(32), 2018. doi:10.21105/joss.01057. → page 121
- [118] K. J. Vicente. *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. CRC Press, 1999. → pages 25, 26, 32
- [119] K. Vredenburg, J.-Y. Mao, P. W. Smith, and T. Carey. A survey of user-centered design practice. *Conference on Human Factors in Computing Systems*, 4:471, 01 2002. doi:10.1145/503457.503460. → page 57
- [120] T. M. Walker, T. A. Kohl, S. V. Omar, J. Hedge, C. Del Ojo Elias, P. Bradley, Z. Iqbal, S. Feuerriegel, K. E. Niehaus, D. J. Wilson, D. A. Clifton, G. Kapatai, C. L. C. Ip, R. Bowden, F. A. Drobniewski, C. Allix-Béguec, C. Gaudin, J. Parkhill, R. Diel, P. Supply, D. W. Crook, E. G. Smith, A. S. Walker, N. Ismail, S. Niemann, T. E. A.

- Peto, J. Davies, C. Crichton, M. Acharya, L. Madrid-Marquez, D. Eyre, D. Wyllie, T. Golubchik, and M. Munang. Whole-genome Sequencing for Prediction of Mycobacterium tuberculosis Drug Susceptibility and Resistance: A Retrospective Cohort Study. *The Lancet Infectious Diseases*, 15(10):1193–1202, 2015. doi:10.1016/S1473-3099(15)00062-6. → page 51
- [121] M. Waskom. Seaborn. <https://seaborn.pydata.org/>, 2019. Accessed: 2019-03-25. → page 121
- [122] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-SNE effectively. *Distill*, 2016. doi:10.23915/distill.00002. URL <http://distill.pub/2016/misread-tsne>. → pages 84, 278, 279
- [123] H. Wickham. A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics*, 19(1):3–28, Jan. 2010. ISSN 1061-8600, 1537-2715. doi:10.1198/jcgs.2009.07098. → page 106
- [124] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, New York, 2016. ISBN 978-3-319-24277-4. URL <http://ggplot2.org>. → pages 3, 121, 124, 133
- [125] L. Wilkinson. *The Grammar of Graphics*. Statistics and Computing. Springer-Verlag, New York, New York, 2005. ISBN 978-0-387-24544-7. doi:10.1007/0-387-28695-0. → pages 106, 124
- [126] K. M. Winters, D. Lach, and J. B. Cushing. Considerations for characterizing domain problems. *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pages 16–22, 2014. doi:10.1145/2669557.2669573. → pages 30, 39
- [127] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Towards a general-purpose query language for visualization recommendation. In *Workshop on Human-In-the-Loop Data Analytics*, number 4 in HILDA ’16, pages 4:1–4:6. ACM, 2016. ISBN 978-1-4503-4207-0. doi:10.1145/2939502.2939506. → pages 167, 178
- [128] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, D. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization*

and *Computer Graphics*, 22(1):649–658, Jan 2016. ISSN 1077-2626. doi:10.1109/TVCG.2015.2467191. → pages 21, 176, 178

- [129] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *Conference on Human Factors in Computing Systems*, CHI ’17, pages 2648–2659, New York, NY, USA, 2017. ISBN 978-1-4503-4655-9. doi:10.1145/3025453.3025768. → pages 21, 147, 176, 178
- [130] J. Wood, R. Beecham, and J. Dykes. Moving beyond sequential design: Reflections on a rich multi-channel approach to data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2171–2180, Dec 2014. doi:10.1109/TVCG.2014.2346323. → page 153
- [131] S. H. Woolf. The Meaning of Translational Research and Why it Matters. *Journal of the American Medical Association*, 299(2): 211–213, 2008. doi:10.1001/jama.2007.26. → page 40
- [132] World Health Organization. Global tuberculosis report 2013, 2013. URL http://www.who.int/tb/publications/global_report/en/index.html. Accessed: 2016-03-21. → page 42
- [133] P. Wright, C. Jansen, and J. C. Wyatt. How to Limit Clinical Errors in Interpretation of Data. *Lancet*, 352(9139):1539–1543, 1998. doi:10.1016/S0140-6736(98)08308-1. → page 52
- [134] L. Yeganova, W. Kim, S. Kim, and W. J. Wilbur. Retro: Concept-based Clustering of Biomedical TopicalSsets. *Bioinformatics*, 30(22):3240–8, nov 2014. ISSN 1367-4811. doi:10.1093/bioinformatics/btu514. URL <https://bit.ly/2Jpn0nz>. → page 82
- [135] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017. doi:10.1111/2041-210X.12628. → pages 4, 122
- [136] D. A. Zipkin, C. A. Umscheid, N. L. Keating, E. Allen, K. Aung, R. Beyth, S. Kaatz, D. M. Mann, J. B. Sussman, D. Korenstein,

C. Schardt, A. Nagi, R. Sloane, and D. A. Feldstein. Evidence-based Risk Communication: a Systematic Review. *Annals of Internal Medicine*, 161(4):270–80, Aug 2014. ISSN 1539-3704. doi:10.7326/M14-0295. → page 52

Appendix A

Evidence Based Design Supplemental Materials

Contents

1. Supplemental Figure S1
2. Supplemental Tables S1-S6
3. Final Design Walkthrough
4. Survey Instrument 1: Task and Data Questionnaire
5. Survey Instrument 2: Design Choices Questionnaire

A.1 Supplemental Figures

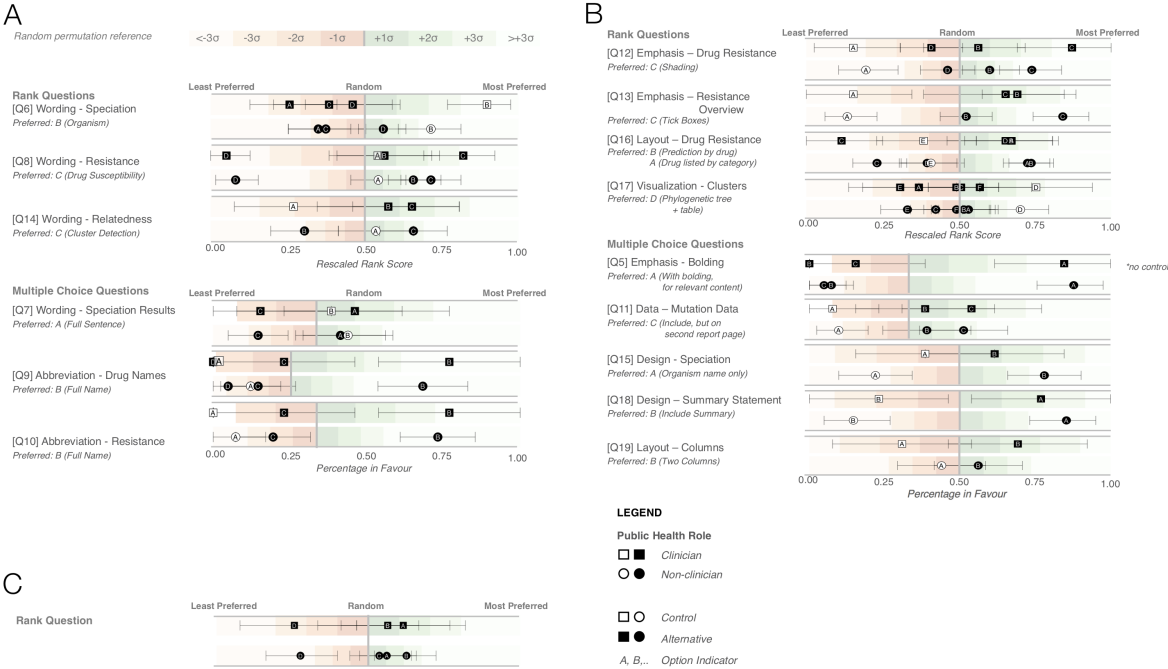


Figure A.1: Survey responses with confidence intervals. (a) Wording choices (b) Design choices (c) Full reports

A.2 Supplemental Tables

Table A.1: Task and data questionnaire study participants.

Note: Participants could select one or more levels of training, thus, rows will *not* add to 100%. * By professional experience, we mean collaborating with others on a project. ** By continuing education, we mean attending workshops, training sessions, or self-directed learning

Subject Area	Level Training				
	None	Undergraduate	Graduate (Masters, PhD); Medical Training	Professional Experience*	Continuing Education**
Molecular Biology or Biochemistry	29.4%	29.4%	47.1%	41.2%	35.3%
Epidemiology	11.8%	5.9%	58.5%	64.7%	41.2%
Biostatistics	58.8%	11.8%	29.4%	23.5%	23.5%
Bioinformatics	52.9%	0.0%	11.8%	35.3%	29.4%
Genomics	23.5%	5.9%	23.5%	47.1%	52.0%
Infectious Disease	5.9%	35.3%	58.8%	76.5%	52.9%
Respiratory Medicine	17.4%	1.4%	29.4%	47.1%	29.4%

Table A.2: Task and Data Questionnaire respondents anticipated future use of molecular/genomic data

Data Type	Extent of usage					
	Never	Rarely	Sometimes	Often	All of the time	Don't know what this is
Patient Information	1 (5.9%)	0 (0.0%)	1 (5.9%)	1 (5.9%)	14 (82.4%)	0 (0.0%)
Patient's own prior TB test result	0 (0.0%)	0 (0.0%)	3 (17.6%)	1 (5.9%)	12 (70.6%)	1 (5.9%)
Requester Identifier	2 (11.8%)	2 (11.8%)	2 (11.8%)	2 (11.8%)	9 (52.9%)	0 (0.0%)
Review identifier	2 (11.8%)	2 (11.8%)	4(23.5%)	0 (0.0%)	8 (47.1%)	1 (5.9%)
Type of sample	0 (0.0%)	0 (0.0%)	1 (5.9%)	5 (24.9%)	11 (64.7%)	0 (0.0%)
Sample collection site	0 (0.0%)	2 (11.8%)	0 (0.0%)	1 (5.9%)	11 (64.7%)	0 (0.0%)
Sample Collection date	0 (0.0%)	0 (0.0%)	2 (11.8%)	2 (11.8%)	13 (76.5%)	0 (0.0%)
Interpretation or comments from reviewer	3 (17.6%)	2 (11.8%)	2 (11.8%)	1 (5.9%)	11 (64.7%)	0 (0.0%)
Tuberculin Skin Test Results	4 (23.5%)	2 (11.8%)	2 (11.8%)	2 (11.8%)	7 (41.2%)	0 (0.0%)
Interferon Gamma Release Assay (IGRA) results	3 (17.6%)	2 (11.8%)	1 (5.9%)	4 (23.5%)	7 (41.2%)	0 (0.0%)
Chest X-ray	3 (17.6%)	2 (11.8%)	0 (0.0%)	3 (17.6%)	9 (52.9%)	0 (0.0%)
Acid Fast Bacilli (AFB) smear status	2 (11.8%)	1 (5.9%)	1 (5.9%)	1 (5.9%)	12 (70.6%)	0 (0.0%)
Culture results	1 (5.9%)	0 (0.0%)	0 (0.0%)	2 (11.8%)	14 (82.4%)	0 (0.0%)
Speciation	0 (0.0%)	0 (0.0%)	1 (5.9%)	0 (0.0%)	16 (94.1%)	0 (0.0%)
Phenotypic Drug susceptibility test (determined by culture)	0 (0.0%)	0 (0.0%)	1 (5.9%)	1 (5.9%)	15 (88.2%)	0 (0.0%)
Molecular Drug susceptibility testing (determine by PCR or Line Probe Assay)	0 (0.0%)	0 (0.0%)	1 (5.9%)	4 (23.5%)	12 (70.6%)	0 (0.0%)
Specific mutations conferring drug resistance (resistotype)	1 (5.9%)	0 (0.0%)	1 (5.9%)	5 (24.9%)	9 (52.9%)	1 (5.9%)
Spoligotype	3 (17.6%)	3 (17.6%)	1 (5.9%)	3 (17.6%)	2 (11.8%)	5 (29.4%)

Table A.2 continued from previous page

	Extent of usage					
MIRU-VNTR	0 (0.0%)	1 (5.9%)	1 (5.9%)	4 (23.5%)	11 (64.7%)	0 (0.0%)
RFLP	3 (17.6%)	6 (35.3%)	1 (5.9%)	2 (11.8%)	1 (5.9%)	4 (23.5%)
Cluster Assignment	0 (0.0%)	2 (11.8%)	2 (11.8%)	1 (5.9%)	12 (70.6%)	0 (0.0%)
SNP/V distance from other isoaltes	1 (5.9%)	3 (17.6%)	1 (5.9%)	2 (11.8%)	9 (52.9%)	1 (5.9%)
Phylogenetic Tree	1 (5.9%)	2 (11.8%)	3 (17.6%)	2 (11.8%)	6 (25.3%)	3 (17.6%)
Laboratory performance measures	2 (11.8%)	3 (17.6%)	1 (5.9%)	5 (24.9%)	5 (29.4%)	1 (5.9%)

Table A.3: Task and Data Questionnaire respondents' confidence in their ability to interpret various types of laboratory data.

	Confidence Interpreting Information				
Data Type	Confident	Somewhat Confident	Not Confident	Don't know what this is	Total Confident
MIRU-VNTR	64.7%	29.4%	5.9%	0.0%	94.1%
Phenotypic drug susceptibility testing from culture	58.8%	23.5%	11.8%	5.9%	82.3%
Molecular drug susceptibitily testing from PCR or LPA	58.8%	23.5%	11.8%	5.9%	82.3%
Genomic Clusters	52.9%	29.4%	11.8%	5.9%	82.3%
SNPS (mutations)	47.1%	35.2%	11.8%	5.9%	82.3%
SNP/V conferring drug resistance	41.2%	29.4%	23.5%	5.9%	70.6%
Genetic distance between isolates measure by SNP/V	35.3%	41.2%	17.6%	5.9%	76.5%
Phylogenetic Tree	35.4%	29.4%	17.6%	17.6%	64.8%

Table A.3 continued from previous page

	Confidence Interpreting Information				
Percentage of Genome Covered	29.4%	29.4%	35.3%	5.9%	58.8%
Genome Sequencing quality metrics	29.4%	29.4%	29.4%	11.8%	58.8%
Number of reads mapped /unmapped	29.4%	29.4%	29.4%	11.8%	58.8%
Depth fo sequencing coverage	29.4%	29.4%	29.4%	11.8%	58.8%
RFLP	29.4%	5.9%	35.3%	29.4%	35.3%
Soligotyping	23.5%	11.8%	23.5%	41.2%	35.3%

Table A.4: Task and Data Questionnaire respondents' confidence in the ability of genomic data to perform various laboratory tasks

Data Types	Task Type	Level of Confidence			
		It can do this	It may be able to do this	It can't do this	Don't know what this is
Organism/Speciation	Diagnosis	76.5%	17.9%	5.4%	0.0%
Diagnose Active TB		29.4%	23.5%	47.1%	0.0%
Predict Drug Susceptibility	Treatment	52.9%	47.1%	0.0%	0.0%
Inform a physician's choice of a therapy regimen		35.3%	64.7%	0.0%	0.0%
Monitor Treatment progress		5.9%	47.1%	41.2%	5.9%
Identify epidemiologically related patients	Surveillance	58.8%	41.2%	0.0%	0.0%
Identify transmission events		41.2%	52.9%	5.9%	0.0%
Rule out transmission events		64.7%	29.4%	5.9%	0.0%
Assign patient to existing TB cluster		70.0%	29.4%	0.0%	0.0%

Table A.5: Task and Data Questionnaire respondents identification of laboratory-associated barriers impacting their work-flows

	Diagnosis	Treatment	Surveillance*
Response	<i>Respondents = 6</i>		<i>Respondents = 5</i>
No Issues	0 (0.0%)	0 (0.0%)	NA
Additional data is needed	0 (0.0%)	2 (33.3%)	3 (60.0%)
Issue with timeliness of results being provided (too slow)	5 (83.3%)	5 (83.3%)	NA
Results provided over multiple unconnected documents	5 (83.3%)	5 (83.3%)	NA
Difficultly interpreting lab results	2 (33.3%)	3 (50.0%)	4 (80.0%)
Lab data is not routinely provided	0 (0.0%)	1 (16.7%)	3 (60.0%)
Lab data is not linked to patient data	1 (16.7%)	3 (50.0%)	1 (20.0%)
Other	2 (33.3%)	1 (16.7%)	NA

Table A.6: Summary of questions asked in the Design Choice Questionnaire, including preferred response.

Question #	Options	Participant Preference	Classification	Question Type
1 to 4	NA	NA	Demographic	NA
5	A - With Bolding B - Without Bolding C - Equally Informative	A - With Bolding	Design	Multiple Choice
6	A- Speciation B - Organism (Control) C - Diagnosis D - Species	B - Organism (Control)	Wording	Rank
7	A - Full Sentence B - Summary	A - Full Sentence	Wording	Rank
8	A - Drug Resistance (Control) B - Drug Sensitivity C - Drug Susceptibility D - Treatment	C - Drug Susceptibility	Wording	Rank
9	A - 3 letter abbreviation (Ex. INH) (Control) B - Full Name (Ex. Isoniazid) C - Show me everything (Ex. Isonizaid (INH,H)) D - The are equally informative	B - Full Name	Wording	Multiple Choice
10	A - 1 letter abbreviation (Ex. S,R,U) (Control) B - Full Name (Ex. Susceptible, Resistant, Unknown) C - They are equally informative	B - Full Name	Wording	Multiple Choice
11A	A - No, I am not interested in mutation data B - Yes on the same table with drug susceptibility data (Control) C - Yes, but on the other side of the report	C - Yes, but on the other side of the report	Design	Multiple Choice

Table A.6 continued from previous page

Question #	Options	Participant Preference	Classification	Question Type
12	A - Basic (Control) B - Alert Glyphs C - Shaded D - Bolded	D - Shaded	Design	Rank
13	A - Basic (Control) B - Summary Sentence C - Tick Boxes	C - Tick boxes	Design	Rank
14	A - Relatedness (Control) B - Epidemiology C - Cluster Detection	C - Cluster Detection	Wording	Rank
15	A - Percent Match (Control) B - Organism Name	B - Organism Name	Design	Multiple Choice
16	A - Drugs listed by category B - Prediction by drug C- Summary Sentence D - Drugs listed by category bin E - Abbreviated prediction by drug (Control)	A - Drugs listed by category B - Prediction by drug	Design	Rank
17	A - # of cases with spark line B - # of isolates related table C- Table + Graph of # of isolates by SNP distance D - Table + Phylogenetic Tree E- Related isolates with SNP difference details F - Summary with related isolates per year	D - Table + Phylogenetic Tree	Design	Rank

Table A.6 continued from previous page

Question #	Options	Participant Preference	Classification	Question Type
18	A - Summary Statement B - No summary statement	A - Summary Statement	Design	Rank
19	A - One column B- Two column	B - Two column	Design	Rank
21 to 23	NA	NA	Full Report	Likert
24	A- Dark Heading B- Gray Heading C- Light Heading D- Pictures		Full Report	Rank

A.3 Justification for Final Design Choices by Section

A.3.1 Analysis of Quantitative and Qualitative Results

Shorthand for the different surveys / requirements documents

Abbreviation	Examples
EC: Expert Consults	EC-1 = Expert consult #1
S1: Survey 1 (task survey)	S1-Q10 = Survey 1 question 10
S2: Survey 2 (design survey)	S2-Q11A = Survey 2 question 11A
ISO: ISO15189 requirements document	S2-SR18 = Survey 2 survey respondent 18 (for text answers)

1. Summary Statement

- (a) On first page of report
- (b) Summary sentence
- (c) Bold important terms

2. Organism

- (a) On first page of report
- (b) Section title is Organism (supported by S2-Q6. 31/54 of respondents prefer “Organism” as top choice (42/54 preferred it as one of their top two choices). Many participants (13/54) ranked “Diagnosis” the first choice, over “species” and “speciation”, however, however this trend was driven mainly by non-clinicians (11 non-clinicians ranking diagnosis as their first choice, and only 2 clinicians ranking it as their first choice). In fact, clinicians consistently ranked “Diagnosis” much lower.
- (c) Summary sentence with bolding to emphasize findings

3. Drug Susceptibility: in general, there was not a clear and obvious dislike of the control design (S2-Q16 “Abbreviated prediction by drug”) because it was not consistently ranked as lowest preference, but it was not the most desirable

choice for respondents. Clinicians tended to rank the control design as the lowest preference relative to non-clinicians.

- (a) On first page of report
- (b) Section title is Drug Susceptibility (supported by S2-Q8. Respondents (27/54) preferred “Drug Susceptibility” as their first choice and 41/54 preferred it as one of their top two choices, but other options also selected (Drug Resistance, Drug Sensitivity). Anecdotal and also qualitative evidence indicated that the title predicted drug resistance still controversial.
- (c) Summary sentence to state in silico prediction (not phenotypic)
- (d) Tick boxes (S2-Q13 to indicate mono, multi, or extensive drug resistance (supported by 38/54 who rated tick boxes as preferred choice, and majority rate basic (control report design) as least preferred (43/54). Good comment support for tick boxes too: S2-R5: “[..] Tick box is the most straightforward way [..] summary sentence [..]likely will be ignored”; S2-R23: “the less risk of misinterpretation of test data the better”. There was some different between clinician and non-clinician preferences, but we opted to use the tick boxes with additional annotations to more clearly indicate when no resistance was detected.
- (e) Table listing predictions for drug susceptibility (supported by responses for S2-Q16. Many respondents felt that an organized table/bins would be the best, and when including the resistance information the table was the easiest choice.)
 - i. Categorize drugs by class
 - ii. Categorize drugs by susceptible or resistant using full term (S2-Q16 top choices were to “list prediction by drug” (21/54) and also to “list prediction by category” (17/54). The design choices offered didnt quite do both, but the final design does. It categories drugs according to first and second line (not test on S2) and then

by Sensitive / Resistant and finally lists each drug line by line.)

iii. Full name (no abbreviation) for drugs

iv. Highlight resistant drugs by shading (supported by S2-Q12 where majority preferred shading (33/54) over other options. Clear that basic (no emphasis on resistance) least preferred (36/54 ranked it last). Number of comments were made for showing resistance: S2-SR3 “report must call attention to drug resistance”; S2-R18 “MDR-TB should be flagged”, S2-R11 “best highlights the MDR-TB”, S2-SR16 “better to highlight what is working instead of what is not working”, S2-SR24 “Bold gets confused with column headers”) Indicate resistance prediction source (see 4. Resistance Information)

4. Resistance Information: Only 5/54 participants didn't want to see any genomic mutation information at all, but participants were split as to how this information should be prioritized. 28/54 wanted to see this information on the second page (not front of mind) while 21/54 wanted to see this information on the front page. In the end, we put this information on the front page because it worked well with the design (see rationale in main paper), but we reduced the amount of genomic information shown so as not to overwhelm the reader.

(a) Incorporated into Drug Susceptibility table

(b) Column header: Resistance (Mutation)

(c) Resistance indicated by Gene (Amino Acid Change) or “No mutation detected”. (S2-Q11. 46/54 wanted gene abbreviation (i.e. katG) info included when resistance is detected. But participants were less enthusiastic about addition information. A total of 25/54 participants wanted to see base pair changes, 27/54 wanted to see amino acid changes, and (this is a bit odd) 29/54 wanted to see read support for a mutation (but not the total number of reads sequenced (wanted by only 14/54 participants)). We chose to show the amino acid change. Other data

suggest clinicians in particular do want to see this kind of laboratory data (see 7. Laboratory Quality Data).

5. Cluster Detection: concerns raised about the relevance of this information at all: S2-SR18 “Cluster detection would only be fine for those who already know what a cluster is”, S2-SR9 “Not sure what this conveys [...] What is the clinical action?”

- (a) On second page of report

- (b) Section title is Cluster Detection (supported by S2-Q14. All respondents ranked “cluster detection” as top choice (25/54) or top two choices (46/54), compared to 18/54 ranking the control design (“Relatedness”) first, or 36/54 ranking it among their top two choices. Also “cluster detection” or “epidemiology” was the most preferred by clinicians, while “relatedness” was the least preferred. Support also from comments: S2-SR23 “When I see this I think epidemiology and clusters; not relatedness”, S2-SR11 “Cluster detection is important clinically and epidemiologically.”)

- (c) Table with phylogenetic tree (control option preferred)

6. Laboratory Quality Data: concerns raised about the relevance of this information at all: S2-SR18 “Cluster detection would only be fine for those who already know what a cluster is”, S2-SR9 “Not sure what this conveys [...] What is the clinical action?”

7. Laboratory Quality Data

- (a) Do not include laboratory (sample & sequence) QC data on report (Compared to the original report, this report does not have the laboratory technical details (i.e. percent mapping to reference, genome coverage, reference genome information etc.) because this was deemed not necessary information for any of the tasks that stakeholders (but especially clinicians) used to conduct their activities (S1). Including laboratory technical data considered harmful (“Why would the lab put

out poor quality results for me to interpret?”, “Isn't that up to the lab?” (EC)). This doesn't mean the data isn't collected and stored but that the data isn't presented on the clinical report. It can be moved to the second page of the report if necessary, but should not be featured on the front page.

A.3.2 ISO15189 Requirements

BSI Standards - BS EN ISO 15189:2012 Medical Laboratories - Requirements for quality and competence.

5.8 Reporting of Results

- 5.8.1 General

- The results of each examination shall be reported accurately, clearly, unambiguously and in accordance with any specific instructions in the examination procedures.
- The laboratory shall define the format and medium of the report (i.e. electronic or paper) and the manner in which it is to be communicated from the laboratory.
- The laboratory shall have a procedure to ensure the correctness of transcription of laboratory results.
- Reports shall include the information necessary for the interpretation of the examination results.
- The laboratory shall have a process for notifying the requester when an examination is delayed that could compromise patient care.

- 5.8.2 Report attributes

- The laboratory shall ensure that the following report attributes effectively communicate laboratory results and meet the users needs:

- * Comments on sample quality that might compromise examination results
 - * Comments regarding sample suitability with respect to acceptance/rejection criteria
 - * Critical results, where applicable
 - * Interpretive comments on results, where applicable, which may include the verification of the interpretation of automatically selected and reported results (see 5.9.1) in the final report.
- 5.8.3 Report content
 - The report shall include, but not be limited to, the following:
 - * A clear, unambiguous identification of the examination including, where appropriate, the examination procedure; the identification of the laboratory that issued the report; Will this be Oxford or Birmingham?
 - * Identification of all examinations that have been performed by a referral laboratory
 - * Patient identification and patient location on each page
 - * Name or other unique identifier of the requester and the requesters contact details
 - * Date of primary sample collection (and time, when available and relevant to patient care)
 - * Type of primary sample
 - * Measurement procedure, where appropriate
 - * Examination results reported in SI units, units traceable to SI units, or other applicable units

- * Biological reference intervals, clinical decision values, or diagrams/nomograms supporting clinical decision values, where applicable;
- * Other comments such as cautionary or explanatory notes (e.g. quality or adequacy of the primary sample which may have compromised the result, results/interpretations from referral laboratories, use of developmental procedure
- * Identification of examinations undertaken as part of a research or development program and for which no specific claims on measurement performance are available
- * Identification of the person(s) reviewing the results and authorizing the release of the report (if not contained in the report, readily available when needed)
- * Date of the report, and time of release (if not contained in the report, readily available when needed); o page number to total number of pages (e.g. Page 1 of 5, Page 2 of 5, etc.).

•

5.9 Reporting of Results

• 5.9.1 General

- The laboratory shall establish documented procedures for the release of examination results, including details of who may release results and to whom. The procedures shall ensure that the following conditions are met.
- When the quality of the primary sample received is unsuitable for examination, or could have compromised the result, this is indicated in the report.
- When examination results fall within established alert or critical

intervals:

- * a physician (or other authorized health professional) is notified immediately [this includes results received on samples sent to referral laboratories for examination (see 4.5)];
 - * records are maintained of actions taken that document date, time, responsible laboratory staff member, person notified and examination results conveyed, and any difficulties encountered in notifications.
 - Results are legible, without mistakes in transcription, and reported to persons authorized to receive and use the information.
 - When results are transmitted as an interim report, the final report is always forwarded to the requester.
 - There are processes for ensuring that results distributed by telephone or electronic means reach only authorized recipients. Results provided orally shall be followed by a written report. There shall be a record of all oral results provided.
 - * NOTE 1 For the results of some examinations (e.g. certain genetic or infectious disease examinations) special counselling may be needed. The laboratory should endeavour to see that results with serious implications are not communicated directly to the patient without the opportunity for adequate counselling.
 - * NOTE 2 Results of laboratory examinations that have been separated from all patient identification may be used for such purposes as epidemiology, demography or other statistical analyses.
- 5.9.2 Automated selection and reporting of results
 - If the laboratory implements a system for automated selection and reporting of results, it shall establish a documented procedure to ensure that:

- * The criteria for automated selection and reporting are defined, approved, readily available and understood by the staff.
 - NOTE Items for consideration when implementing automated selection and reporting include changes from previous patient values that require review and values that require intervention by laboratory personnel, such as absurd, unlikely or critical values.
- * There is a process for indicating the presence of sample interferences (e.g. haemolysis, icterus, lipaemia) that may alter the results of the examination;
- * There is a process for incorporating analytical warning messages from the instruments into the automated selection and reporting criteria, when appropriate;
- * Results selected for automated reporting shall be identifiable at the time of review before release and include date and time of selection;
- * There is a process for rapid suspension of automated selection and reporting.
- * When an original report is revised there shall be written instructions regarding the revision so that:
 - The revised report is clearly identified as a revision and includes reference to the date and patients identity in the original report
 - The user is made aware of the revision
 - The revised record shows the time and date of the change and the name of the person responsible for the change
 - The original report entries remain in the record when revisions are made

- Results that have been made available for clinical decision making and revised shall be retained in subsequent cumulative reports and clearly identified as having been revised
- When the reporting system cannot capture amendments, changes or alterations, a record of such shall be kept.

A.4 Task and Data Questionnaire Online Survey

Questionnaire submitted online for the task and data survey



COMPASS-TB Report Design Questionnaire

Page 1

Description and Consent

Many public health agencies are starting to use whole genome sequencing (reading every letter of an organism's DNA) as a tool for diagnosing infections, predicting what antibiotics an organism is sensitive or resistant to, and identifying closely related isolates that might suggest an outbreak. Last year, [a study in The Lancet Infectious Diseases](#), showed that when this technique is used in the tuberculosis laboratory, we can generate all the usual results that one has come to expect from a reference mycobacteriology lab, but we can do so much faster and at lower cost. As a result of this study, groups like Public Health England, the BC Centre for Disease Control, and the US Centers for Disease Control and Prevention are all using genomics to analyze their incoming mycobacterial isolates.

Sequencing a bacterial genome generates a lot of information, only some of which might be needed to manage a patient's infection. We are interested in designing a new lab report form that will help to communicate tuberculosis genomic data in a clear, concise, and meaningful way that will help those in the tuberculosis community - clinicians, epidemiologists, laboratory scientists, and more - in their daily work. There is a large field of research into how to present data in a way that makes it easily interpretable - we will be using principles from this field in designing our new report format, which will be shared with public health laboratories so that they may choose to use it in their own reporting.

By participating in this survey, you will help us better understand how you use lab data in your daily tuberculosis-related work. The answers from this survey will help us to design a series of sample reports, which we will test later in the year through a second survey.

Today's survey is divided into several parts. We'd like everyone to complete Parts I and II, which ask questions about your job and your familiarity with concepts and data types. Part III, on tasks related to diagnosis and treatment, will only be asked to physicians/clinicians. Part IV, on contact tracing and outbreak management, will be asked of all participants. Part V, on surveillance, will only be asked of epidemiologists, surveillance analysts, and researchers. All participants will be asked for (optional) email contact information in Part VI.

Consent for Participation**STUDY PROCEDURES:**

If you agree to voluntarily participate in this research, your participation will include the following online survey (estimated completion time 15-30 minutes) in which you will be asked questions about how you use TB laboratory data in your work. At the end of the survey, you may choose to provide an email address if you'd like to be entered into a draw for an Apple Store gift card, or receive the final results of the study.

There are no known or anticipated risks to you by participating in this research. An optional benefit is receiving the results of the study via an emailed report at the project's conclusion, which will include a template for the final report design that participants may use in their own work. Study results will be also shared with the research community through open-access publications, conference reports, tweets and other social media postings.

MEASURES TO MAINTAIN CONFIDENTIALITY

Data from this study will be coded anonymously: a unique anonymous identifier will be used in place of the optional email addresses, which will be saved separately for the purposes of the gift card draw and sending information about the final report to participants. After analysis, the anonymized data will be saved in electronic format and made publicly available online for use by the research community.

CONTACTS FOR COMPLAINTS OR CONCERNS

Geoff McKee is a resident physician in Public Health and Preventive Medicine at the University of British Columbia and you may contact him if you have any further questions by email at

If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics

Taking part in this study is entirely up to you. You have the right to refuse to participate in this study. If you decide to take part, you may choose to pull out of the study at any time without giving a reason.

By completing the questionnaire, you are consenting to participate in this research.

PRINCIPAL INVESTIGATOR:

Jennifer Gardy, School of Population & Public Health

PART I – OCCUPATION AND SUBJECT AREA KNOWLEDGE QUESTIONS

All participants are asked to complete this first part of the survey: we'd like to find out more about you, your background, and your general attitude towards genomics in public health.

1. What is your role in tuberculosis diagnosis, treatment, management, and/or surveillance? You may select more than one role.

[Select as many as apply]

- ☐ Clinical management - I work directly with TB patients, providing care and/or case management
- ☐ Laboratory work – I work in a mycobacteriology laboratory setting where I am involved with lab testing for TB
- ☐ Surveillance/epidemiology - I work with TB data to understand patterns in disease occurrence
- ☐ Research - I carry out academic research into TB
- ☐ Other, please specify...

Type here

What is your clinical role?

[Select one option]

- ☐ Physician/Clinician
- ☐ Nurse
- ☐ Other, please specify...

Type here

2. Who is your primary employer?

[Select as many as apply]

- ☐ Public Health Organization - e.g. Public Health England, CDC
- ☐ Private Clinic/Primary Care - e.g. a doctor's office
- ☐ Hospital
- ☐ Academic Institution
- ☐ Other, please specify...

Type here

3. In what country do you work?

[Select one option]

- ☐ England
- ☐ Canada
- ☐ USA
- ☐ Other, please specify...

Type here

4. How many years of experience do you have working in the field of tuberculosis?

[Number of years]

Type here

5. Please indicate the highest level of training (if any) you have in the following subject areas:

* By professional experience, we mean collaborating with others on a project
** By continuing education, we mean attending workshops, training sessions, or self-directed learning

	None	Undergraduate	Graduate Masters, PhD, Medical Training	Professional Experience*	Continuing Education**
Molecular Biology or Biochemistry	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Epidemiology	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Biostatistics	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Bioinformatics	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Genomics	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Infectious Diseases	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Respiratory Medicine	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

6. Have you ever heard of or been involved in a research project that used whole genome sequencing data to diagnose or characterize tuberculosis infections or understand tuberculosis epidemiology?

[Select one option]

- ☐ Yes - I have heard about these sorts of studies but have not been involved in one
- ☐ Yes - I have worked on one of these studies
- ☐ No - I am not familiar with TB genomics studies

7. How enthusiastic are you about public health agencies using genome sequencing to understand and diagnose infectious diseases?

[Select one option]

- ☐ Very enthusiastic – we should be using genomics now
- ☐ Enthusiastic – genomics has a lot of potential, but still needs to be validated for clinical use
- ☐ Neutral - I don't have a strong opinion on genomics in public health
- ☐ Skeptical – genomics may be useful, but there is no clear application
- ☐ It's all hype – genomics hasn't proven itself to be more useful than the techniques we currently use

PART II – FAMILIARITY WITH DATA TYPES

All participants are asked to complete this second part of the survey: we'd like to hear about the many types of TB laboratory data you might encounter in your work.

8. How frequently do you foresee yourself using the following data types in your future, routine work?

[Select one option per data type]

	Never	Rarely	Sometimes	Often	All the time	I Don't Know What This Is
Patient identifiers (Name, age, location)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Patient's own prior tuberculosis test results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Requester identifiers (Name, contact, copy to etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reviewer identifiers (Name, position etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Type of sample (Sputum, fine needle aspirate etc)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sample collection site (lymph node, peripheral blood draw etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sample collection date	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interpretation or comments from reviewer	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Tuberculin Skin Test Results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interferon Gamma Release Assay (IGRA) results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Chest X-ray results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Acid Fast Bacilli (AFB) Smear results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Culture results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Speciation (M. tuberculosis, MAC, M. bovis etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Phenotypic drug susceptibility testing - determined by culture	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Molecular drug susceptibility testing - determined by PCR or Line Probe Assay (LPA)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Specific mutations conferring drug resistance (Resistotype)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Spoligotype	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
MIRU-VNTR	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Restriction fragment length polymorphisms (RFLP)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cluster Assignment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Single Nucleotide Polymorphism/Variant distance from other isolates	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Phylogenetic Tree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Laboratory performance measures (Sequence quality, coverage etc.)

Never Rarely Sometimes Often All the time I Don't Know What this is

9. How would you describe your ability to interpret the following data?

To help you choose your answers, we suggest the following scheme:

- *Don't know what it is:* you are unaware of this data type
- *Not confident:* you know what these data are, but you are not certain how to interpret the data for clinical management, surveillance, or research.
- *Somewhat confident:* you know what these data are and are capable of interpreting it, but you usually seek out a confirmation for your interpretation
- *Confident:* you understand how to interpret this data and are confident in using it in your practice

	Don't know what this is	Not Confident	Somewhat Confident	Confident
Spoilgotyping	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
RFLP	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
MIRU-VNTR	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Single Nucleotide Polymorphisms (mutations)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Phenotypic Drug Susceptibility Testing from culture	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Molecular Drug Susceptibility Testing from PCR or LPA	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Single nucleotide polymorphisms/variants (mutations) conferring drug resistance	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Phylogenetic Tree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Genetic distance between cases measured in Single Nucleotide Polymorphisms/Variants (mutations)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Genomic Clusters	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Genome sequencing quality metrics	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Number of reads mapped/unmapped	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Percentage of Genome Covered	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Depth of sequencing coverage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

10. How confident are you that genomic data can be used to correctly perform the following tasks?

	Don't know what this is	It can't do this	It may be able to do this	It can do this
Organism Speciation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Diagnose active tuberculosis	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Predict Drug Susceptibility	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Inform a physician's choice of a therapeutic regimen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Monitor treatment progress	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Identify epidemiologically related patients	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	Don't know what this is	It can't do this	It may be able to do this	It can do this
Identify transmission events	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rule out transmission events	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Assign patient to existing tuberculosis cluster	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

PART III – TASKS RELATED TO DIAGNOSIS & TREATMENT

Only physicians/clinicians are asked to complete this part: our initial assessment indicated that only clinicians are involved in diagnosis and treatment, these questions should not be answered by nurses, researchers, epidemiologists, or biostatisticians as they are not directly involved in diagnosis and treatment.

11. Are you involved in the diagnosis and treatment of tuberculosis?

Yes No

12. What types of samples do you requisition or send to the laboratory?

[Select as many as apply]

- ☐ Sputum
- ☐ Bronchoscopy Wash
- ☐ Fine Needle Aspirate
- ☐ Biopsy
- ☐ Urine
- ☐ Other, please specify...

Type here

13. Do you want to know any laboratory or bioinformatics quality metrics that may be associated with that data being reported to you?

[Select one option]

- ☐ Yes – I want to always want to have data quality metrics
- ☐ No – Data quality results are not relevant, the lab would not release low quality data and I trust their processes
- ☐ I don't know
- ☐ Other, please specify...

Type here

14. In what format do you currently receive this data?

[Select as many as apply]

- ☐ Physical report mailed or faxed to me (hard copy)
- ☐ PDF report in electronic health record system (soft copy)
- ☐ Extracted data in electronic health record system (soft copy)
- ☐ Other, please specify...

Type here

15. In the following question you will be provided with several clinical tasks in the form of narratives and be asked what data you would use to complete the task.

A. [Diagnose Latent Tuberculosis] You receive a laboratory report for a patient screened for tuberculosis who recently immigrated from India. Which of the following data types would you use / be required to make a diagnosis of latent tuberculosis?

B. [Diagnose Active Tuberculosis] You receive a laboratory report for a patient recently hospitalized with respiratory and constitutional symptoms suggestive of tuberculosis. Which of the following data types would you use / be required to make a diagnosis of active tuberculosis?

C. [Reactivation vs. New Acquisition] You receive a laboratory report for a patient confirming active tuberculosis. Which of the following data types would you use / be

required to differentiate between reactivation and new acquisition of tuberculosis?

D. [Characterize Transmission Risk] You have just diagnosed a patient with active tuberculosis and are determining what steps are necessary to prevent transmission to others. What data would you use / be required to characterize the patient's risk of transmission?

[Select as many as apply]

	A. Diagnose Latent Tuberculosis	B. Diagnose Active Tuberculosis	C. Reactivation vs. New Acquisition	D. Characterize Transmission Risk
Patient identifiers (Name, age, location)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Patient's own prior tuberculosis test results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Requester identifiers (Name, contact, copy to etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Reviewer identifiers (Name, position etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Type of sample (Sputum, fine needle aspirate etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sample collection site (lymph node, peripheral blood draw etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sample collection date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Report release date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Interpretation or comments from reviewer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Tuberculin Skin Test Results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Interferon Gamma Release Assay (IGRA) results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Chest X-ray results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Acid Fast Bacilli Smear results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Culture results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Speciation (m. tuberculosis, MAC, m. bovis etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Phenotypic drug susceptibility testing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Predicted (in silico) drug susceptibility testing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Specific Mutations conferring drug resistance (Resistotype)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Spoligotype	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MIRU-VNTR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Restriction fragment length polymorphisms (RFLP)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cluster assignment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Single Nucleotide Polymorphism/Variant distance from other isolates	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Phylogenetic tree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Laboratory performance measures (Sequence quality, coverage etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

16. When you are using laboratory data to diagnose a patient with active TB, you encounter the following challenges:

[Select as many as apply]

- ☐ No challenges - the lab data I currently receive is sufficient
- ☐ The lab data I currently receive does not help me to make a diagnosis
- ☐ I would like to receive data faster to make a more timely diagnosis
- ☐ Important results come at different times and/or in different documents
- ☐ I find it difficult to interpret the lab results I receive
- ☐ I am not regular receiving data that would help me to make a diagnosis
- ☐ The lab data I receive is not routinely linked to patient data
- ☐

Other, please specify...

Type here

17. In the following question you will be provided with several clinical tasks in the form of narratives and be asked what data you would use to complete the task.

A. [Choose Medications] You are managing a patient who has just been diagnosed with active tuberculosis. What data would you use / be required to decide what medications should be prescribed for the patient?

B. [Choose Duration of Treatment] You are managing a patient who has just been diagnosed with active tuberculosis. What data would be required to decide the duration of treatment for the patient?

C. [Assess Responsiveness to Treatment] You continue to follow the patient as they proceed with the therapeutic regimen for active tuberculosis. What data would be required to assess their responsiveness to treatment?

[Select as many as apply]

	A. Choose Medications	B. Choose Duration of Treatment	C. Assess Responsiveness to Treatment
Patient identifiers (Name, age, location)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Patient's own prior tuberculosis test results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Requester identifiers (Name, contact, copy to etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Reviewer identifiers (Name, position etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Type of sample (Sputum, fine needle aspirate etc)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sample collection site (lymph node, peripheral blood draw etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sample collection date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Report release date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Interpretation or comments from reviewer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Tuberculin Skin Test Results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Interferon Gamma Release Assay (IGRA) results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Chest X-ray results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Acid Fast Bacilli Smear results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Culture results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Speciation (m. tuberculosis, MAC, m. bovis etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Phenotypic drug susceptibility testing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Predicted (in silico) drug susceptibility testing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Specific Mutations conferring drug resistance (Resistotype)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Spoligotype	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MIRU-VNTR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Restriction fragment length polymorphisms (RFLP)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cluster assignment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Single Nucleotide Polymorphism/Variant distance from other isolates	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Phylogenetic tree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Laboratory performance measures (Sequence quality, coverage etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

18. What are the main barriers for improving the efficiency of active TB treatment through the use of molecular laboratory data?

[Select as many as apply]

- ☐ There aren't any barriers
- ☐ Additional laboratory data is needed
- ☐ Timeliness of results being provided (too slow)

10 of 17

- ☐ Results provided over multiple unconnected documents
- ☐ Difficulty interpreting lab results
- ☐ Lab data is not routinely provided
- ☐ Lab data is not routinely linked to patient data
- ☐ Other, please specify...

19. Do you have any additional comments you wish to make on the use of genomic and molecular data for active TB diagnosis and treatment?

Type here

PART IV – CONTACT TRACING AND OUTBREAK MANAGEMENT

All participants are asked to complete this part: Contact tracing and outbreak management are performed by nurses, clinicians, epidemiologists, and sometimes also researchers.

20. Are you involved in the epidemiological aspects of TB management, including contact tracing and/or managing outbreak?

Note that surveillance - collating data for regional or national-level efforts - is not included here. It will be covered in the next section.
[Select only one]

Yes

No

21. During your epidemiological work, do you directly review original lab reports?

[Select only one]

Yes

No

Do you get aggregate extracted data?

[Select only one]

Yes

No

22. In the following question you will be provided with several clinical tasks in the form of narratives and be asked what data you would use to complete the task.

A. [Guide Contact Tracing] You have been tasked with tracing potential contacts of a patient recently diagnosed with active tuberculosis. Which of the following data types would be useful in guiding contact tracing?

B. [Report to Public Health] You are a clinician managing several new cases of active tuberculosis and are concerned that they may represent a cluster. What data would influence your decision to report your concerns to public health?

C. [Define a Cluster] You are investigating increased incidence of tuberculosis in a rural community. What laboratory data would be required to define a cluster of tuberculosis cases?

D. [Connect Case to Existing Cluster] Following the identification of a cluster, new cases have been reported in a nearby community. What data would be required to connect these new cases to the existing cluster?

E. [Guide Public Health Response] What data would assist in guiding the public health response to the newly identified cluster?

[Select as many as apply]

	A. Guide Contact Tracing	B. Report to Public Health	C. Define a Cluster	D. Connect Case to Existing Cluster	E. Guide Public Health Response
Patient identifiers (Name, age, location)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Patient's own prior tuberculosis test results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Requester identifiers (Name, contact, copy to etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Reviewer identifiers (Name, position etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Type of sample (Sputum, fine needle aspirate etc)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sample collection site (lymph node, peripheral blood draw etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Sample collection date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Report release date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Interpretation or comments from reviewer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Tuberculin Skin Test Results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Interferon Gamma Release Assay (IGRA) results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Chest X-ray results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Acid Fast Bacilli Smear results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Culture results	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Speciation (m. tuberculosis, MAC, m. bovis etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Phenotypic drug susceptibility testing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Predicted (in silico) drug susceptibility testing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Specific Mutations conferring drug resistance (Resistotype)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Spoligotype	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
MIRU-VNTR	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Restriction fragment length polymorphisms (RFLP)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Cluster assignment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Single Nucleotide Polymorphism/Variant distance from other isolates	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Phylogenetic tree	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Laboratory performance measures (Sequence quality, coverage etc.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

PART V – SURVEILLANCE

Only epidemiologists, surveillance analysts, and researchers are asked to complete this part of the survey.

23. Are you involved in tuberculosis surveillance?

Yes No

24. What data does your institution currently use as part of its surveillance practices?

[Select as many as apply]

- ☐ Patient identifiers (Name, age, location)
- ☐ Patient's own prior tuberculosis test results
- ☐ Requester identifiers (Name, contact, copy to etc.)
- ☐ Reviewer identifiers (Name, position etc.)
- ☐ Type of sample (Sputum, fine needle aspirate etc)
- ☐ Sample collection site (lymph node, peripheral blood draw etc.)
- ☐ Sample collection date
- ☐ Report release date
- ☐ Interpretation or comments from reviewer
- ☐ Tuberculin Skin Test Results
- ☐ Interferon Gamma Release Assay (IGRA) results
- ☐ Chest X-ray results
- ☐ Acid Fast Bacilli Smear results
- ☐ Culture results
- ☐ Speciation (m. tuberculosis, MAC, m. bovis etc.)
- ☐ Phenotypic drug susceptibility testing
- ☐ Predicted (in silico) drug susceptibility testing
- ☐ Specific Mutations conferring drug resistance (Resistotype)
- ☐ Spoligotype
- ☐ MIRU-VNTR
- ☐ Restriction fragment length polymorphisms (RFLP)
- ☐ Cluster assignment
- ☐ Single Nucleotide Polymorphism/Variant distance from other isolates
- ☐ Phylogenetic tree
- ☐ Laboratory performance measures (Sequence quality, coverage etc.)

25. Is your institution planning to use more genomic data in the future?

[Select only one]

- ☐ Yes – we're looking into it right now
- ☐ Not yet – but we'd like to incorporate genomic data in the future
- ☐ No and we have no plans to do so in the near future

How do envision genomic data being part of future surveillance efforts?

Type here

26. What is the main barrier of using genomic data more routinely as part of surveillance?

[Select as many as apply]

- ☐ Data is not consistently accessible
- ☐ Data are not consistently linked to relative patient data
- ☐ It is not clear how this data is useful for surveillance
- ☐ It is not clear how to interpret this data for surveillance purposes
- ☐ Difficulty interpreting lab results
- ☐ Other, please specify...

Type here

PART VI – CONTACT INFORMATION

All participants are asked to complete this part of the survey.

Would you like to provide an email address so that we can contact you for the post-survey gift card draw and/or later email with the results of this survey? This contact information will be removed when we anonymize the survey data before making it available to other researchers.

[Select as many as apply]

- ☐ Yes, please enter me into the gift card draw for participants who complete this survey
- ☐ Yes, please send me the final results of this study

Email Address:

Type here

A.5 Design Choice Questionnaire Online Survey

Questionnaire submitted online for the design choice survey

COMPASS-TB Report Design: Second Survey

0%

DESCRIPTION AND CONSENT

Many public health agencies are starting to use whole genome sequencing (reading every letter of an organism's DNA) as a tool for diagnosing infections, predicting what antibiotics an organism is sensitive or resistant to, and identifying closely related isolates that might suggest an outbreak. Last year, [a study in The Lancet Infectious Diseases](#) showed that when this technique is used in the tuberculosis laboratory, we can generate all the usual results that one has come to expect from a reference mycobacteriology lab, but we can do so much faster and at lower cost. As a result of this study, groups like Public Health England, the BC Centre for Disease Control, and the US Centers for Disease Control and Prevention are all using genomics to analyze their incoming mycobacterial isolates.

Sequencing a bacterial genome generates a lot of information, only some of which might be needed to manage a patient's infection. We are interested in designing a new lab report form that will help to communicate tuberculosis genomic data in a clear, concise, and meaningful way that will help those in the tuberculosis community - clinicians, epidemiologists, laboratory scientists, and more - in their daily work. There is a large field of research into how to present data in a way that makes it easily interpretable - we will be using principles from this field in designing our new report format, which will be shared with public health laboratories so that they may choose to use it in their own reporting.

By participating in this survey, you will help us better understand how lab data should be represented and what design elements should be used in the final report. The results of this survey will be used to construct a final prototype report that will be tested in a third and final survey later this year.

[Consent for Participation](#)

STUDY PROCEDURES:

If you agree to voluntarily participate in this research, your participation will include the following online survey (estimated completion time 15-30 minutes) in which you will be asked to compare different visual representations of genomic data and choose your preferred design. At the end of the survey, you may choose to provide an email address if you'd like to be entered into a draw for an Amazon gift card.

There are no known or anticipated risks to you by participating in this research, and the benefit is receiving the results of the study via an emailed report at the project's conclusion, which will include a template for the final report design that participants may use in their own work. Study results will be shared with the research community through open-access publications, conference reports, tweets and other social media postings.

MEASURES TO MAINTAIN CONFIDENTIALITY

Data from this study will be coded anonymously.

CONTACTS FOR COMPLAINTS OR CONCERNS

Geoff McKee is a resident physician in Public Health and Preventive Medicine at the University of British Columbia and you may contact him if you have any further questions by email

If you have any concerns or complaints about your rights as a research participant and/or your experiences while participating in this study, contact the Research Participant Complaint Line in the UBC Office of Research Ethics

Taking part in this study is entirely up to you. You have the right to refuse to participate in this study. If you decide to take part, you may choose to pull out of the study at any time without giving a reason.

By completing the questionnaire, you are consenting to participate in this research.

PRINCIPAL INVESTIGATOR:
Jennifer Gardy, School of Population & Public Health

CO-INVESTIGATORS:
Geoff McKee, School of Population and Public Health,
Anamaria Crisan, School of Population and Public Health

COMPASS-TB Report Design: Second Survey

16%

PART I – DEMOGRAPHICS

First, we have a few short questions about your background.

1. Do you work with tuberculosis patients or the *Mycobacterium tuberculosis* bacterium at all?

[Select one option]

☐ Yes ☐ No

1B. What is your role in tuberculosis diagnosis, treatment, management, and/or surveillance?

[Select as many as apply]

- ☐ Physician - I work directly with TB patients, providing care and/or case management
- ☐ Nurse - I work directly with TB patients, providing care and/or case management
- ☐ Laboratory work – I work in a mycobacteriology laboratory setting where I am involved with lab testing for TB
- ☐ Surveillance/epidemiology - I work with TB data to understand patterns in disease occurrence
- ☐ Research - I carry out academic research into TB and/or *M. tuberculosis*
- ☐ Other, please specify...

2. Do you work in public health microbiology or microbial genomics, whether on TB or another pathogen?

[Select one option]

☐ Yes ☐ No

2B. What is your role in public health microbiology or microbial genomics?

[Select as many as apply]

- ☐ Clinical – I am directly involved in patient care and/or case management
- ☐ Bioinformatics – I use computational tools to analyse genomic data from pathogens
- ☐ Laboratory work – I am involved in directly handling and/or testing specimens
- ☐ Surveillance/epidemiology – I work with data to understand patterns in disease occurrence
- ☐ Research – I carry out academic research in public health and/or microbial genomics
- ☐ Other, please specify...

2C. What pathogens do you work on?

[Select as many as apply]

☐ Respiratory infections (e.g. influenza, pertussis)

☐ Enteric infections (e.g. Salmonella, E. coli)

☐ Vector-borne disease (e.g. malaria, Zika)

☐ Blood-borne disease (e.g. HIV, hepatitis)

☐ Other, please specify...

Type here

3. Who is your primary employer?

[Select as many as apply]

☐ Public Health Organization - e.g. Public Health England, CDC

☐ Private Clinic/Primary Care - e.g. a doctor's office

☐ Hospital

☐ Academic Institution

☐ Other, please specify...

Type here

4. In what country do you work?

[Select one option]

☐ United Kingdom

☐ Canada

☐ USA

☐ Other, please specify...

Type here

Back

Next

Administrator

Page 2 of 2



- About UBC

Contact UBC

About the University

News

Events

Careers

Make a Gift

Search UBC.ca
- UBC Campuses

Vancouver Campus

Okanagan Campus

UBC Sites

Robson Square

Great Northern Way

Faculty of Medicine Across BC

Asia Pacific Regional Office

COMPASS-TB Report Design: Second Survey

33%

PART II – Design Elements

Laboratory results are usually communicated to end-users like doctors or public health officials in the form of a brief one- or two-page report. There are many different styles of lab report, from simple text documents to colourful pictorial reports. We are interested in understanding what sort of design choices can make a TB genomic laboratory report easy for end-users to read and to act upon. The report will contain information on what mycobacterial species a patient is infected with, what antibiotics their TB infection is susceptible or resistant to, and whether or not their TB isolate is related to other isolates and might be part of an outbreak.

Throughout the rest of the survey, we will be showing you some designs that show these different data – speciation, resistance, and epidemiological relatedness – in different ways. We want to find out which designs you prefer, so that these design elements can be incorporated into a final report design later in our project.

First, we will look at small elements of the report design.

5A. You are reading a summary of a patient's lab test results. Which of the following summary statement formats is better at communicating the information you need to know to do your job?

A

Summary

The specimen is positive for *Mycobacterium tuberculosis*. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.

B

Summary

The specimen is positive for *Mycobacterium tuberculosis*. It is resistant to isoniazid and rifampin. It belongs to a cluster, suggesting recent transmission.

[Select one option]

- ☐ A (with bolding)
- ☐ B (without bolding)
- ☐ They are equally informative.

5B. Please explain your choice or provide feedback.

[Optional]

Type here

6A. One section of the report will describe which mycobacterial species a patient was diagnosed with. Which headline best describes this section of the report?

A

Speciation

The specimen is positive for *Mycobacterium tuberculosis*.

B

Organism

The specimen is positive for *Mycobacterium tuberculosis*.

C

Diagnosis

The specimen is positive for *Mycobacterium tuberculosis*.

D

Species

The specimen is positive for *Mycobacterium tuberculosis*.

[Please rank your choices]

A (Speciation)	1	1
B (Organism)	2	2
C (Diagnosis)	3	3
D (Species)	4	4

6B. Please explain your choice or provide feedback.

[Optional]

Type here

7A. Which wording best conveys tuberculosis speciation results?

A

Speciation

The specimen is positive for *Mycobacterium tuberculosis*.

B

Speciation

Organism: *Mycobacterium tuberculosis*

[Select one option]

- ☐ A (Full sentence)
- ☐ B (Summary)

☐ They are equally informative

7B. Please explain your choice or provide feedback.

[Optional]

Type here

8A. The presence of particular mutations in a TB genome can be used to predict whether a specimen is sensitive or resistant to specific antibiotics. Which headline best describes this section of the report?

A

Drug Resistance

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

B

Drug Sensitivity

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

C

Drug Susceptibility

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

D

Treatment

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

[Please rank your choices]

A (Drug Resistance)

11

B (Drug Sensitivity)

22

C (Drug Susceptibility)

33

D (Treatment)

44

8B. Please explain your choice or provide feedback.

[Optional]

Type here

9A. There are many ways to represent a TB drug's name, from a single letter to a full name. Which naming scheme is most useful on a report?

[Select one option]

- ☐ Full Name (Ex. isoniazid)
- ☐ 3-letter abbreviation (Ex. INH)
- ☐ 1-letter abbreviation (Ex. H)
- ☐ Show me everything - (Ex. Isoniazid (INH, H))
- ☐ They are equally informative

9B. Please explain your choice or provide feedback.

[Optional]

Type here

10A. A specimen can be described as susceptible to an antibiotic (high likelihood of clinical success), resistant to an antibiotic (low likelihood of clinical success), intermediate (clinical success uncertain), or unknown (not enough information to draw a conclusion). Which naming scheme is most useful on a report?

[Select one option]

- ☐ Full Name (Ex. Susceptible, Resistant, Unknown)
- ☐ 1-letter abbreviation (Ex. S, R, U)
- ☐ They are equally informative

10B. Please explain your choice or provide feedback.

[Optional]

Type here

11A. Drug resistance in TB is caused by point mutations – single base-pair changes that alter the normal function of a gene or the protein it encodes. If a resistance phenotype is predicted from genomic data, would you want to know the exact mutation that caused it?

[Select one option]

- ☐ Yes – on the same table with the drug susceptibility data
- ☐ Yes, but on the other side of the report
- ☐ No – I am not interested in the mutation data

11B. What types of information related to the point mutation would you want to see?

[Select as many as apply]

- ☐ Gene abbreviation (e.g. katG, inhA)
- ☐ Base pair change (e.g. A1562C)
- ☐ Amino acid change (e.g. S531T)
- ☐ Number of sequencing reads that position (e.g. 48x)
- ☐ Number of reads supporting the mutation/coverage (e.g 47/48)

12A. Here are four ways of showing a result in which a specimen is resistant to two drugs. Which one is easiest for you to interpret?

A

Drug Susceptibility

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

B

Drug Susceptibility

Drug	Prediction
Isoniazid	Resistant ⚠
Rifampin	Resistant ⚠
Ethambutol	Sensitive
Pyrazinamide	Sensitive

C

Drug Susceptibility

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

D

Drug Susceptibility

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

[Please rank your choices]

A (Basic)	1	1
B (Alert Glyphs)	2	2
C (Shaded)	3	3
D (Bolded)	4	

12B. Please explain your choice or provide feedback.

[Optional]

Type here

13A. Depending on the resistance mutations observed, an isolate might be identified as having multidrug-resistant TB (MDR-TB). There are many ways this could be noted on the report.

A

Drug Susceptibility

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

B

Drug Susceptibility

Based on predicted antibiotic sensitivities, this individual has multidrug-resistant (MDR) TB.

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

C

Drug Susceptibility

Mono-resistant☐

Multidrug-resistant (MDR)☒

Extremely Drug Resistant (XDR)☐

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

[Please rank your choices]

A (Basic)

11

B (Summary Sentence)

22

C (Tick Boxes)

33

13B. Please explain your choice or provide feedback.

[Optional]

Type here

14A. One section of the report will describe whether a patient's specimen is closely related to any specimens that were previously sequenced, suggesting the cases might be part of a cluster or outbreak. Which headline best describes this section of the report?

A

Relatedness

	Likely Related (less than 5 SNP Difference)	Possibly Related (6-30 SNP Differences)
Number of Isolates	2	6

B

Epidemiology

	Likely Related (less than 5 SNP Difference)	Possibly Related (6-30 SNP Differences)
Number of Isolates	2	6

C

Cluster Detection

	Likely Related (less than 5 SNP Difference)	Possibly Related (6-30 SNP Differences)
Number of Isolates	2	6

[Please rank your choices]

A (Relatedness)	1	1
B (Epidemiology)	2	2
C (Cluster Detection)	3	3

14B. Please explain your choice or provide feedback.

Optional

Type here

Back

Next

COMPASS-TB Report Design: Second Survey

50%

PART III – Report Sections

Now that we’ve looked at some individual design elements, we will next look at each of the three sections of the report: what organism is this, what antibiotics is it sensitive to, and is it related to other specimens. For each section, we will show you a few different representations of the same dataset; we want to know which one you prefer. Factors such as ease of readability, time taken to interpret the result, and aesthetics may all influence your choice

15A. Data on speciation and diagnosis is presented below in two different formats. Which do you find most interpretable?

A

Speciation

Organisms	Percent Match
M. tuberculosis	100%
M. canettii	40%
Mycobacterium Absum Complex	20%

B

Speciation

The specimen is positive for *Mycobacterium tuberculosis*.

[Select one option]

- ☐ A (Percent match)
- ☐ B (Organism name)

15B. Please explain your choice or provide feedback.

[Optional]

Type here

16A. Data on drug susceptibility is presented below in a number of different formats. Which do you find most interpretable?

A Drug Susceptibility

Prediction	Drugs
Sensitive	Ethambutol, Pyrazinamide
Resistant	Isoniazid, Rifampin
Indeterminate	-

B Drug Susceptibility

Drug	Prediction
Isoniazid	Resistant
Rifampin	Resistant
Ethambutol	Sensitive
Pyrazinamide	Sensitive

C Drug SusceptibilityThe specimen is resistant to isoniazid, rifampin. It is sensitive to **ethambutol** and pyrazinamide.**D** Drug Susceptibility**E** Drug Susceptibility

[Please rank your choices]

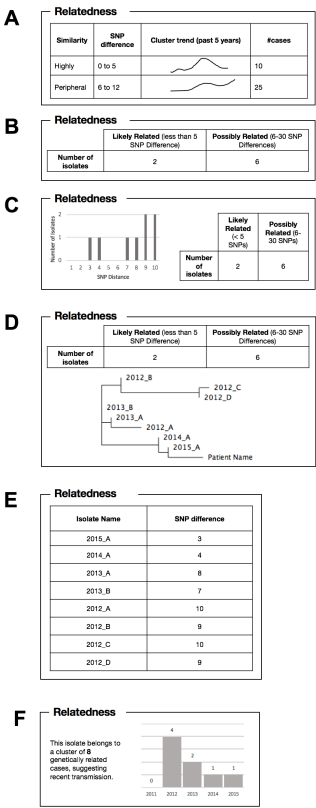
A (Drugs listed by category)	1	1
B (Prediction by drug)	2	2
C (Summary sentence)	3	3
D (Drugs listed by category bin)	4	4
E (Abbreviated prediction by drug)	5	5

16B. Please explain your choice or provide feedback.

[Optional]

Type here

17A. Data on relatedness to other isolates/clusters is presented below in a number of different formats. Which do you find most interpretable?



[Please rank your choices]

A (# of cases with spark line)	1 1
B (# of isolates related table)	2 2
C (Table + Graph of # of isolates by SNP distance)	3 3

Click on images to zoom



COMPASS-TB Report Design: Second Survey

66%

PART IV – Report Feedback

In the last part of the survey, we will show you four potential prototype reports. You will have seen some of the elements already – things like speciation and resistance prediction – but you'll also see new information, such as a quality report describing the genome sequencing analysis. The reports have been organized such that the most critical information appears on page one, with expanded details on page two. Please read carefully through both pages before answering the questions.

20A. Please review the following report and select the response indicating your agreement with the corresponding statements.

Click on images to zoom

[illegible]

103

103

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
This report is easy to read.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I know what the information in this report means.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can read this report and get the information I need quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel that I can accurately interpret the information on this report.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

20B. Please provide any additional comments you may have on the report.

[Optional]

21A. Please review the following report and select the response indicating your agreement with the corresponding statements.

Click on images to zoom

BC Tuberculosis Genome Sequencing Results Page 1 of 2

Patient Information

Patient Name	John Doe	Sample Type	Sputum
Patient ID	12345678	Sample Date	01-01-2016
Patient Sex	Male	Sample Site	01-01-2016
Location	Colomb	Specimen ID	101001234

Summary of Findings

Based upon analysis of the specimen's genomic data, this patient has **mycobacterium tuberculosis** that is predicted to be resistant to 2 antibiotics (**Isoniazid, Rifampin**). This case belongs to a cluster of cases with similar genomic findings.

Diagnosis

Molecularly genetic data from the specimen was compared to mycobacterium and non-mycobacterium reference genomes for identification and resistance.

The specimen was sequenced as **mycobacterium tuberculosis**.

Treatment

Molecularly genetic data from the specimen was compared to mycobacterium and non-mycobacterium reference genomes for identification and resistance.

The specimen was considered to be **multi-drug resistant (MDR) TB**.

Summary of resistance findings:

Drug	Resistance	Status	Comment
Isoniazid	Resistant	1	Gene: hspR , Rifamycin Acid Change: 300T
Rifampin	Resistant	1	Gene: hspR , Rifamycin Acid Change: 300T
Fluoroquinolone	Sensitive	✓	
Pyrazinamide	Sensitive	✓	
ETH	Sensitive	✓	
BM	Sensitive	✓	
MDR	Sensitive	✓	

BC Tuberculosis Genome Sequencing Results Page 2 of 2

Epidemiologic Summary

Molecularly genetic data from the specimen was compared to mycobacterium and non-mycobacterium reference genomes for identification and resistance.

The specimen belongs to a **previously existing cluster**.

Severity	SNP difference	Cluster trend (past 5 years)	Members (Total)
Highly	0 to 5		2
Prevalent	6 to 10		6

Quality Summary

The sequencing quality analysis of this sample was completed (99.99% coverage) as the number of reads was greater than 47 million with 99.99% mapped and a coverage of 99.99%.

Comments

None

Author Dr. John Doe
Position Laboratory Director
Signature [Signature]
Date 01-01-2016

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
This report is easy to read.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I know what the information in this report means.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can read this report and get the information I need quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel that I can accurately interpret the information on this report.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

21B. Please provide any additional comments you may have on the report.

[Optional]

22A. Please review the following report and select the response indicating your agreement with the corresponding statements.

Click on images to zoom

[illegible]

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
This report is easy to read.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I know what the information in this report means.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can read this report and get the information I need quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel that I can accurately interpret the information on this report.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

22B. Please provide any additional comments you may have on the report.


[Optional]


23A. Please review the following report and select the response indicating your agreement with the corresponding statements.


Click on images to zoom


MYCOBACTERIAL GENOME SEQUENCING REPORT


Reported by: **OSORD** Report Date: **1 Jan 1990**


1  **PATIENT INFORMATION**
 Name: **John A. Johnson** Accession: **123456789**
 Date: **1 Jan 1990** Gender: **Male** Date: **1 Jan 1990**
 Location: **Mykology** Contact: **John Doe**

2  **SPECIES IDENTIFIED BY SEQUENCING**
 100% identity to *Mycobacterium tuberculosis*

3  **PREDICTED ANTIBIOTIC RESISTANCE**
 Resistant to: **vancomycin, clindamycin**


4  **EPIDEMIOLOGICAL RELATIONSHIPS**
 Grouping of isolates: **4** genetically related strains, suggesting recent transmission.

5  **SEQUENCING QUALITY**
 Coverage: **4,849,783 bp** in **10,000,000 bp** (94.99%); **4,738 reads**.
 Consensus: **99.99%** (99.99% in **4,849,783 bp**)

6  **COMMENTS**
 The sample was sequenced twice. The resulting sequences are not quite high quality and should be discarded.

MYCOBACTERIAL GENOME SEQUENCING REPORT

Reported by: **OSORD** Report Date: **1 Jan 1990**

7  **The value of the report depends on the information provided on the first page.**

Resistotype

The resistotype describes the resistances that have been tested on this organism.

Drug	Test	Resistance	Colony	Control	Report
Vancomycin	yes	yes	100%	yes	100% yes

Related Isolates

The following table lists other strains isolates that have been identified as being genetically similar to the patient's isolate.

	Isolate	CFR	CFR Values
1	201A	201A	4
	201B	201B	4
	201C	201C	7
	201D	201D	7
	201E	201E	9
2	201F	201F	14
	201G	201G	9

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
This report is easy to read.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I know what the information in this report means.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can read this report and get the information I need quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel that I can accurately interpret the information on this report.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

23B. Please provide any additional comments you may have on the report.

[Optional]

24A. The previous 4 report prototypes demonstrate different ways of presenting lab data from whole genome sequencing of a tuberculosis isolate. Which of the reports do you prefer?

Please see previous questions for enlarged images.

[illegible]

B Tuberculosis Genome Sequencing Results Page 1 of 2

Patient Information

First Name	John	Second Name	Smith
Age	45	Gender	Male
Address	123 Main St, Vancouver, BC V6A 1A1		

Summary of Findings

Genome sequencing was performed on the patient's sputum sample. The results show a high degree of similarity to the reference genome, indicating a high level of genetic relatedness. The patient is currently on treatment for tuberculosis.

Diagnosis

The patient was diagnosed with Tuberculosis (Tb) based on clinical findings and laboratory results.

Treatment

The patient is currently on treatment with the following regimen: Isoniazid, Rifampin, Pyrazinamide, and Ethambutol.

Drug	Dose	Frequency	Comments
Isoniazid	300 mg	Once daily	Continue
Rifampin	600 mg	Once daily	Continue
Pyrazinamide	900 mg	Once daily	Continue
Ethambutol	150 mg	Once daily	Continue

Page 1 of 2

B Tuberculosis Genome Sequencing Results Page 2 of 2

Epidemiologic Summary

The patient's isolate is a primary, drug-resistant strain.

Resistance	Level	Resistance Level	Resistance Level
Isoniazid	High	High	High
Rifampin	High	High	High
Pyrazinamide	High	High	High
Ethambutol	High	High	High

Quality Summary

The patient's isolate is a primary, drug-resistant strain.

Comments

The patient's isolate is a primary, drug-resistant strain.

Page 2 of 2

C Mycobacterial Genome Sequencing Results Page 1 of 2

Patient Information

First Name	John	Second Name	Smith
Age	45	Gender	Male
Address	123 Main St, Vancouver, BC V6A 1A1		

Summary

Genome sequencing was performed on the patient's sputum sample. The results show a high degree of similarity to the reference genome, indicating a high level of genetic relatedness. The patient is currently on treatment for tuberculosis.

Diagnosis

The patient was diagnosed with Tuberculosis (Tb) based on clinical findings and laboratory results.

Treatment

The patient is currently on treatment with the following regimen: Isoniazid, Rifampin, Pyrazinamide, and Ethambutol.

Drug	Dose	Frequency	Comments
Isoniazid	300 mg	Once daily	Continue
Rifampin	600 mg	Once daily	Continue
Pyrazinamide	900 mg	Once daily	Continue
Ethambutol	150 mg	Once daily	Continue

Page 1 of 2

C Mycobacterial Genome Sequencing Results Page 2 of 2

Epidemiologic Summary

The patient's isolate is a primary, drug-resistant strain.

Resistance	Level	Resistance Level	Resistance Level
Isoniazid	High	High	High
Rifampin	High	High	High
Pyrazinamide	High	High	High
Ethambutol	High	High	High

Quality Summary

The patient's isolate is a primary, drug-resistant strain.

Comments

The patient's isolate is a primary, drug-resistant strain.

Page 2 of 2

D Mycobacterial Genome Sequencing Report Page 1 of 2

Patient Information

First Name	John	Second Name	Smith
Age	45	Gender	Male
Address	123 Main St, Vancouver, BC V6A 1A1		

Summary

Genome sequencing was performed on the patient's sputum sample. The results show a high degree of similarity to the reference genome, indicating a high level of genetic relatedness. The patient is currently on treatment for tuberculosis.

Diagnosis

The patient was diagnosed with Tuberculosis (Tb) based on clinical findings and laboratory results.

Treatment

The patient is currently on treatment with the following regimen: Isoniazid, Rifampin, Pyrazinamide, and Ethambutol.

Drug	Dose	Frequency	Comments
Isoniazid	300 mg	Once daily	Continue
Rifampin	600 mg	Once daily	Continue
Pyrazinamide	900 mg	Once daily	Continue
Ethambutol	150 mg	Once daily	Continue

Page 1 of 2

D Mycobacterial Genome Sequencing Report Page 2 of 2

Epidemiologic Summary

The patient's isolate is a primary, drug-resistant strain.

Resistance	Level	Resistance Level	Resistance Level
Isoniazid	High	High	High
Rifampin	High	High	High
Pyrazinamide	High	High	High
Ethambutol	High	High	High

Quality Summary

The patient's isolate is a primary, drug-resistant strain.

Comments

The patient's isolate is a primary, drug-resistant strain.

Page 2 of 2

[Please rank your choices]

A (Dark heading)		1	1
B (Gray heading)		2	2
C (Light Heading)		3	3
D (Pictures)		4	4

24B. Please explain your choice or provide feedback.

[Optional]

Type here

Back

Next

Administrator

Page 6 of 6



About UBC
Contact UBC
About the University
News
Events
Careers
Make a Gift
Search UBC.ca

UBC Campuses
Vancouver Campus
Okanagan Campus
UBC Sites
Robson Square
Great Northern Way
Faculty of Medicine Across BC
Asia Pacific Regional Office

COMPASS-TB Report Design: Second Survey

83%

PART V – CONTACT INFORMATION

Thank you so much for taking part in our survey! Your responses will help us create a better, more interpretable laboratory report. You can follow our project's progress at [Public Health InfoVis](#) – we will be collating the results of this survey and releasing a summary report on the blog shortly. We are also happy to email you a copy of the report.

Don't forget, by having completed the survey, you are eligible to enter our draw for an Amazon gift card. To enter the draw, please enter an email address below.

25. Would you like to provide an email address so that we can contact you for the post-survey gift card draw and/or later email with the results of this survey?

This contact information will be removed when we anonymize the survey data before making it available to other researchers.

- ☐ Yes, please enter me into the gift card draw for participants who complete this survey
- ☐ Yes, please send me the final results of this study

Email Address:

Type here

Back

Submit

[Administrator](#)

Page 6 of 9



- About UBC

Contact UBC

About the University

News

Events

Careers

Make a Gift

Search UBC.ca
- UBC Campuses

Vancouver Campus

Okanagan Campus

UBC Sites

Robson Square

Great Northern Way

Faculty of Medicine Across BC

Asia Pacific Regional Office

Appendix B

Adjutant Supplemental Materials

Adjutant is primarily a graphical user interface that implements a standard text mining workflow in addition to t-SNE and hdbscan under the hood to rapidly analyze a corpus of PubMed articles and to derive topic clusters in an unsupervised manner. This document details Adjutants implementation and then investigates the quality of Adjutants clustering abilities using both synthetic and real data.

Please note that this appendix is written in the conversational style of a tutorial and is available as an R markdown notebook at:

<https://github.com/amcrisan/adjutant/important-adjutant-details>

B.1 Adjutant Implementation Details

This section provides low-level details of Adjutant's algorithmic implementation.

Querying PubMed and assembling a document corpus. Given a PubMed-compatible search query, Adjutant uses the `RISmed` package [59] to obtain a summary of articles, including PubMed ID, Journal, Article Title, Authors, Abstract, Public Date, and MeSH terms. It then uses the `jsonlite` package [84] and the E-Utils eSummary API to query and extract additional metadata, including PubMed

Central (PMC) ID, article DOI, PMC citation count, article type (i.e. Journal Article, Review, Meta-Analysis), and language. Both RISmed and jsonlite are used because of the differing outputs from the E-Utils eFetch and eSummary APIs.

Data wrangling. Adjutant decomposes the PubMed document corpus into single-word entities extracted from article titles and abstracts and converts it to a tidy format for further analysis using the `tidytext` package [109]. All words are stemmed using Porter’s algorithm [116], from the `SnowballC` package [10], after which common (stemmed) stop words are removed. Again, using `tidytext` package resources, Adjutant next calculates the term frequency inverse document frequency (tf-idf) metric, and then filters terms that are too infrequent (fewer than 1% of all documents) or too frequent (more than 70%). Finally, Adjutant generates a document term matrix (DTM), with articles as rows, stemmed single words as columns, and tf-idf as the relevant analytic metric.

Unsupervised Topic Clustering. The multidimensional DTM is decomposed into two dimensions using the Barnes-Hut t-SNE [115] implementation from the `Rtsne` package [60]. We use default t-SNE parameters, except when the document corpus contains more than 1000 articles and when the t-SNE the perplexity parameter is set to 50 [122]; however, Adjutant also allows users to modify the perplexity and theta t-SNE parameters after an initial analysis is complete. Next, Adjutant derives clusters using the `hdbscan` algorithm [15] from the `dbscan` package [49]. Adjutant will attempt to automatically calculate the optimal `hdbscan` minimum cluster points (`minPts`) parameter, with optimal being defined as the fewest number of clusters that best fits the t-SNE data. Adjutant identifies the optimal `minPts` parameters by leveraging goodness-of-fit measurements derived from linear models, specifically the adjusted R^2 and the Bayesian Information Criteria (BIC); thus, each `minPts` parameter value tested will have an associated R^2 and BIC measure. Adjutant makes this calculation by fitting separate linear models to each of the two t-SNE dimensions, where for each linear model the t-SNE component co-ordinates are used as the dependent variable and the clusters are used as the independent variables. Each cluster is a vector of membership probabilities, from 0 (not in the cluster) to 1 (definitely a cluster member). The adjusted R^2 between the two component models are multiplied, and the BICs are averaged. To choose the optimal `minPts`

parameters, Adjutant identifies all minPts values with an adjusted R^2 within 0.05 of the best performing minPts value, and among those different options selects the minPts value with the lowest BIC. The clusters resulting from the optimal minPts value are named using the two most commonly occurring terms within the cluster.

B.2 Adjutant in Action

This section provides an overview of Adjutant’s functionality using both real and synthetic data. These are the libraries needed to run the analysis.

```
library(MASS)
library(ggplot2)
library(adjutant)
library(dplyr)
library(Rtsne)
library(dbSCAN)
library(tidytext)
library(reshape)
library(ggthemes)
library(cowplot)
library(topicmodels)
library(stringr)
library(SnowballC)
```

This section provides an overview of Adjutant’s functionality using both real and synthetic data.

B.2.1 t-SNE with Simulated Data

Prior to testing Adjutant with a real data set, we will explore the limitations of Adjutant with some generated data, thus knowing the ground truth. Note that the Distill Pub article on t-SNE [122] does this as well, however, here we are testing the specific configuration and environmental dependencies upon which Adjutant is built.

A Very Simple Example

The Wikipedia article for t-SNE advises against clustering on t-SNE dimensionally reduced data points and for rationale references a stack exchange discussion as the basis of this conclusion. Here we begin by examining the rationale of that stack exchange argument.

```
set.seed(416)

# function to generate multivariate normal distributions
multiNormGen<-function(mu = rep(0,2),
                        Sigma=matrix(c(10,3,3,2),2,2),
                        grpName = NULL,n=1000){

  values<-mvrnorm(n = n, mu = mu, Sigma = Sigma)
  if(!is.null(grpName)){
    return(data.frame(x=values[,1],
                      y=values[,2],
                      grp=rep(grpName,nrow(values))))
  }else{
    return(data.frame(x=values[,1],y=values[,2]))
  }
}

# generating the distributions
sampleDat<-rbind(multiNormGen(n=250,mu=c(-2,0),
                              Sigma = matrix(c(1,0,0,1),2,2),
                              grpName = "grp1"),
                 multiNormGen(n=750,mu=c(2,0),
                              Sigma = matrix(c(1,0,0,1),2,2),
                              grpName = "grp2"))

sampleDat$PMID<-1:nrow(sampleDat)

ggplot(sampleDat,aes(x=x,y=y,colour=grp))+
  geom_point(alpha=0.7)+
  theme_bw()
```

Running t-SNE

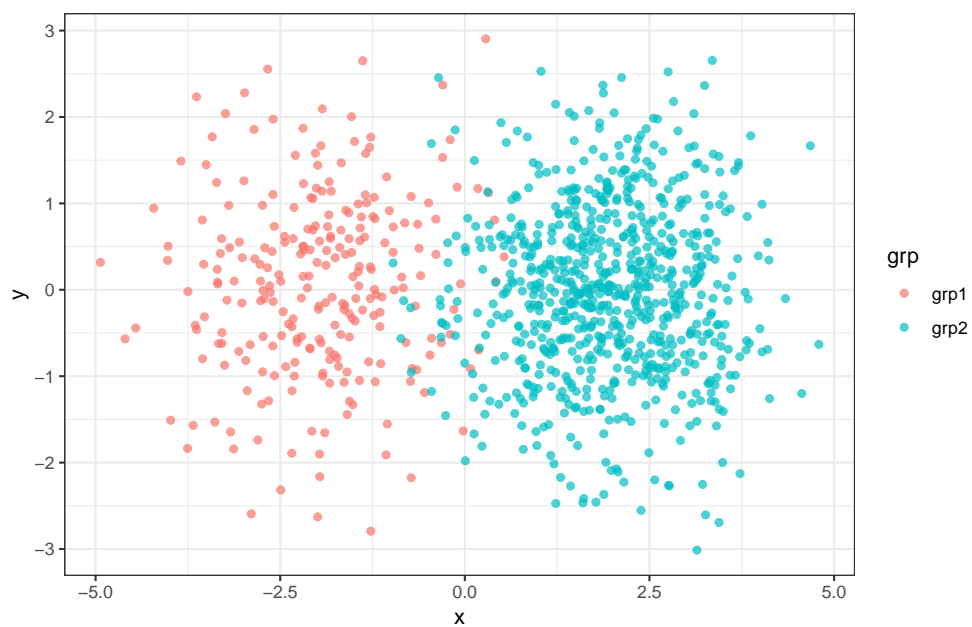


Figure B.1: Simple example of two class simulated distributions

Adjutant expects a Document Term Matrix (DTM) as input, so instead I have supplied the defaults Adjutant uses to choose a perplexity parameters for t-SNE.

```

# first, adjutant selects a perplexity
# depending upon the number of articles

tsnePer <- 30 #default adjutant value
if(nrow(sampleDat)>=1000){
  tsnePer<-50
}else if(nrow(values$corpus)<=100){
  tsnePer<-5
}

# then it runs t-SNE
tsneObj<-Rtsne(sampleDat[,c("x","y")],
               perplexity = tsnePer)

# some cleaning up and renaming to satisfy the next steps
df<-data.frame(cbind(1:nrow(tsneObj$Y),tsneObj$Y),
               stringsAsFactors = F)
colnames(df)<-c("PMID",paste("tsneComp",
                             1:(ncol(tsneObj$Y)),sep=""))

# let's take a look at the t-SNE output
df<-inner_join(sampleDat,df)

ggplot(df,aes(x=tsneComp1,y=tsneComp2,color=grp))+
  geom_point(alpha=0.7)+
  theme_bw()

```

It is possible to see that the larger cluster is spread out quite a bit so as to make it look like there are many other smaller clusters within it. This observation is inline with what the Stats Exchange comment brings up as well. Before addressing what it's like to cluster on this data, I will first run other perplexity parameter values to gauge how good Adjutant's default choices are.

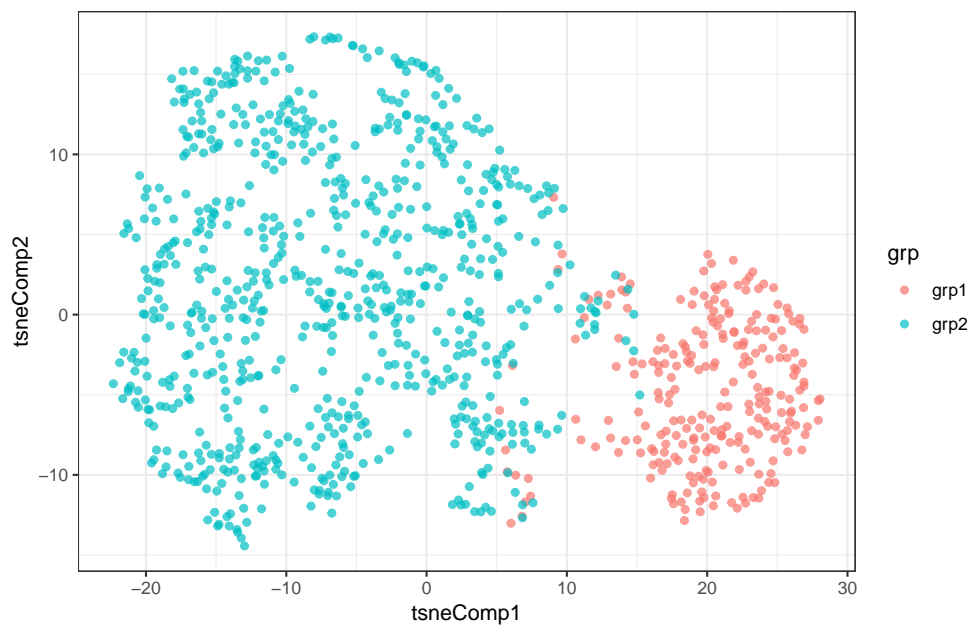


Figure B.2: Simple example: applying t-SNE

```
df<-c()
for(tsnePer in c(2,5,30,50,100)){
  tsneObj<-Rtsne(sampleDat[,c("x","y")],
                 perplexity = tsnePer)
  df<-rbind(df,cbind(1:nrow(tsneObj$Y),
                    rep(tsnePer,nrow(tsneObj$Y)),tsneObj$Y))
}

df<-data.frame(df,stringsAsFactors = FALSE)
colnames(df)<-c("PMID","perplexity","tsneComp1","tsneComp2")

# let's take a look at the t-SNE output
df<-inner_join(sampleDat,df)

ggplot(df,aes(x=tsneComp1,y=tsneComp2,color=grp))+
  facet_grid(~perplexity)+
  geom_point(alpha=0.7)+
  theme_bw()+
  labs(title="RTsne with variable perplexity parameters",
```

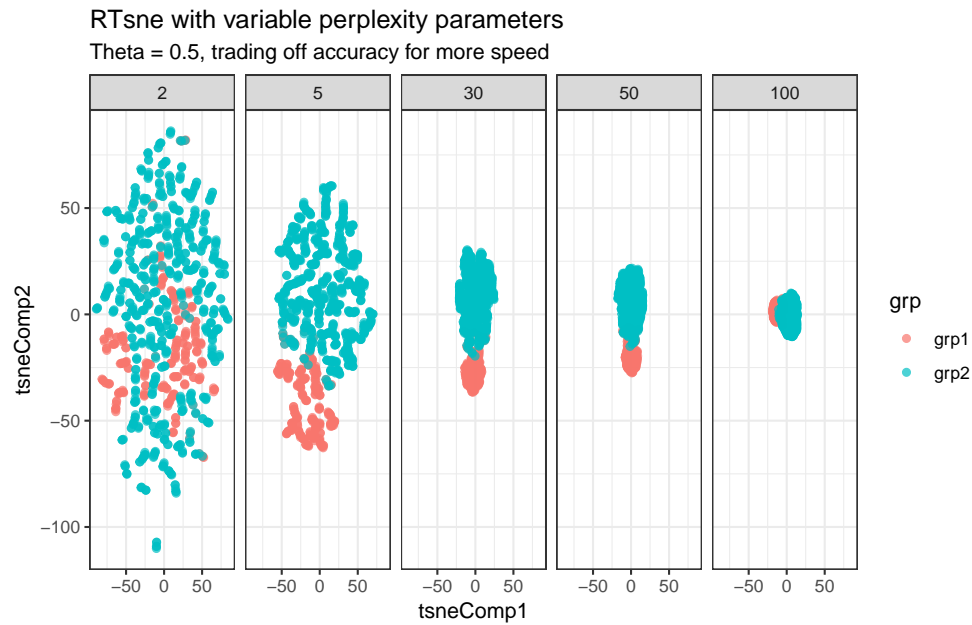


Figure B.3: Simple example: t-SNE results with varying parameters

```
subtitle="Theta = 0.5, trading off accuracy for speed")
```

What the figure above shows is that with a greater perplexity parameter the “closer” we get to resolving two clusters. This is true of what was shown in the Distill Pub t-SNE article as well - so while the exact original spatial orientation and density is not the same the original, the fact that are two slightly overlapping clusters is clear. What is notable is that we are not performing the fully accurate t-SNE here either, since the rTsne package has a theta parameter to speed up t-SNE at the cost of some accuracy. The default value of theta is 0.5, however, we can use a theta value of 0.0 to get the classical version of t-SNE:

```

df<-c()
for(tsnePer in c(2,5,30,50,100)){
  tsneObj<-Rtsne(sampleDat[,c("x","y")],
                 perplexity = tsnePer,theta=0)
  df<-rbind(df,cbind(1:nrow(tsneObj$Y),
                    rep(tsnePer,nrow(tsneObj$Y)),tsneObj$Y))
}

df<-data.frame(df,stringsAsFactors = FALSE)
colnames(df)<-c("PMID","perplexity","tsneComp1","tsneComp2")

# let's take a look at the t-SNE output
df<-inner_join(sampleDat,df)

ggplot(df,aes(x=tsneComp1,y=tsneComp2,color=grp))+
  facet_grid(~perplexity,scales="free")+
  geom_point(alpha=0.7)+
  theme_bw()+
  labs(title="RTsne with variable perplexity parameters",
       subtitle="Theta = 0.0, defaulting to classical t-SNE")

```

Although the pictures are slightly different, the results are the same, that with greater perplexity some greater discernability that there are indeed two clusters (note we've freed-up the x-axis scales to make it easier to see that last group).

Running Adjutant's hdbscan procedure

But the big question is can hdbscan reliably cluster on these data? The hdbscan algorithm requires that the user specifies the minimum number of points (minPts) in a cluster. It can be difficult to decide the best minPts parameter value, and so Adjutant automatically tries several different cluster sizes and selects the best one based upon the procedure specified in the implementation details.

```

# Under Adjutant, a tsne perplexity parameter of 50
# would have been selected for the data.
tsneObj<-Rtsne(sampleDat[,c("x","y")],
               perplexity = 50)
df<-data.frame(PMID = 1:nrow(tsneObj$Y),

```

RTsne with variable perplexity parameters

Theta = 0.0, defaulting to classical t-SNE

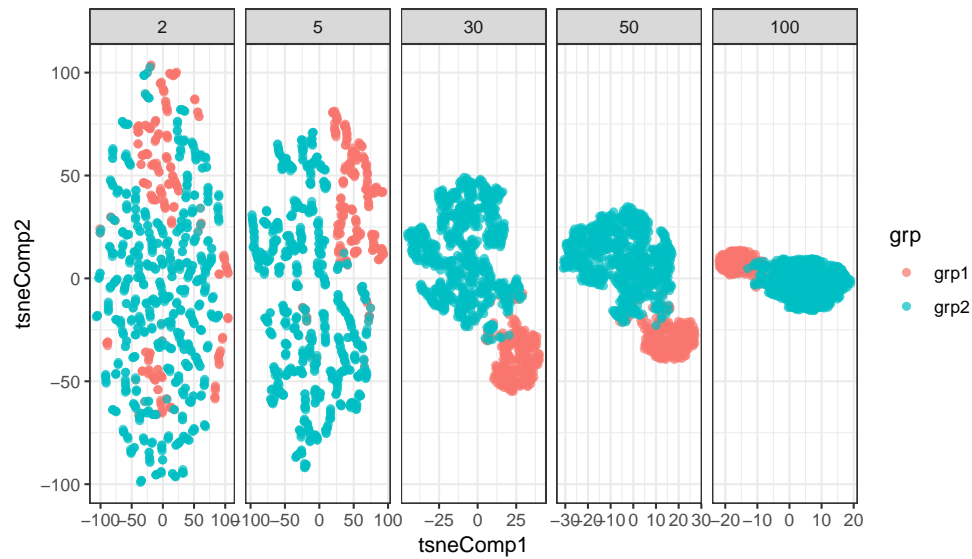


Figure B.4: Simple example: t-SNE results with varying parameters II

```
tsneComp1 = tsneObj$Y[,1],
tsneComp2 = tsneObj$Y[,2])

# now we can run the optimal params method,
# which runs HDBSCAN and picks the best parameters
optOut<-optimalParam(df)

#we can see all the choices adjutant cycles through
pList<-lapply(optOut$altChoices,function(x){
  x$fitPlot +
  labs(title = paste("minPts=",x$minPt)) +
  theme(legend.position="none",
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        axis.title = element_blank())
})

cowplot::plot_grid(plotlist = pList, nrow = 2)
```

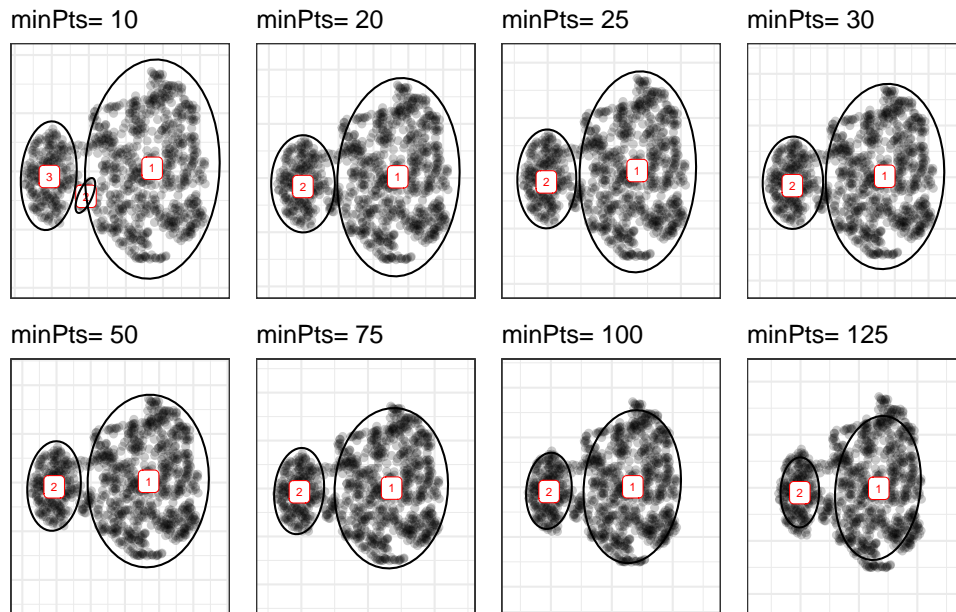


Figure B.5: Simple example: hdbscan on dimensionally reduced data

The above figures show how the data would be classified using different hdbscan parameters. Generally, hdbscan finds two clusters, with the exception of when smaller cluster sizes are permitted ($\text{minPts} = 10$) when three clusters are detected. Note that not all data points are assigned to clusters, some are assigned to noise category. This is not shown here, but is shown below, when using the “optimal” parameters selected automatically by Adjutant.


```

sampleDat$PMID<-factor(sampleDat$PMID)
df$PMID<-factor(df$PMID)

tmp<-inner_join(sampleDat,optOut$retItems)
df<-inner_join(df,tmp)

# add a group for noise
df<-df %>% mutate(grpRev= ifelse(tsneCluster == "0",
                                "Not-Clustered",
                                paste(grp)))

# get cluster of co-ordinates
clusterNames <- df %>%
  dplyr::group_by(grpRev) %>%
  dplyr::summarise(medX = median(tsneComp1),
                  medY = median(tsneComp2)) %>%
  dplyr::filter(grpRev != "Not-Clustered")

resolved<-ggplot(df,aes(x=tsneComp1,y=tsneComp2,group=grpRev))+
  geom_point(alpha=0.7,aes(colour=grpRev))+
  scale_colour_manual(values=c(scales::hue_pal()(2),"lightgray"))+
  stat_ellipse(aes(alpha = grpRev))+
  geom_label(data=clusterNames,
             aes(x=medX,y=medY,label=grpRev),
             size=2,
             colour="black")+
  scale_alpha_manual(values = c(0.7,0.7,0.0))+
  theme_bw()+
  labs(title="Resolved Plot",
       subtitle="Data after t-SNE and HDBSCAN.
       \nEllipses denote HDBSCAN clusters on t-SNE data")

original<-ggplot(df,aes(x=x,y=y,color=grp))+
  geom_point(alpha=0.7)+
  theme_bw()+
  labs(title="Original Plot",
       subtitle="Data before t-SNE and HDBSCAN")

cowplot::plot_grid(original,resolved,nrow=1)

```

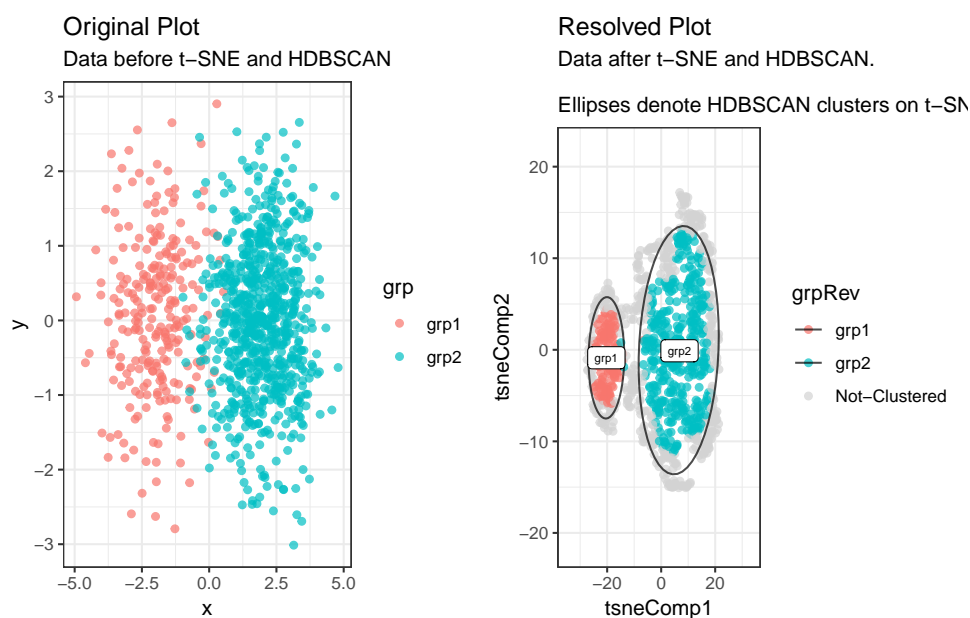


Figure B.6: Simple example: clusters resolved by Adjutant. The black ellipses indicate the cluster that Adjutant would automatically select.

The above figure shows the original data as well as the resolved clusters once they've been run through Adjutant. A couple things are notable in the above figure. First, Adjutant correctly suggests that there are two clusters in the data. The spatial orientation and the point density of clusters are not identical to the original, but the structure of groups holds (i.e. points from grp 1 continue to cluster together after being run through Adjutant's procedures). The next thing to note is that not all points could be classified. The hdbscan algorithm allows some points to be labelled as noise (unclassified to any distribution, indicated in grey). When the ground truth is known this is not a desirable outcome, however, in situations where we are less certain about what the shape and structure of the data should be we argue this is a useful feature. There is no doubt that the two distributions overlap, and that there are some points at the peripheries of both distributions – those are points the clustering procedure is less sure about and so are relegated to noise. What Adjutant does successfully cluster has a very strong signal. There are two ways to address the issue of noise, one is to force Adjutant to cluster as many of the points as possible and choose the minPts value that does just that – there are downsides to this approach

since sometimes ambiguity or uncertainty is warranted. The second is to use the more confidently clustered data as a prior to other techniques that can classify items relegated to noise. It is possible to conclude from this analysis that while t-SNE does misrepresent the original spatial positions of points it is still possible to find the correct number of clusters in the data. Adjutant's goal is to suggest clusters and in this task it's algorithmic procedure does a good job. We do advise caution in overly interpreting spatial positions of clusters, although we will show soon spatial positions are not entirely irrelevant either.

A More Complex Example

Instead of using two distributions, we'll now simulate a more complex scenario and later even add some noise.

Generating sample data

```
sampleDat<-c()

count = 1
prevPoint<-c()
while(count<10){
  n=sample(30:250,1)
  mu<-c(sample(-20:20,1),sample(-20:20,1))
  sC<-sort(sample(0:5,4,replace=TRUE),decreasing = T)

  Sigma<-matrix(c(sC[1],sC[3],sC[4],sC[2]),2,2)

  tmp<-rbind(prevPoint,mu)

  # make sure that distributions are
  # not sitting RIGHT on top of each other.
  if(sum(dist(tmp)<10)<1){
    prevPoint <- rbind(prevPoint,mu)
  }

  sampleDat<-rbind(sampleDat,
                    multiNormGen(n = n,mu = mu,
                                Sigma = Sigma,
```

```

        grpName = paste("grp", count)))
    count = count + 1
  }
}

# let's add some noise in the middle to
# keeps things interesting. Generally,
# I have found that noisy articles hang out
# in the middle and really clearly resolveable
# clusters out to the sides.
sampleDat$PMID<-1:nrow(sampleDat)

clusterNames <- sampleDat %>%
  dplyr::group_by(grp) %>%
  dplyr::summarise(medX = median(x),
                  medY = median(y))

ggplot(sampleDat, aes(x=x, y=y, colour=grp, group=grp)) +
  geom_point(alpha=0.7) +
  geom_label(data=clusterNames,
            aes(x=medX, y=medY, label=grp),
            colour="black") +
  stat_ellipse() +
  scale_colour_manual(values=c(
    tableau_color_pal("Classic 10 Medium")(10),
    "black")) +
  theme_bw()

```

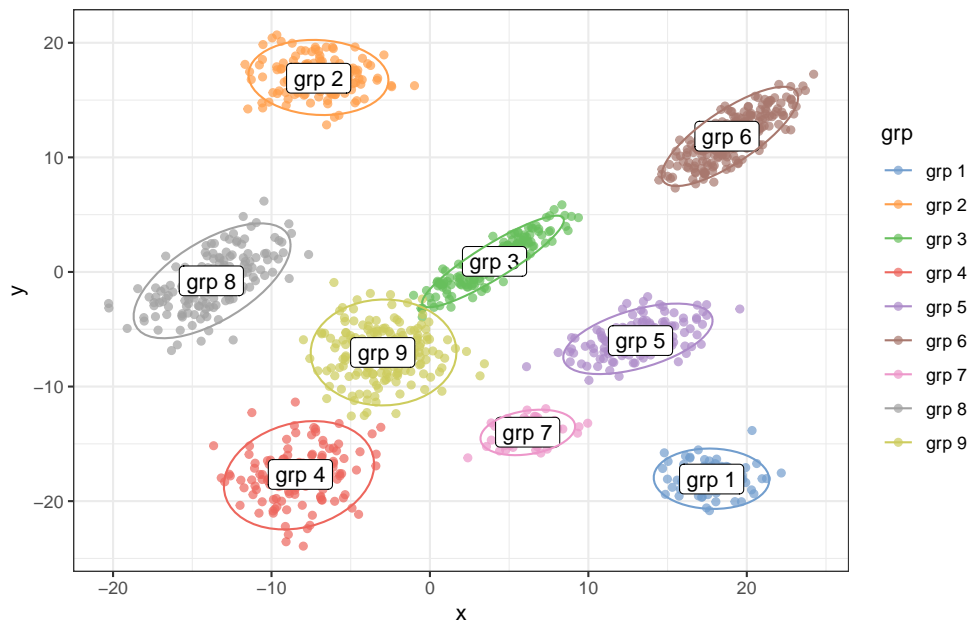


Figure B.7: Complex example: multiclass simulated distributions

Running *t*-SNE

```
#t-Sne resolution at different perplexity parameters
df<-c()
for(tsnePer in c(2,5,30,50,100)){
  tsneObj<-Rtsne(sampleDat[,c("x","y")],
    perplexity = tsnePer)

  df<-rbind(df,cbind(1:nrow(tsneObj$Y),
    rep(tsnePer,nrow(tsneObj$Y)),
    tsneObj$Y))
}

df<-data.frame(df,stringsAsFactors = FALSE)
colnames(df)<-c("PMID","perplexity","tsneComp1","tsneComp2")

#let's take a look at the t-SNE output
df<-inner_join(sampleDat,df)

ggplot(df,aes(x=tsneComp1,y=tsneComp2,color=grp))+
```

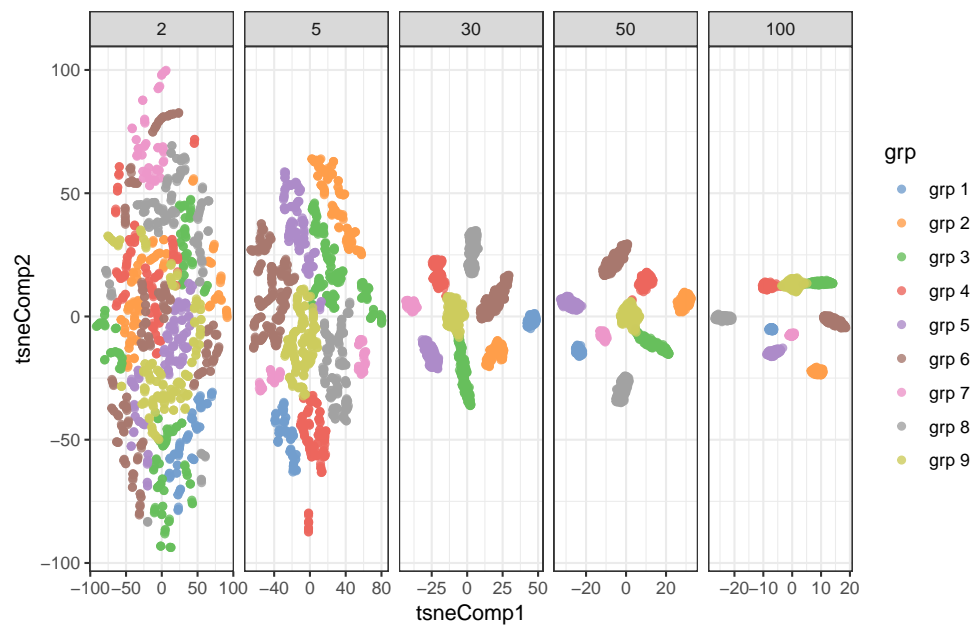


Figure B.8: Complex example:t-SNE results with varying parameters II

```
facet_grid(~perplexity,scales="free")+
geom_point(alpha = 0.8)+
theme_bw()+
scale_colour_manual(values=c
  tableau_color_pal("Classic 10 Medium")(10),

labs(title="RTsne with various perplexity parameters")
```

As with the simpler two distributions the higher the perplexity parameter the easier better t-SNE and resolving the individual clusters. Again, even a t-SNE parameter of just five starts to optimally separate the data points.

Running Adjutant's HDBSCAN Procedure

```
# under Adjutant, a tsne perplexity parameter
# of 50 would have been selected for the data.
tsneObj<-Rtsne(sampleDat[,c("x","y")],
               perplexity = 50)

df<-data.frame(PMID = 1:nrow(tsneObj$Y),
               tsneComp1 = tsneObj$Y[,1],
               tsneComp2 = tsneObj$Y[,2])

# now we can run the optimal params method,
# which runs HDBSCAN and picks the best parameters
optOut<-optimalParam(df)

#we can see all the choices adjutant cycles through
pList<-lapply(optOut$altChoices,function(x){
  x$fitPlot +
    scale_colour_manual(values=c(
      tableau_color_pal("Classic 10 Medium")(10),
      "black"))+
    labs(title = paste("minPts=",x$minPt)) +
    theme(legend.position="none",
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          axis.title = element_blank())
})

cowplot::plot_grid(plotlist = pList, nrow = 2)
```

Adjutant favours choosing a parameter that most spatially separates clusters and uses the minimum amount of clusters.

```
sampleDat$PMID<-factor(sampleDat$PMID)
df$PMID<-factor(df$PMID)

sampleDat<-inner_join(sampleDat,optOut$retItems)
df<-inner_join(df,sampleDat)

#add a group for noise
df<-df %>% mutate(grpRev= ifelse(tsneCluster == "0",
```

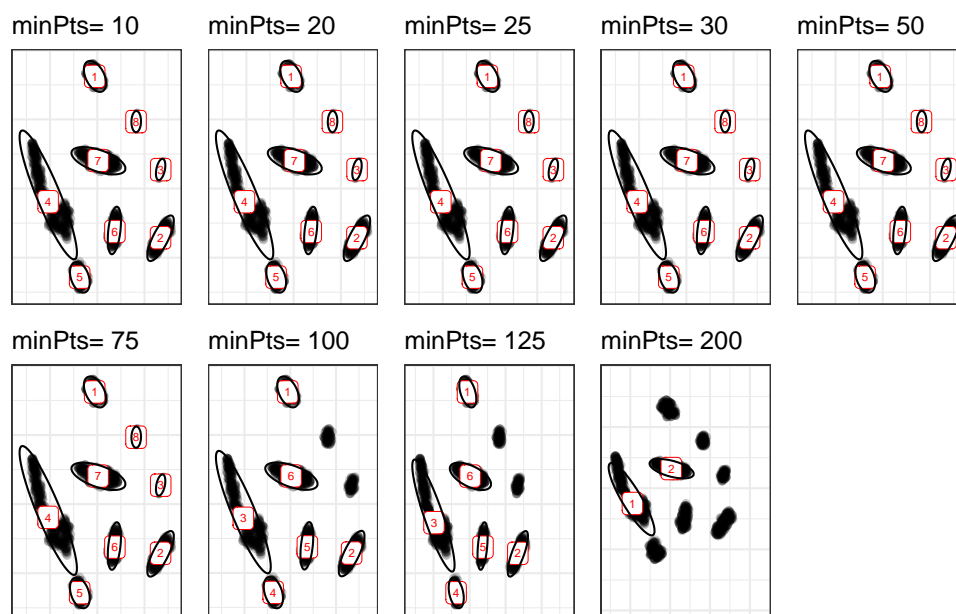


Figure B.9: Complex example: hdbscan on dimensionally reduced data

```

                                "Not-Clustered",
                                paste("hdbscan clust",
                                        tsneCluster)))

#get cluster of co-ordinates
clusterNames <- df %>%
  dplyr::group_by(grpRev) %>%
  dplyr::summarise(medX = median(tsneComp1),
                  medY = median(tsneComp2)) %>%
  dplyr::filter(grpRev != "Not-Clustered")

df<-df %>%
  mutate(isNoise = ifelse(grp== "Not-Clustered",
                          "Not-Clustered",
                          "Signal"))

resolved<-ggplot(df, aes(x=tsneComp1, y=tsneComp2, group=grpRev)) +

```



```

geom_point(aes(colour=grp,alpha = isNoise))+
stat_ellipse()+
theme_bw()+
scale_alpha_manual(values = c(0.2,0.7))+
scale_colour_manual(values=c(
  tableau_color_pal("Classic 10 Medium")(10),
  "black")))+
labs(title="Resolved Plot",
      subtitle="Data after t-SNE and HDBSCAN.
      \nEllipses denote HDBSCAN clusters on t-SNE data")

original<-ggplot(df,aes(x=x,y=y,color=grp))+
  geom_point(aes(alpha = isNoise))+
  theme_bw()+
  scale_alpha_manual(values = c(0.2,0.7))+
  scale_colour_manual(values=c(
    tableau_color_pal("Classic 10 Medium")(10),
    "black")))+
  labs(title="Original Plot",
        subtitle="Data before t-SNE and HDBSCAN")

cowplot::plot_grid(original,resolved,nrow=1)

```

Comparing the original plot against the t-SNE cluster plot a few things are noteworthy. The t-SNE co-ordinates *do not* recover the identical spatial positions (and by extension point density) of the original plot, but it is clear that Adjutant's procedures are none-the-less able to recover the original number of clusters even in this more complex example – which again is the goal of this tool. It turns out that in this example when the clusters are actually separable and not overlapping that none of the points are classified as noise.

This example also shows that the t-SNE spatial positions are not totally irrelevant. For example, the red, brown, and blue groups are close to each like in the original, as are the gray, purple, and orange clusters. The proximity of the blue cluster to the orange cluster is misleading, but this is likely an artifact of preserving its proximity to other clusters. It also appears that no documents are assigned to noise, meaning they all are assigned to some cluster.

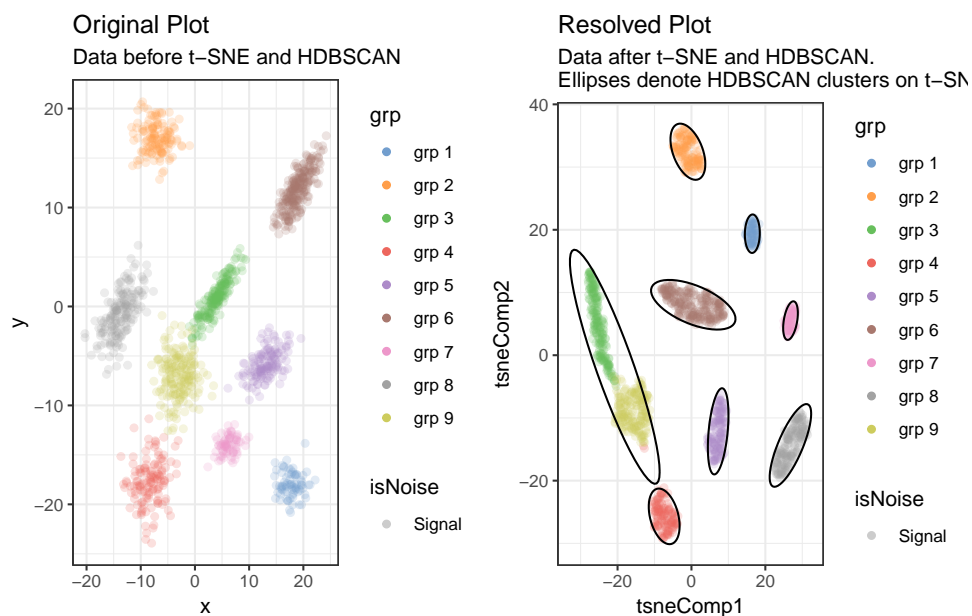


Figure B.10: Complex example: clusters resolved by Adjutant. The black ellipses indicate the clusters that Adjutant would automatically select.

Adding Noise

Next, we'll test the limits of Adjutant's algorithms to see how well it does with some messier, but still synthetic, data, by adding about 1000 data points of noise.

```
sampleDat<-dplyr::select(sampleDat,x,y,grp)

sampleDat<-rbind(
  sampleDat,
  multiNormGen(n=1000,mu=c(0,0),
    Sigma = matrix(c(60,35,30,55),2,2),
    grpName = "Not-Clustered"))

sampleDat$PMID<-1:nrow(sampleDat)
```

When adding the noise, it becomes clear that several clusters caught within it are lost within the central noise blob. Clusters on the periphery (yellow, green, purple, pink, and brown) do indeed still form their own clusters and appear to also pick up some points from the “noise” distribution. The yellow, orange, and blue clusters that



Figure B.11: Complex example with added noise

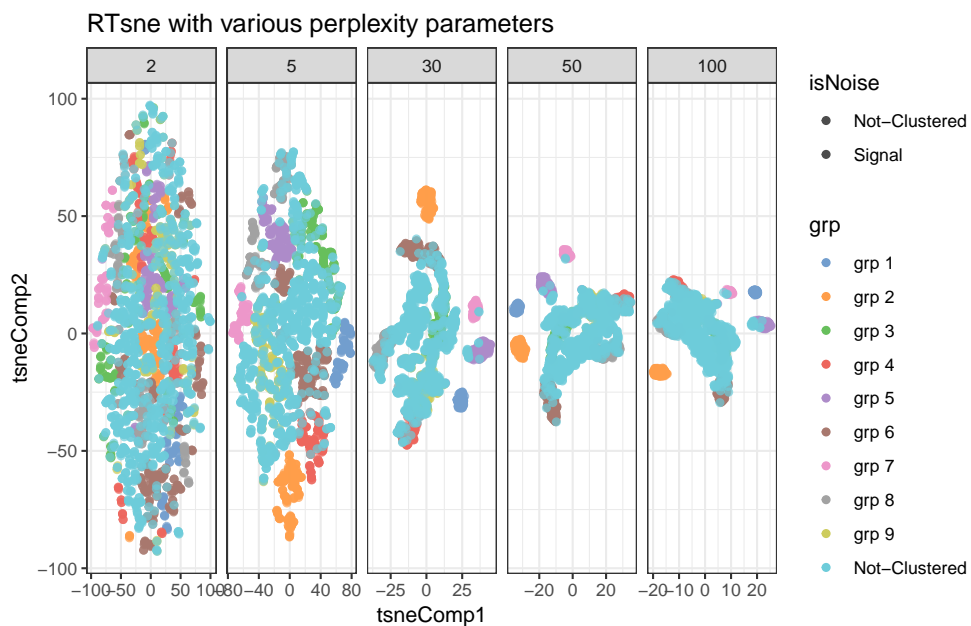


Figure B.12: Complex example with noise: t-SNE results with varying parameters

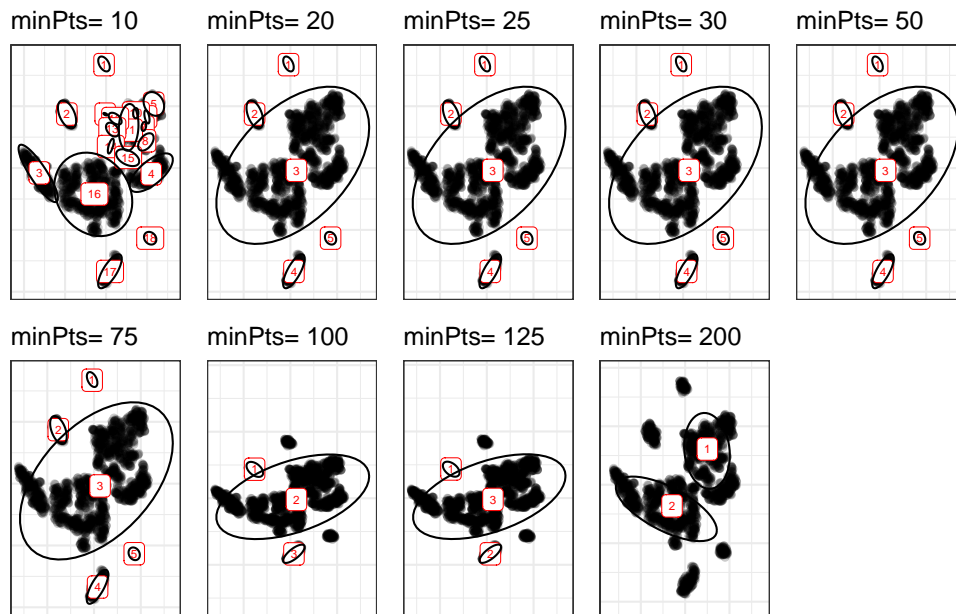


Figure B.13: Complex example with noise: chdbscan on dimensionally reduced data

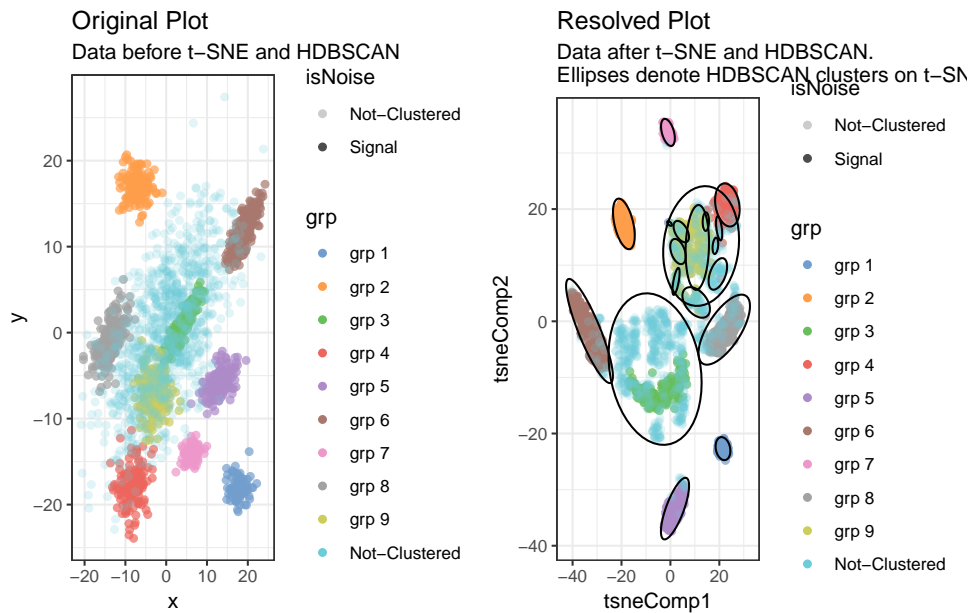


Figure B.14: Complex example with noise: clusters resolved by Adjutant. The black ellipses indicate the clusters that Adjutant would automatically select.

are totally overlapped by the noise all get clumped together. If the noise is diffuse it's possible that dense pockets of points within it are classifiable, but this is not the case here.

B.2.2 Investigating Adjutant with Real Data

We'll now obtain a real-world dataset, one where we do not know the ground truth, and to see what Adjutant would come up with. We'll take at articles pertaining to clinical and public health genomics genomic sequencing.

Downloading 20,000 articles + their metadata from using a single core takes about 15 minutes and then an additional 5 or some minutes for t-SNE and hdbscan to run. For this reason we've saved the analysis and do not run this code in the notebook.

```
df<-processSearch('(("whole genome"
                    OR "next generation"
                    OR "high throughput"))
                    AND "sequencing")
                    AND ("medicine" or "public health")',
                    retmax=20000)

#running the tidy corpus step
tidy_df<-tidyCorpus(corpus = df)

#running t-SNE
tsneObj<-runTSNE(tidy_df, check_duplicates=FALSE,
                 perplexity=100)

df<-inner_join(df, tsneObj$Y, by="PMID")

#running hdbscan
optClusters <- optimalParam(df)

save(df, tidy_df, tsneObj, optClusters,
      file="exampleGenomeDataAnalysis.Rdata")
```

Having run t-SNE, we can now explore how many different clusters Adjutant would suggest for this dataset for approximately 19 thousand documents.

```

load("exampleGenomeDataAnalysis.Rdata")
#let's look through the cluster changes via HDBSCAN
pList<-lapply(optClusters$altChoices,function(x){
  x$fitPlot +
    labs(title = paste("minPts=",x$minPt)) +
    theme(legend.position="none",
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          axis.title = element_blank())
})

cowplot::plot_grid(plotlist = pList, nrow = 2)

```

This figure shows that clusters forming toward the periphery of the t-SNE plot tend to be consistently grouped, and that the changing minPts parameter appears to have the greatest effect on documents clustered toward the center of the plot (similar to what was shown with the noisy synthetic data). More specifically as the minimum cluster size passed to hdbscan gets larger, we lose the ability to resolve clusters within the the middle blob. Why are those clusters in the middle blob resolvable at all? Its likely because the density within that blob is diffuse enough that smaller, but dense, clusters within it are still resolvable. But caution is warranted when interpreting clusters found within what appears to be a noisy blob.

Now we can take a look at the clusters Adjutant suggests for this data.

```

load("exampleGenomeDataAnalysis.Rdata")

#add the new cluster ID's the running dataset
df<-inner_join(df,optClusters$retItems,by="PMID") %>%
  mutate(tsneClusterStatus = ifelse(tsneCluster == 0,
                                    "not-clustered",
                                    "clustered"))

#now name the clusters
clustNames<-df %>%
  group_by(tsneCluster)%>%
  mutate(tsneClusterNames =
         getTopTerms(clustPMID = PMID,
                     clustValue=tsneCluster,

```

```

                                topNVal = 2,
                                tidyCorpus=tidy_df)) %>%
  select (PMID,tsneClusterNames) %>%
  ungroup ()

#update document corpus with cluster names
df<-inner_join(df,clustNames,by=c("PMID","tsneCluster"))

#re-plot the clusters

clusterNames <- df %>%
  dplyr::group_by(tsneClusterNames) %>%
  dplyr::summarise (medX = median (tsneComp1),
                    medY = median (tsneComp2)) %>%
  dplyr::filter (tsneClusterNames != "Not-Clustered")

ggplot (df,aes (x=tsneComp1,y=tsneComp2,group=tsneClusterNames)) +
  geom_point (aes (colour = tsneClusterStatus),alpha=0.2) +
  stat_ellipse (aes (alpha=tsneClusterStatus)) +
  geom_label (data=clusterNames,
              aes (x=medX,y=medY,label=tsneClusterNames),
              size=3,colour="red") +
  scale_colour_manual (values=c ("black","blue"),
                       name="cluster status") +
  scale_alpha_manual (values=c (1,0),
                      name="cluster status") +
  theme_bw ()

```

Validity of Clusters

Adjutant proposes many clusters for this document corpus, but how is it possible to establish the validity of these clusters without ground truth data? Here, we will look at the distribution of terms across t-SNE coordinates to assess the quality of the clustering. Again, while precise spatial positions from t-SNE should be cautiously interpreted, but we have shown that they are not all together irrelevant either and one would expect that some terms are unique to a specific cluster or set of (ideally) spatially proximal clusters.



Figure B.15: Clusters identified by Adjutant from a real data dataset

We use two methods to check the cluster validity. The first is a simple way looking at just individual words, and the second is a bit more complex that looks overall term frequencies between clusters.

Looking up specific terms

In this example we will assess the distribution of cluster names, since these reflect the top-two most common terms in the cluster itself. Ideally the top two terms

of a cluster name should occur primarily within the cluster of interest. It is crude metric but it's a useful start. We'll also include some common biology terms, specifically "cancer", "gene", "protein", "expression", "mutation", and "study" which we hypothesize should occur across nearly all of the clusters.

```
tidy_df_check <- dplyr::select(df,
                              PMID,
                              tsneComp1,
                              tsneComp2,
                              tsneClusterStatus,
                              tsneClusterNames) %>%

  inner_join(tidy_df)

#bag of terms
clustBag<-clusterNames %>%
  mutate(clusterWords = strsplit(tsneClusterNames, "-")) %>%
  tidyr::unnest(clusterWords)

clustBag<-unique(c(clustBag$clusterWords, c("cancer",
                                           "gene",
                                           "data",
                                           "protein",
                                           "expression",
                                           "mutat",
                                           "studi"))))

tmp<-df %>%
  filter(tsneClusterStatus == "clustered") %>%
  dplyr::select(PMID, tsneComp1, tsneComp2, tsneClusterNames)

tidy_df_check %>%
  filter(wordStemmed %in% clustBag) %>%
  filter(tsneClusterStatus == "clustered") %>%
  ggplot(aes(x=tsneComp1, y=tsneComp2, group=tsneClusterNames)) +
  stat_ellipse(data=tmp, aes(group=tsneClusterNames), col="red") +
  geom_point(alpha=0.2) +
  facet_wrap(~wordStemmed) +
  theme_bw() +
```

```
theme(legend.position="none",
      panel.grid.major= element_blank(),
      panel.grid.minor = element_blank(),
      axis.text = element_blank(),
      axis.title = element_blank(),
      panel.border = element_blank(),
      axis.ticks = element_blank(),
      strip.text = element_text(size=8))
```

In the above figure, each red circle represents a cluster and we've only drawn in the points for articles that contain a specific term. Although it is not very easy to read the individual terms, the pattern that is evident is that there are indeed some terms that tend to be present across all articles in the dataset and others that are present only specific articles within specific regions of the t-SNE plot. For example, “genom”, is a term that essentially present across all the documents in the corpus. Whereas terms like “outbreak”, “vaccine”, “lymphoma”, and even “tumor” tend to be present in specific areas. The clearest example of this separation is articles containing the words “genom”, “strain”, and/or “tumor”:

```
tidy_df_check %>%
  filter(wordStemmed %in% c("genom", "strain", "tumor")) %>%
  filter(tsneClusterStatus == "clustered") %>%
  ggplot(aes(x=tsneComp1, y=tsneComp2, group=tsneClusterNames)) +
  stat_ellipse(data=tmp, aes(group=tsneClusterNames), col="red") +
  geom_point(alpha=0.2) +
  facet_wrap(~wordStemmed) +
  theme_bw() +
  theme(legend.position="none",
        panel.grid.major= element_blank(),
        panel.grid.minor = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank(),
        panel.border = element_blank(),
        axis.ticks = element_blank(),
        strip.text = element_text(size=8))
```

In the above figure, the term “genom” occurs across all articles, while “strain” tends to be present in articles on the lower left quadrant and tumor tends to be present in

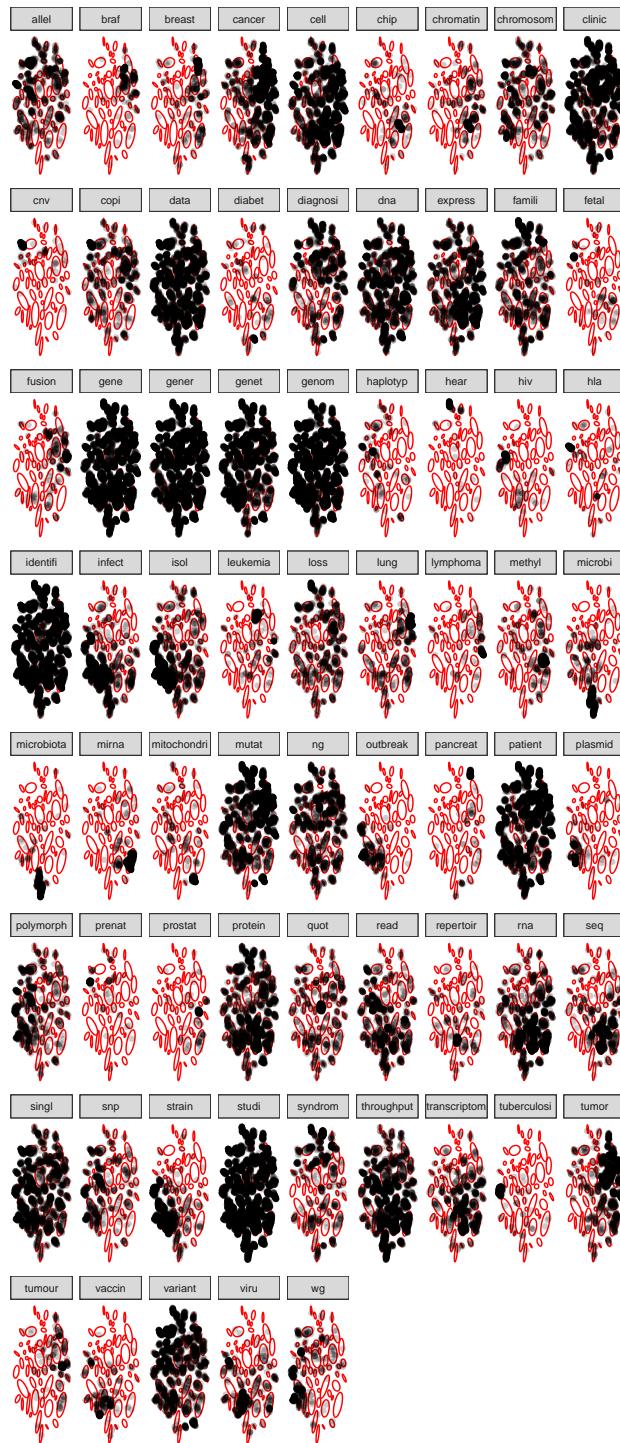


Figure B.16: Real data: distribution of terms across clusters. Red outlines indicate the cluster boundaries. Documents that contain specific terms are shown in black within each facet, while documents that do not contain those terms are not shown.

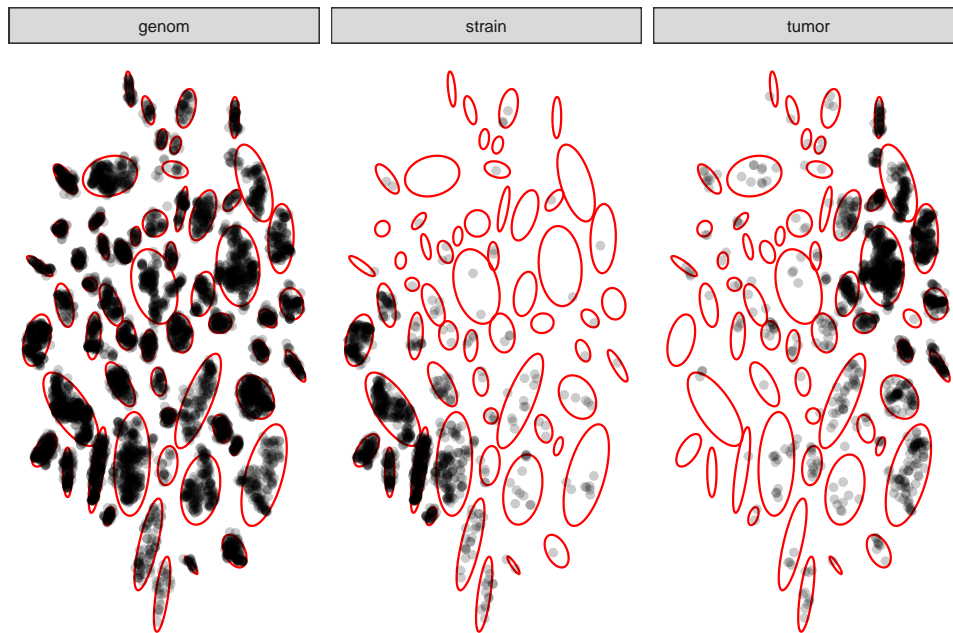


Figure B.17: Real data: distribution of terms across clusters II. Red outlines indicate the cluster boundaries. Documents that contain specific terms are shown in black within each facet, while documents that do not contain those terms are not shown. Close examination of the terms `genom*`, `strain*`, and `tumor`. The “`genom*`” term is distributed across all clusters, while the other two terms are limited to specific clusters.

articles of the upper right quadrant. This suggests only that there is some underlying “method of the madness” that is reasonable, not that the clusters are perfectly correct. But this is still a very simple example and only based on single word matches. What is more useful, and also more complex, is to compare clusters as “bags of words”, and that is precisely what we’ll do now.

Comparing term frequency

Using simple terms gives us a sense that there are indeed some relevant merits of the spatial clustering on t-SNE, however, I will now implement a more sophisticated approach that treats each cluster as a bag of words (derived from its documents) and compares derived term frequencies between clusters using the cosine similarity.

```

# quickly calculat the cosine similiarity
coss <- function(x) {
  crossprod(x) / (sqrt(tcrossprod(colSums(x^2))))
}

# looking at term frequency across clusters
# note that t-SNE is based on the tf_idf metric
# instead, we're comparing term frequencies across documents
clustDTM<-tidy_df_check %>%
  select(-tf,-idf,-tf_idf) %>% #
  group_by(tsneClusterNames,wordStemmed) %>%
  summarise(count = sum(n)) %>%
  ungroup() %>%
  bind_tf_idf(wordStemmed, tsneClusterNames, count) %>%
  cast_dtm(.,tsneClusterNames,wordStemmed,tf) %>%
  as.matrix()

simClust<-reshape::melt(coss(t(clustDTM))) %>% data.frame()
colnames(simClust)<-c("tsneClusterNames","compClust","cosSim")

# draw how similar clusters are to other clusters around
# them based upon the terms that occur within them

pList<-c()
for(clust in unique(as.character(simClust$tsneClusterNames))){
  tmp<-filter(simClust,compClust == clust) %>% select(-compClust)

  tmp<-inner_join(df,tmp) %>%
    mutate(isCluster = ifelse(tsneClusterNames == clust,1,0))%>%
    mutate(cosSimStep=cut(round(cosSim,2),seq(0,1,0.2)))

  p<- tmp %>%
    filter(tsneClusterNames != "Not-Clustered") %>%
    ggplot(aes(x=tsneComp1,
               y=tsneComp2,group=tsneClusterNames,
               colour = factor(isCluster),fill = cosSimStep))+
    stat_ellipse(geom="polygon")+
    scale_fill_brewer(breaks=c("(0,0.2]", "(0.2,0.4]",
                              "(0.4,0.6]", "(0.6,0.8]",

```

```

        "(0.8,1]"), drop=FALSE,
        name="Cosine Similarity")+
    scale_colour_manual(values=c("black", "red"),
        name="Selected Cluster",
        labels=c("other", clust))+
    labs(title=clust)+
    theme_bw()+
    theme(legend.position="none",
        panel.grid.major= element_blank(),
        panel.grid.minor = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank(),
        panel.border = element_blank(),
        axis.ticks = element_blank(),
        strip.text = element_text(size=8))

    pList[[clust]]<-p
}

pList[[1]] + theme(legend.position = "right")

```

Before considering the results of all possible clusters, we'll examine the results of just one to understand what is being shown. The above figure is the result for the cluster “cancer-breast”, and it compares this cluster (circled in red) with all other derived clusters in this data. The comparison is accomplished by visually overlaying the cosine similarity of all clusters relative to the “cancer-breast” cancer. The darker the blue the more similar the clusters; a cosine similarity of one reflects identical clusters, while a cosine similarity of 0 are totally different.

The above figure of the “cancer-breast” cluster shows what we'd kind of hoped for. First our cluster of interest (“cancer-breast” outlined in red) is, appropriately, a dark blue color (cosine similarity of 1). The clusters that are closest to it are not as dark blue, but appear to have a cosine similarity between 0.6 and 0.8 (related but not identical), and as we move further and further away from the cluster we see that the cosine similarity drops. This suggests that there is indeed some sanity both to spatial positions of the t-SNE clustering and also how hdbscan and Adjutant have

cancer-breast

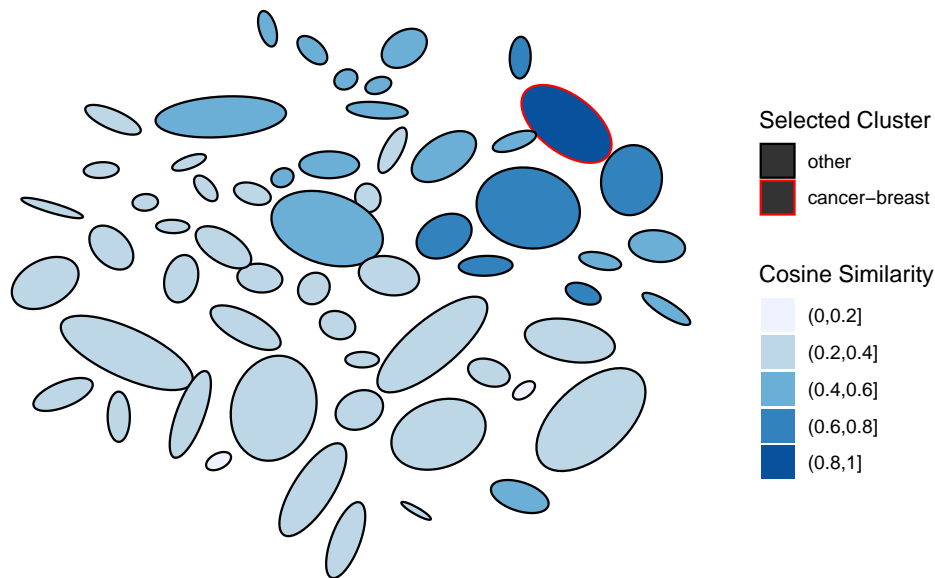


Figure B.18: Real data: distribution of terms across clusters III. Red outlines indicates the cluster with the term “cancer-breast”, while the color of the cluster indicates cosine similarity : white (low similarity) to dark blue (high similarity). Note that points that are spatially proximal to the “cancer-breast” cluster are more similar relative to distant clusters.

chosen the cluster boundaries.

Let’s look at all the other clusters now, to examine if the same trend occurs.

Looking at all of the clusters essentially recapitulates what’s in the previous single cluster figure. This means that t-SNE spatial positions are reasonable and driven by common terms in documents and are not arbitrarily placed. It’s also interesting to consider the “Noise” data (row 4, last column), which has high cosine similarity with pretty much every cluster - this is another nice sanity check as these articles had a difficult time clustering with any particular articles. It may be possible to use Adjutant’s derived clusters to classify those documents that have been assigned to the noise, or even to discover some fuzzy documents between clusters that are correctly not classified to one group or another. Adjutant doesn’t support that kind of clustering because it is up to a user what they’d like to do next with

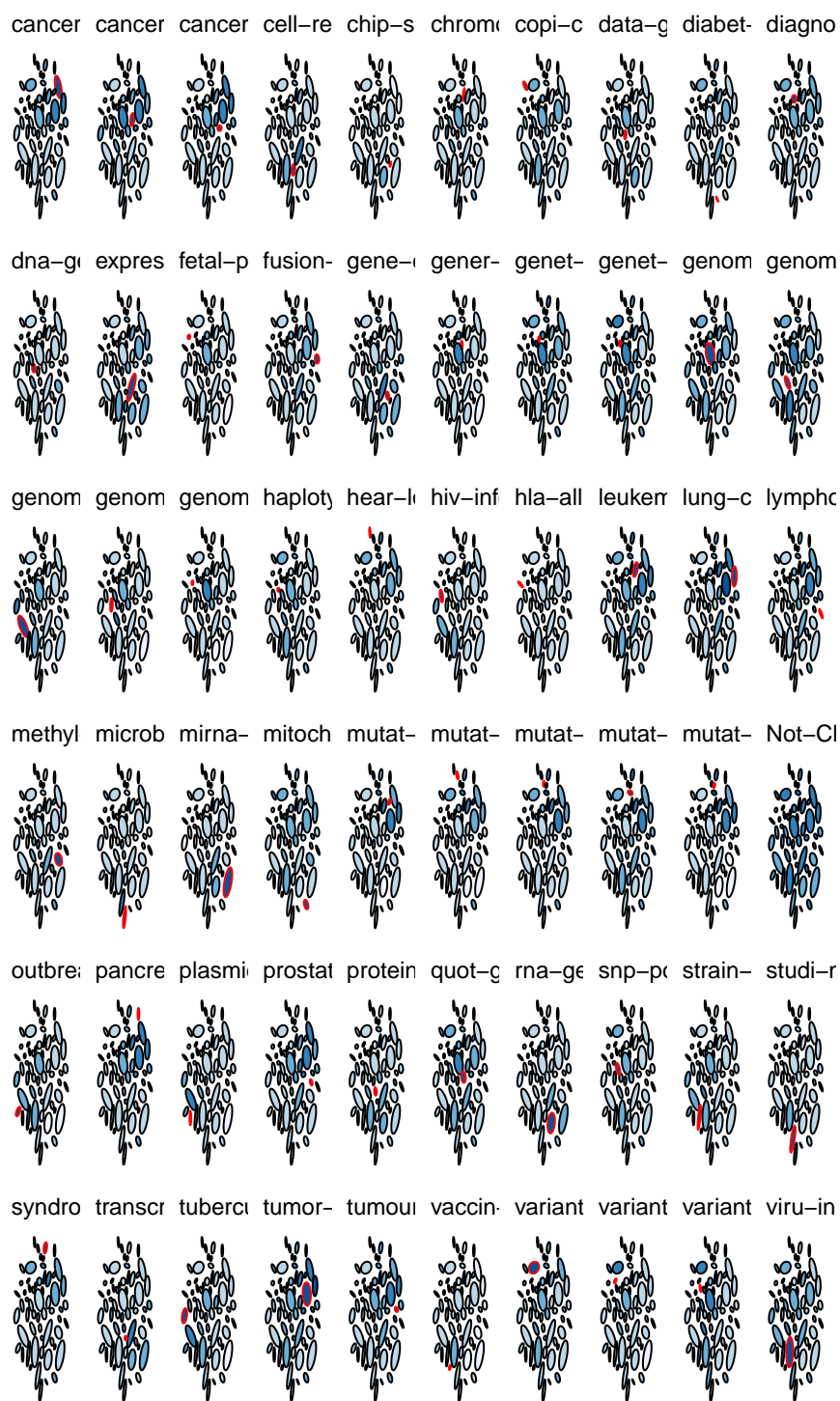


Figure B.19: Cluster similarity across all clusters.

analysis. However Adjutant's R compatible outputs allow the user to leverage to full complement of R's analytic tools to further explore this document corpus from a more informed initial point.

B.2.3 Alternative Approaches

Topic modeling is a broad and has had various advances in a number of different disciplines. Here we'll compared against a standard in R, which is the Latent Diriechilet Allocation (LDA) analysis presented in the tidytext manual. While many alternative approaches exist in R and beyond, we feel that is this is the most relevant comparison since we are developing an R based tool and use tidytext's features within Adjutant's development. One challenge of LDA, and many clustering methods, is that it's not very easy to establish the right initial parameters that should be provided to the method. Adjutant tries to scan for optimal parameters, but the best way to do this remains an active area of research. The Tidytext manual example begins with a priori knowledge that there are four books, and thus initializes the K parameter (number of clusters) to 4. But we don't know how many clusters we should have because we're taking a purely unsupervised approach. However, we do know that Adjutant would suggests around 60 clusters for this document corpus, so we can begin LDA with that. This is a nice example of the way that Adjutant can be used with other topic models.

Input to LDA will be a document term matrix using the tf.idf for analysis and an initial cluster size of 60. We will then look at the top 5 most common words within each cluster, shown below. The beta from the LDA analysis provide a sense of how important a word is to a particular topic cluster.

Running this step take a very long time on a document corpus of this size (> 10min). We have saved this analysis so that it can be quickly re-run here, but it emphasizes that rather large speed up we get with Adjutant's approach.

```
dtm<- tidy_df %>%  
  tidytext::cast_dtm(PMID, wordStemmed, n)  
  
dtm<-tidy_df_check %>%
```

```

ungroup() %>%
  cast_dtm(PMID, wordStemmed, n)

#suggested # of topics from Adjutant
adjClust<-length(unique(df$tsneClusterNames)) #60

df_lda <- LDA(dtm, k = 60, control = list(seed = 1234))

#Extracting topics
df_topics <- tidy(df_lda, matrix = "beta")

#top words within each topic
top_terms <- df_topics %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

#let's look at this as a tile plot
ggplot(top_terms, aes(x = topic, y = term, colour=beta))+
  geom_point()+
  theme_bw()+
  scale_colour_continuous(name=expression(beta))+
  scale_x_continuous(breaks=1:60)+
  theme(axis.text.y=element_text(size=8),
        axis.text.x=element_text(size=8, angle = 90, hjust=0.5))

```

From the previous figure, it is evident that some words belong to multiple topics – this is not surprising as we’ve already seen that terms like “genom” can be readily found across many documents. However, the beta values are all dark blue (closer to zero), implying that there is not a strong association between some particular word and a topic cluster. This is different than what the t-SNE analysis revealed, which showed that there indeed some individuals words that could be quite prominent in some clusters, or spatially proximal clusters, and absent or less present in other clusters.

It is also possible to analyze which topics individual articles belong to by using the gamma values from the LDA.

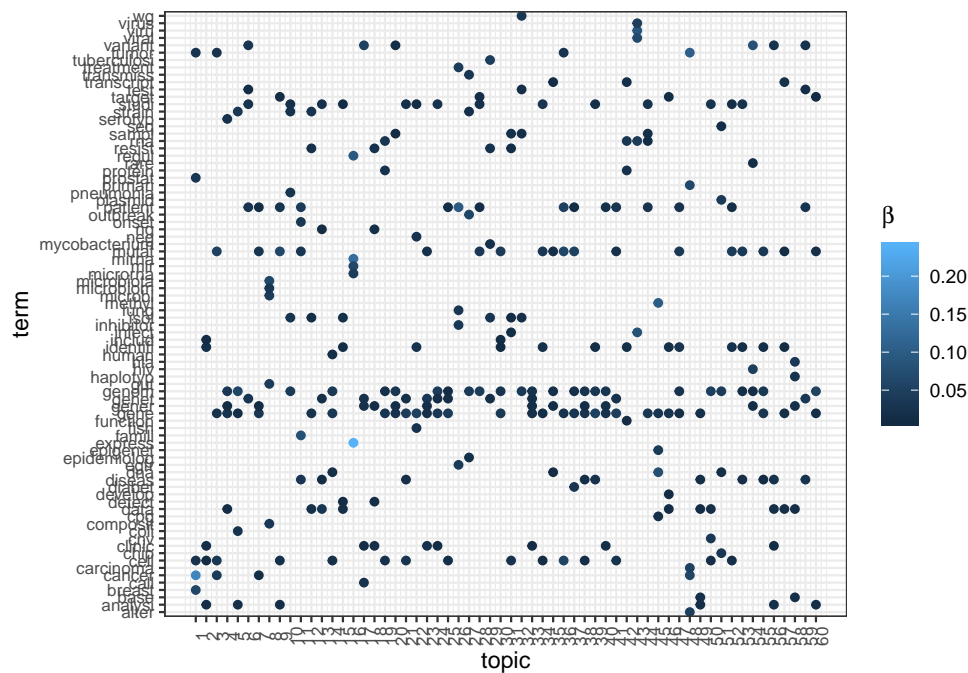


Figure B.20: Real data: LDA clusters and topics

```
df_gamma<- tidy(df_lda, matrix = "gamma")

#next we'll get the consensus topic for each article
df_classifications <- df_gamma %>%
  group_by(document) %>%
  top_n(1, gamma) %>%
  ungroup()

ggplot(df_gamma, aes(gamma)) +
  geom_histogram() +
  scale_y_log10() +
  labs(title = "Distribution of probabilities for all topics",
        y = "Number of documents", x = expression(gamma))
```

Ideally, this gamma distribution should go from 0 to 1, the results here suggests that it's hard for LDA to place documents within topics. With nearly 19,000 documents there is a huge range of potential values that can be passed to LDA's 'k' parameter

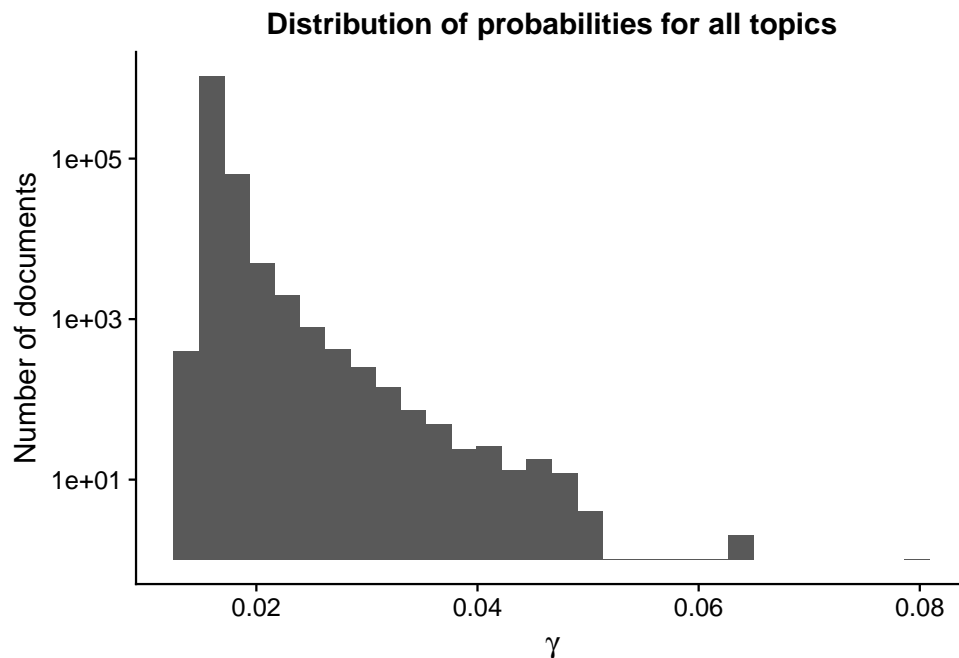


Figure B.21: Real data: LDA gamma distribution

function. The problem may be that the initial parameter value is not useful, or it may be that it's simply difficult for LDA to cluster these data – it's possible to try different alternatives but it quickly becomes cumbersome. One thing to note about the hdbscan procedure is that it allows some articles to not be classified whereas LDA actually tries to classify every document. This difference may be why Adjutant finds distinct clusters in this data and LDA does not.

We can also see that the combination of t-SNE and hbscan allowed us to reason about the topic clusters a little but more easily, since they can be visually inspected (as we showed earlier) and since “minimum cluster size” is a far easier parameter to reason about than “number of clusters”. It may be possible to improve the LDA results by trying different parameter combinations, cleaning the data in different ways, or using other packages. But herein lies the problem Adjutant is trying to address. The combination of t-SNE and hbscan do a reasonably good first pass, which can help an individual to reason about the next steps from a more information position.

Appendix C

GEViT Supplemental Materials

Contents

1. Supplemental Methods for Visualization Analysis
2. Supplemental Figures
3. Supplemental Tables

A reminder that analysis notebooks are also available at:

<https://github.com/amcrisan/GEViTAnalysisRelease>

C.1 Supplemental Methods for Visualization Analysis

We applied qualitative analysis techniques in order to consistently describe and compare aspects of our corpus of literature-derived data visualizations. We used a Grounded Theory approach, which refers to a general set of techniques used by qualitative researchers to inductively analyze and construct a theory about some phenomenon that is “grounded” in data [57]. Grounded Theory is conceptually similar to unsupervised analysis methods used in quantitative research [77], since both approaches rely on emergent pattern matching that is found within human-

curated and labelled data rather than applying a specific hypothesis or theory; in qualitative methods, the human resolves the relevant patterns, in quantitative methods, the algorithm does. 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 [20]. Two widely used approaches are open and axial coding. In open coding, text is read multiple times to identify emergent themes – these are captured as codes. In axial coding, a researcher develops 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 internal validity is a recognized concept within qualitative research and there exist agreed-upon conventions to assess this validity (see [70], Chapter 6), which we have employed here.

GTM is used in the field of information visualization (infovis), though we note that the application of GTM is different between the social sciences and human-computer interaction (HCI). HCI and infovis researchers frequently apply GTM to text [41], video, and image data [17], whereas social scientists tend to primarily use interview text, although some examples of image analysis with social sciences exist [65]. Our application of GTM, and especially open and axial coding, is drawn from the HCI and infovis research traditions, and we build upon established terminology and ideas from Munzner’s Visualization Analysis and Design [79]. As our team comprises primarily quantitative researchers, we apply a specific interrogative lens to the way we use GTM. There exists a fascinating and broader discussion about mixed methods approaches that best combine qualitative and quantitative research methods [22], which is beyond the application of this work but that the reader should be aware of.

C.2 Supplemental Figures

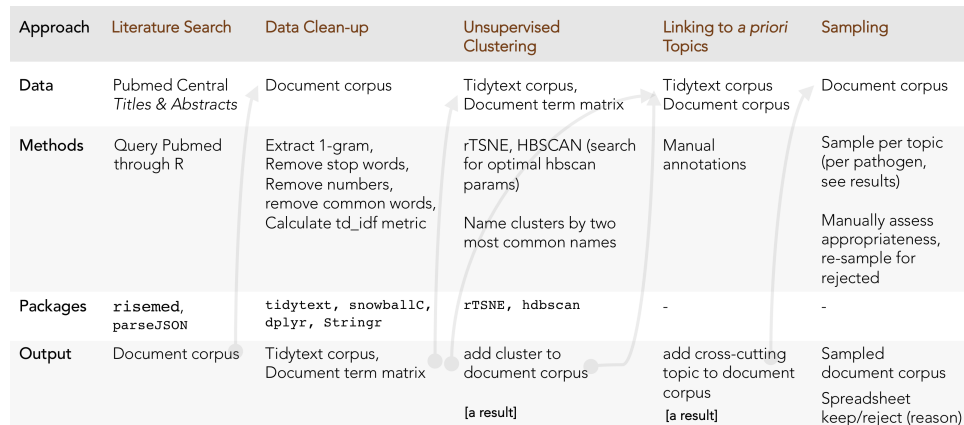


Figure C.1: Literature mining methods

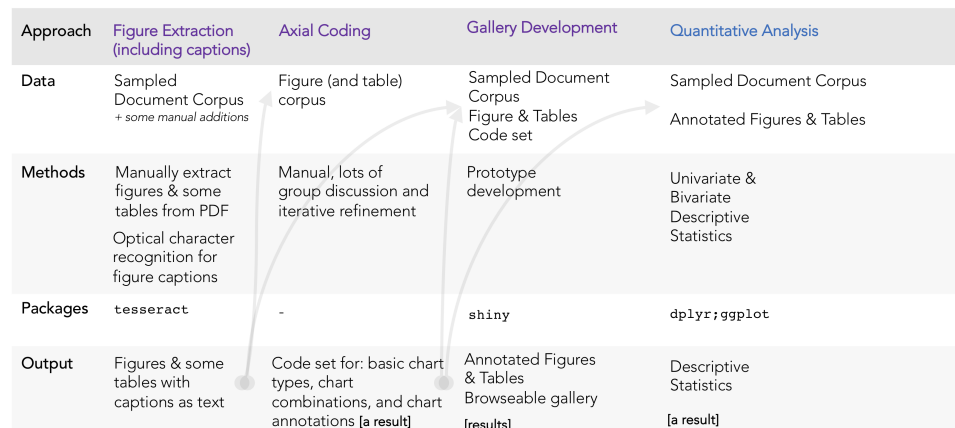


Figure C.2: Qualitative and quantitative visualization analysis methods

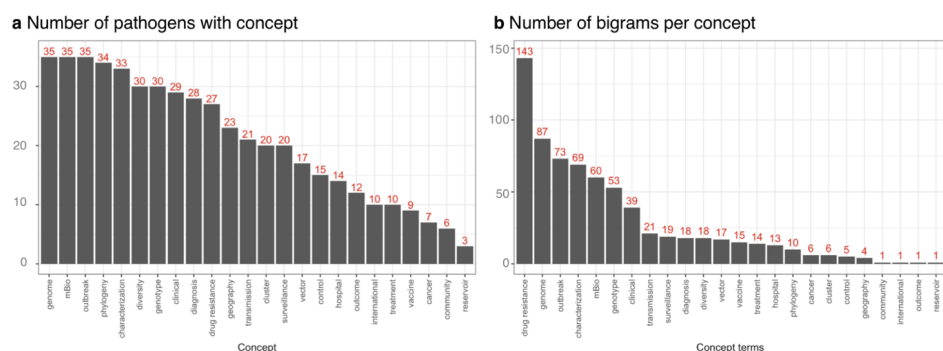


Figure C.3: A priori concepts distributed among pathogens (a) and the number of bigrams assigned to each concept (b)

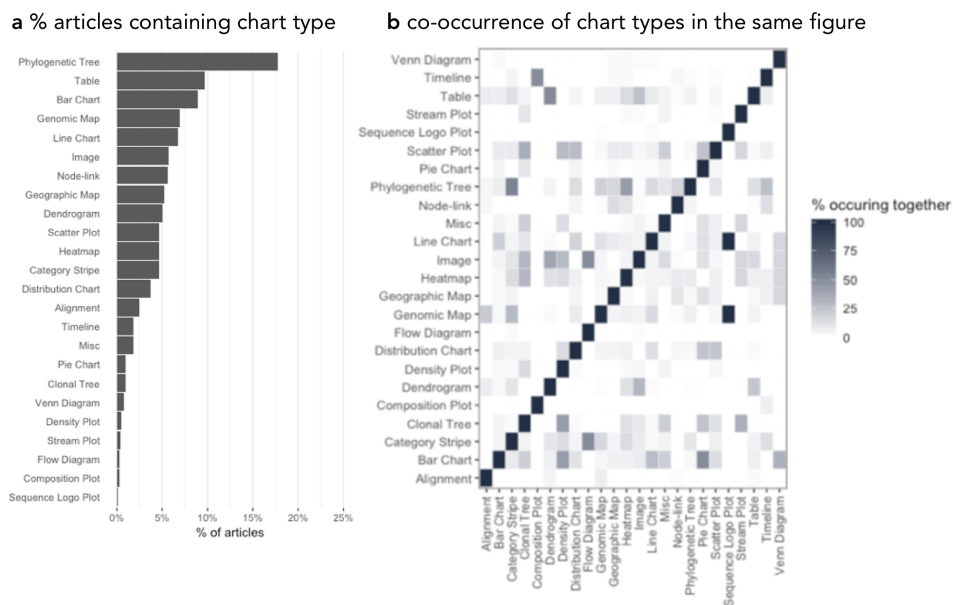


Figure C.4: Distribution of chart types across articles (a) and the co-occurrence of chart types with figures (b)

C.3 Supplemental Tables

Table C.1: 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.

Pathogen Bigram in Corpus	Pathogen
enceph_viru	Ignore
fever_viru	Ignore
respiratori_syndrom	Ignore
hemorrhag_fever	Ignore
porcin_reproduct	Ignore
porcin_epidem	Ignore
viru_pedv	Ignore
middl_east	Ignore
east_respiratori	Ignore
cov_infect	Ignore
escherichia_coli	Escherichia coli
mycobacterium_tuberculosi	Mycobacterium tuberculosis
staphylococcu_aureu	Staphylococcus aureus
influenza_viru	Influenza Virus
influenza_virus	Influenza Virus
vibrio_cholera	Vibrio cholerae
viru_hbv	Hepatitis B
immunodefici_viru	Human Immunodeficy Virus
human_immunodefici	Human Immunodeficy Virus
viru_hcv	Hepatitis C
salmonella_enterica	Salmonella Enterica
klebsiella_pneumonia	Klebsiella pneumonia
hiv_infect	Human Immunodeficy Virus
human_papillomaviru	Human Papillomavirus
hbv_infect	Hepatitis B
tuberculosi_isol	Mycobacterium tuberculosis
mrsa_isol	Staphylococcus aureus
coli_isol	Escherichia coli
acinetobact_baumannii	Acinetobact baumannii
dengu_viru	Dengue virus
pseudomona_aeruginosa	Pseudomona aeruginosa

Table C.1 continued from previous page

Pathogen Bigram in Corpus	Pathogen
produc_escherichia	Escherichia coli
aureu_isol	Staphylococcus aureus
sar_cov	SARS Coronavirus
helicobact_pylori	Helicobacter pylori
papillomaviru_hpv	Human Papillomavirus
enterococcu_faecium	Helicobact pylori
viru_hev	Hepatitis E
west_nile	West nile virus
nile_viru	West nile virus
neisseria_meningitidi	Neisseria meningitidis
yersinia_pesti	Yersinia pestis
sar_coronaviru	SARS Coronavirus
syndrom_coronaviru	SARS Coronavirus
coronaviru_sar	SARS Coronavirus
hpv_infect	Human Papillomavirus
cholera_isol	Vibrio cholerae
baumannii_isol	Acinetobact baumannii
vibrio_parahaemolyticu	Vibrio parahaemolyticus
aeruginosa_isol	Pseudomona aeruginosa
listeria_monocytogen	Listeria monocytogenes
clostridium_difficil	Clostridium difficile
produc_klebsiella	Klebsiella pneumonia
hev_infect	Hepatitis E
ebola_viru	Ebola virus
human_rotaviru	Human rotavirus
mer_cov	MERS coronavirus
viru_hav	Hepatitis A
viral_hepat	Heptatis - General
legionella_pneumophila	Legionella pneumophila
salmonella_typhimurium	Salmonella typhimurium
zika_viru	Zika virus
chlamydia_trachomati	Chlamydia trachomatis
coronaviru_mer	MERS coronavirus
viru_denv	Dengue virus
herp_simplex	Human herpesvirus
bacillu_anthraci	Bacillus anthracis

Table C.2: Mapping of Bigrams to *a priori* Concepts. Bigrams from the document corpus are in the first column, the occurrence of these Bigrams across clusters (n) and their frequency in the document corpus are also report. The final column contains author annotations that that assign bigram to *a priori* concepts, or indicates that a bigram is not-usefully assigned to any concept (i.e. because it is too general, belongs to something irrelevant, ect.)

bigram	n	freq	Annotate
acinetobact_baumannii	1	0.035714286	pathogen
acut_gastroenter	2	0.071428571	disease
acut_hepat	1	0.035714286	disease
acut_respiratori	2	NA	not-useful
aed_aegypti	1	0.035714286	reservoir
aeruginosa_isol	1	0.035714286	pathogen
ag_research	1	0.035714286	not-useful
allel_frequenc	1	0.035714286	population
amino_acid	7	NA	not-useful
analysi_mlva	1	0.035714286	genotype
analysi_reveal	1	NA	not-useful
anti_hcv	1	0.035714286	not-useful
antibiot_resist	4	0.142857143	resistance
antigen_hbsag	1	0.035714286	molecular biology
antimicrobi_resist	2	0.071428571	resistance
antimicrobi_suscept	1	0.035714286	resistance
associ_studi	2	0.071428571	not-useful
attenu_vaccin	1	0.035714286	vaccine
aureu_isol	1	0.035714286	pathogen
aureu_mrsa	1	0.035714286	resistance
aureu_strain	1	0.035714286	molecular biology
avian_influenza	1	0.035714286	reservoir
basal_core	1	0.035714286	disease
baumannii_clinic	1	0.035714286	pathogen
baumannii_isol	1	0.035714286	pathogen
baumannii_strain	1	0.035714286	characterization
beij_famili	1	NA	not-useful
beij_strain	1	NA	not-useful
beta_lactamas	1	0.035714286	resistance
bla_ctx	1	0.035714286	resistance
bla_oxa	1	0.035714286	resistance
breast_cancer	1	0.035714286	disease

Table C.2 continued from previous page

bigram	n	freq	Annotate
cancer_patient	1	0.035714286	disease
cancer_risk	1	0.035714286	disease
candid_gene	1	0.035714286	genome
capsid_protein	1	0.035714286	molecular biology
carbapenem_resist	2	0.071428571	resistance
carbapenemas_gene	1	0.035714286	resistance
carcinoma_hcc	1	0.035714286	disease
care_unit	1	0.035714286	outbreak
cassett_chromosom	1	0.035714286	molecular biology
cell_carcinoma	1	0.035714286	disease
cervic_cancer	1	0.035714286	disease
cervic_lesion	1	0.035714286	disease
chain_reaction	9	NA	not-useful
chain_reaction	5	0.178571429	not-useful
cholera_epidem	1	0.035714286	outbreak
cholera_isol	1	0.035714286	pathogen
cholera_outbreak	1	0.035714286	outbreak
cholera_pandem	1	0.035714286	outbreak
cholera_strain	1	0.035714286	characterization
cholera_toxin	1	0.035714286	molecular biology
chromosom_mec	1	0.035714286	resistance
chronic_hbv	1	0.035714286	disease
chronic_hcv	1	0.035714286	disease
chronic_hepat	3	0.107142857	disease
circoviru_type	1	0.035714286	characterization
circul_recombin	1	0.035714286	molecular biology
clinic_isol	5	0.178571429	not-useful
clinic_sampl	1	NA	not-useful
clonal_complex	1	0.035714286	molecular biology
close_relat	9	NA	not-useful
close_relat	7	0.25	not-useful
coli_isol	2	NA	NA
coli_isol	1	0.035714286	pathogen
coli_stec	1	0.035714286	resistance
coli_strain	1	0.035714286	characterization
colorect_cancer	1	0.035714286	disease
commun_acquir	1	0.035714286	outbreak

Table C.2 continued from previous page

bigram	n	freq	Annotate
compar_genom	1	NA	genome
complet_genom	5	NA	genome
complet_genom	4	0.142857143	genome
confid_interv	2	0.071428571	not-useful
core_promot	1	0.035714286	molecular biology
coronaviru_mer	1	0.035714286	pathogen
coronaviru_sar	1	0.035714286	pathogen
cov_genom	1	0.035714286	genome
cov_infect	1	0.035714286	pathogen
dengu_epidem	1	0.035714286	outbreak
dengu_fever	1	0.035714286	disease
dengu_infect	1	0.035714286	pathogen
dengu_outbreak	1	0.035714286	outbreak
dengu_viru	1	0.035714286	pathogen
dengu_virus	1	0.035714286	pathogen
denv_serotyp	1	0.035714286	clinical
develop_countri	1	0.035714286	not-useful
diarrhea_viru	1	0.035714286	disease
discriminatory_power	1	0.035714286	not-useful
divers_index	1	0.035714286	population
dna_level	1	0.035714286	not-useful
drug_resist	3	NA	resistance
drug_suscept	1	0.035714286	resistance
drug_user	1	0.035714286	not-useful
east_respiratori	1	0.035714286	disease
electrophoresi_pfge	4	0.142857143	not-useful
enterica_serotyp	1	0.035714286	clinical
enterica_serovar	1	0.035714286	clinical
enterica_subsp	1	0.035714286	molecular biology
enterococcu_faecium	1	0.035714286	pathogen
epidem_diarrhea	1	0.035714286	disease
epidemiolog_investig	1	0.035714286	outbreak
epidemiolog_studi	1	0.035714286	not-useful
esbl_produc	1	0.035714286	resistance
escherichia_coli	2	0.071428571	pathogen
extend_spectrum	1	0.035714286	resistance
field_gel	8	NA	not-useful

Table C.2 continued from previous page

bigram	n	freq	Annotate
form_crf	1	NA	not-useful
fragment_length	1	0.035714286	not-useful
gastric_cancer	1	0.035714286	disease
gastroenter_outbreak	1	0.035714286	outbreak
gel_electrophoresi	8	NA	not-useful
gene_encode	1	0.035714286	genome
gene_express	1	0.035714286	genome
gene_segment	2	0.071428571	genome
genet_divers	9	NA	population
genet_divers	7	0.25	genetic-diversity
genet_factor	1	0.035714286	genome
genet_variant	1	0.035714286	genome
genet_variat	1	0.035714286	genome
genom_constel	1	0.035714286	genome
genom_copi	1	NA	not-useful
genom_epidemiologi	1	0.035714286	not-useful
genom_island	1	0.035714286	molecular biology
genom_segment	1	0.035714286	not-useful
genom_sequenc	24	NA	not-useful
genom_sequenc	22	0.785714286	not-useful
genom_wide	4	NA	not-useful
genom_wide	3	0.107142857	genome
genotyp_constel	1	0.035714286	genotype
genotyp_gii	1	0.035714286	genotype
gii_gii	1	0.035714286	molecular biology
hand_foot	1	0.035714286	not-useful
hav_genom	1	0.035714286	genome
hav_rna	1	NA	clinical
hbeag_neg	1	0.035714286	clinical
hbsag_posit	1	0.035714286	clinical
hbv_carrier	1	0.035714286	clinical
hbv_dna	1	0.035714286	not-useful
hbv_genom	1	0.035714286	genome
hbv_genotyp	1	0.035714286	genotype
hbv_infect	1	0.035714286	pathogen
hbv_strain	1	0.035714286	characterization
hcv_antibodi	1	0.035714286	molecular biology

Table C.2 continued from previous page

bigram	n	freq	Annotate
hcv_genom	1	0.035714286	genome
hcv_genotyp	1	0.035714286	genotype
hcv_infect	1	0.035714286	disease
hcv_rna	1	0.035714286	molecular biology
helicobact_pylori	1	0.035714286	pathogen
hemolyt_urem	1	0.035714286	disease
hepatocellular_carcinoma	2	0.071428571	disease
hev_infect	1	0.035714286	pathogen
hev_strain	1	0.035714286	characterization
highli_pathogen	1	0.035714286	not-useful
hiv_infect	1	0.035714286	pathogen
hiv_type	1	0.035714286	characterization
hospit_children	1	NA	outbreak
hpv_dna	1	0.035714286	not-useful
hpv_genom	1	0.035714286	genome
hpv_genotyp	1	0.035714286	genotype
hpv_infect	1	0.035714286	pathogen
hpv_neg	1	0.035714286	clinical
hpv_posit	1	0.035714286	clinical
hpv_type	1	0.035714286	characterization
human_bocaviru	1	NA	pathogen
human_immunodefici	1	0.035714286	disease
human_influenza	1	0.035714286	disease
human_metapneumoviru	1	NA	not-useful
human_noroviru	1	0.035714286	pathogen
human_papilloma	1	0.035714286	pathogen
human_papillomaviru	1	0.035714286	pathogen
human_papillomavirus	1	0.035714286	pathogen
human_rotaviru	1	0.035714286	pathogen
immun_respons	3	0.107142857	not-useful
immunodefici_viru	1	0.035714286	pathogen
increas_risk	1	0.035714286	not-useful
infect_individu	1	0.035714286	not-useful
infect_patient	2	0.071428571	not-useful
influenza_viru	1	0.035714286	pathogen
influenza_virus	1	0.035714286	pathogen
insert_sequenc	1	0.035714286	molecular biology

Table C.2 continued from previous page

bigram	n	freq	Annotate
intens_care	1	0.035714286	outbreak
interspers_repetit	1	0.035714286	genotype
klebsiella_pneumonia	1	0.035714286	pathogen
lactamas_esbl	1	0.035714286	resistance
lactamas_produc	1	0.035714286	resistance
length_genom	1	0.035714286	not-useful
length_polymorph	1	0.035714286	not-useful
live_attenu	1	0.035714286	vaccine
liver_diseas	3	0.107142857	disease
locu_variabl	1	0.035714286	not-useful
locu_vntr	1	0.035714286	genotype
lower_respiratori	1	NA	disease
mec_sccmec	1	0.035714286	resistance
mer_cov	1	0.035714286	pathogen
meta_analysi	1	0.035714286	not-useful
metapneumoviru_hmpv	1	NA	pathogen
methicillin_resist	1	0.035714286	resistance
methicillin_suscept	1	0.035714286	resistance
middl_east	1	0.035714286	not-useful
miru_vntr	1	0.035714286	genotype
mlva_genotyp	1	0.035714286	genotype
mlva_method	1	0.035714286	not-useful
mlva_profil	1	0.035714286	genotype
mlva_type	1	0.035714286	genotype
molecular_character	1	0.035714286	not-useful
molecular_epidemiologi	9	NA	not-useful
molecular_epidemiologi	6	0.214285714	not-useful
molecular_type	1	0.035714286	not-useful
mouth_diseas	1	0.035714286	disease
mrsa_clone	1	0.035714286	resistance
mrsa_isol	1	0.035714286	pathogen
mrsa_strain	1	0.035714286	characterization
multi_locu	1	0.035714286	not-useful
multidrug_resist	5	0.178571429	resistance
multilocu_sequenc	5	0.178571429	not-useful
multilocu_variabl	1	0.035714286	not-useful
multipl_locu	1	0.035714286	not-useful

Table C.2 continued from previous page

bigram	n	freq	Annotate
mycobacteri_interspers	1	0.035714286	genotype
mycobacterium_tuberculosi	1	0.035714286	pathogen
nasopharyng_aspir	1	NA	not-useful
neutral_antibodi	1	NA	not-useful
noroviru_genom	1	0.035714286	genome
noroviru_genotyp	1	0.035714286	genotype
noroviru_gii	1	0.035714286	characterization
noroviru_infect	1	0.035714286	pathogen
noroviru_nov	1	0.035714286	pathogen
noroviru_outbreak	1	0.035714286	outbreak
noroviru_strain	1	0.035714286	characterization
nosocomi_infect	1	0.035714286	disease
nucleotid_polymorph	4	NA	not-useful
nucleotid_polymorph	3	0.107142857	not-useful
nucleotid_sequenc	7	NA	not-useful
nucleotid_sequenc	5	0.178571429	not-useful
odd_ratio	3	0.107142857	not-useful
pandem_influenza	1	0.035714286	outbreak
panton_valentin	1	0.035714286	resistance
papilloma_viru	1	0.035714286	pathogen
papillomaviru_hpv	1	0.035714286	pathogen
papillomaviru_type	1	0.035714286	characterization
papillomavirus_hpv	1	0.035714286	pathogen
pathogen_avian	1	0.035714286	reservoir
pathogen_island	1	0.035714286	molecular biology
pcr_assai	1	NA	not-useful
pcr_method	1	NA	not-useful
pedv_strain	1	0.035714286	characterization
phage_type	1	0.035714286	characterization
phylogenet_analysi	14	NA	population
phylogenet_analysi	11	0.392857143	phylogeny
pig_farm	1	0.035714286	zoonotic
plasmid_mediat	1	NA	resistance
pneumonia_isol	1	0.035714286	disease
pneumonia_strain	1	0.035714286	characterization
polymeras_chain	8	NA	not-useful
polymeras_chain	5	0.178571429	not-useful

Table C.2 continued from previous page

bigram	n	freq	Annotate
polymorph_rflp	1	0.035714286	genotype
polymorph_snp	1	0.035714286	genotype
popul_structur	1	NA	not-useful
porcin_circoviru	1	0.035714286	reservoir
porcin_epidem	1	0.035714286	pathogen
porcin_reproduct	1	0.035714286	pathogen
posit_sampl	1	NA	not-useful
produc_escherichia	1	0.035714286	pathogen
produc_klebsiella	1	0.035714286	pathogen
promot_bcp	1	0.035714286	resistance
prostat_cancer	1	0.035714286	disease
pseudomona_aeruginosa	1	0.035714286	pathogen
public_health	2	NA	population
public_health	1	0.035714286	not-useful
puls_field	8	NA	not-useful
puls_field	7	0.25	not-useful
rapid_detect	1	NA	not-useful
reaction_pcr	2	NA	not-useful
reaction_pcr	1	0.035714286	not-useful
read_frame	1	0.035714286	not-useful
real_time	2	NA	not-useful
real_time	1	0.035714286	not-useful
reassort_event	1	0.035714286	molecular biology
recombin_form	1	0.035714286	molecular biology
repeat_analysi	1	0.035714286	not-useful
repeat_vntr	2	NA	genotype
repeat_vntr	1	0.035714286	genotype
repetit_unit	1	0.035714286	genotype
resist_acinetobact	1	0.035714286	resistance
resist_determin	3	NA	resistance
resist_determin	2	0.071428571	resistance
resist_enterococci	1	0.035714286	resistance
resist_enterococcu	1	0.035714286	resistance
resist_gene	5	0.178571429	resistance
resist_isol	1	0.035714286	not-useful
resist_klebsiella	1	NA	resistance
resist_mdr	1	0.035714286	resistance

Table C.2 continued from previous page

bigram	n	freq	Annotate
resist_salmonella	1	0.035714286	resistance
resist_staphylococcu	1	0.035714286	resistance
resist_tuberculosi	1	0.035714286	resistance
respiratori_infect	1	NA	disease
respiratori_syncyti	1	NA	not-useful
respiratori_syndrom	2	0.071428571	disease
respiratori_tract	1	NA	not-useful
restrict_fragment	1	0.035714286	clinical
revers_transcript	4	NA	molecular biology
revers_transcript	2	0.071428571	not-useful
risk_factor	4	NA	not-useful
risk_factor	3	0.107142857	not-useful
risk_hpv	1	0.035714286	clinical
risk_human	1	0.035714286	not-useful
rotaviru_infect	1	0.035714286	pathogen
rotaviru_strain	1	0.035714286	characterization
rotaviru_vaccin	1	0.035714286	vaccine
rva_strain	1	0.035714286	characterization
salmonella_enterica	1	0.035714286	pathogen
salmonella_genom	1	0.035714286	genome
salmonella_isol	1	0.035714286	pathogen
salmonella_serovar	1	0.035714286	reservoir
salmonella_strain	1	0.035714286	characterization
salmonella_typhimurium	1	0.035714286	pathogen
sar_coronaviru	1	0.035714286	pathogen
sar_cov	1	0.035714286	pathogen
sar_epidem	1	0.035714286	outbreak
sar_outbreak	1	0.035714286	outbreak
sccmec_type	1	0.035714286	characterization
sequenc_analysi	5	NA	not-useful
sequenc_analysi	4	0.142857143	not-useful
sequenc_type	8	NA	not-useful
sequenc_type	7	0.25	not-useful
serovar_enteritidi	1	0.035714286	reservoir
serovar_typhimurium	1	0.035714286	reservoir
serum_sampl	2	0.071428571	not-useful
seventh_pandem	1	0.035714286	not-useful

Table C.2 continued from previous page

bigram	n	freq	Annotate
sever_acut	1	0.035714286	not-useful
shiga_toxin	1	0.035714286	molecular biology
signific_associ	1	0.035714286	not-useful
singl_nucleotid	4	NA	not-useful
singl_nucleotid	3	0.107142857	not-useful
spa_type	1	0.035714286	genotype
spectrum_beta	1	0.035714286	resistance
squamou_cell	1	0.035714286	disease
staphylococc_cassett	1	0.035714286	molecular biology
staphylococcu_aureu	1	0.035714286	pathogen
statist_signific	1	0.035714286	not-useful
stool_sampl	3	NA	not-useful
stool_sampl	2	0.071428571	not-useful
strain_circul	1	0.035714286	epidemiology
strain_detect	1	0.035714286	not-useful
strain_isol	7	NA	not-useful
strain_isol	6	0.214285714	not-useful
studi_design	1	NA	not-useful
studi_gwa	1	0.035714286	not-useful
subsp_enterica	1	0.035714286	characterization
surfac_antigen	1	0.035714286	molecular biology
syncyti_viru	1	NA	not-useful
syndrom_coronaviru	1	0.035714286	disease
syndrom_sar	1	0.035714286	disease
syndrom_viru	1	0.035714286	not-useful
tandem_repeat	3	0.107142857	genotype
tandem_repeat	2	NA	not-useful
time_pcr	1	NA	not-useful
time_revers	1	NA	not-useful
tor_biotyp	1	0.035714286	pathogen
tor_strain	1	0.035714286	characterization
toxin_produc	1	0.035714286	not-useful
tract_infect	1	NA	not-useful
transcript_polymeras	1	NA	not-useful
tuberculosi_complex	1	0.035714286	characterization
tuberculosi_isol	1	0.035714286	pathogen
tuberculosi_strain	1	0.035714286	characterization

Table C.2 continued from previous page

bigram	n	freq	Annotate
type_method	1	0.035714286	not-useful
type_mlst	2	0.071428571	genotype
typhimurium_isol	1	0.035714286	pathogen
uniqu_recombin	1	NA	not-useful
unit_variabl	1	0.035714286	not-useful
urem_syndrom	1	0.035714286	disease
vaccin_candid	1	0.035714286	vaccine
vaccin_develop	1	0.035714286	vaccine
vaccin_strain	1	0.035714286	characterization
valentin_leukocidin	1	0.035714286	resistance
vancomycin_resist	1	0.035714286	resistance
vibrio_cholera	1	0.035714286	pathogen
viral_gastroenter	1	0.035714286	not-useful
viral_genom	5	NA	not-useful
viral_genom	4	0.142857143	genome
viral_hepat	1	0.035714286	pathogen
viral_load	2	0.071428571	clinical
viru_denv	1	0.035714286	pathogen
viru_genotyp	2	0.071428571	genotype
viru_hav	1	0.035714286	pathogen
viru_hbv	1	0.035714286	pathogen
viru_hcv	1	0.035714286	pathogen
viru_hev	1	0.035714286	pathogen
viru_infect	3	0.107142857	not-useful
viru_isol	4	NA	not-useful
viru_isol	3	0.107142857	not-useful
viru_pedv	1	0.035714286	pathogen
viru_prsv	1	0.035714286	pathogen
viru_serotyp	1	0.035714286	clinical
viru_type	2	0.071428571	characterization
virul_factor	1	0.035714286	molecular biology
virul_gene	2	0.071428571	genome
virus_isol	1	0.035714286	not-useful
vntr_analysi	1	0.035714286	genotype
vntr_loci	1	0.035714286	genotype
vntr_type	1	0.035714286	genotype
wide_associ	2	0.071428571	not-useful

Table C.2 continued from previous page

bigram	n	freq	Annotate
wide_signific	1	0.035714286	not-useful
wild_type	2	0.071428571	not-useful

Table C.3: Master List of Sampled Articles

This table is too large to print and is available online at:
<https://doi.org/10.1093/bioinformatics/bty832>

Table C.4: Final set of pathogens and pathogen clusters

Pathogen	Pathogen Cluster
Enterovirus D68	Other
Acinetobact baumannii	Acinetobact baumannii
Bacillus anthracis	Other
Clostridium difficile	Other
Dengue virus	Dengue virus
Ebola virus	Other
Enterococcus faecium	Enterococcus faecium
Escherichia coli	Escherichia coli
Helicobacter pylori	Other
Hepatitis B	Hepatitis B
Hepatitis C	Hepatitis C
Hepatitis E	Other
Human herpesvirus	Other
Human Immunodeficy Virus	Human Immunodeficy Virus
Human Papillomavirus	Human Papillomavirus
Human rotavirus	Other
Influenza Virus	Influenza Virus
Klebsiella pneumonia	Klebsiella pneumonia
Legionella pneumophila	Other
Listeria monocytogenes	Other
MERS coronavirus	Other
Microbiota	Microbiota
Mycobacterium tuberculosis	Mycobacterium tuberculosis
Neisseria gonorrhoeae	Other
Neisseria meningitidis	Other
Pseudomona aeruginosa	Pseudomona aeruginosa
Salmonella Enterica	Salmonella Enterica
Salmonella typhimurium	Other
SARS Coronavirus	SARS Coronavirus
Staphylococcus aureus	Staphylococcus aureus
Vibrio cholerae	Vibrio cholerae
Vibrio parahaemolyticus	Other

Table C.4 continued from previous page

Pathogen	Pathogen Cluster
Zika virus	Other

Appendix D

minCombinR Supplemental Materials

This document demonstrates how to implement and plot different chart types using minCombinR. This document assumes that you have already run the “Getting started with minCombinR” and that the necessary data has already been loaded into your R workspace.

D.1 Generating Simple Charts with minCombinr

```
devtools::load_all()
library(dplyr)
library(shiny)

# Tabular Data
tab_dat <- input_data(file = system.file("extdata",
                                          "ebov_metadata.csv",
                                          package = "mincombinr"),
                      dataType = "table")

# Tree data
tree_dat<-input_data(file = system.file("extdata",
                                         "ebov_tree.nwk",
```

```

                                package = "mincombinr"),
                                dataType = "tree")
# Genomic data
genomic_dat<-input_data(file = system.file("extdata",
                                "ebov_GIN_genomic_FIXED.fasta",
                                package = "mincombinr"),
                                dataType = "dna")

#Shape files
#Shape files require that .shp,.shx,and .prj
# files at a minimun to be in the same directory
#to add metadata to the shape file, you can also add .dbf files
gin_file<-"gin_admbnda_adml_ocha_itos.shp"
lbr_file<-"lbr_admbnda_adml_ocha.shp"
sle_file<-"sle_admbnda_adml_lm_gov_ocha_20161017.shp"

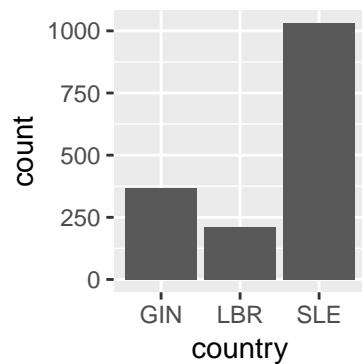
gin_shape_dat<-input_data(file =
                                system.file("./inst/extdata/",
                                gin_file,
                                package = "mincombinr"),
                                dataType = "spatial")
lbr_shape_dat<-input_data(file =
                                system.file("./inst/extdata/",
                                lbr_file,
                                package = "mincombinr"),
                                dataType = "spatial")
sle_shape_dat<-input_data(file =
                                system.file("extdata/",
                                sle_file,
                                package = "mincombinr")
                                ,dataType = "spatial")

# Put all the individual shape files together so that the system
# the system knows to try visualize it all together.
shape_dat <- join_spatial_data(gin_shape_dat, lbr_shape_dat
                                , sle_shape_dat)

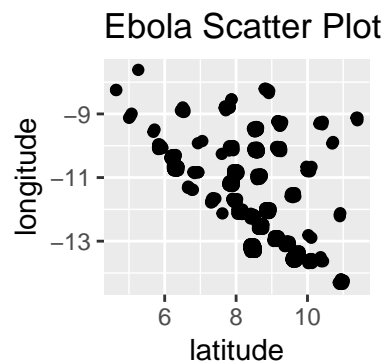
```

D.1.1 Common Statistical Charts

```
# Let's specify and plot some single charts.  
  
# Bar chart:  
bar_chart <- specify_single(chart_type = "bar",  
                             data = "tab_dat",  
                             x = "country")  
  
plot(bar_chart)
```

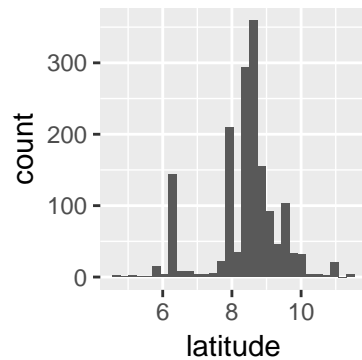


```
# Scatter plot with a title:  
scatter_chart <- specify_single(chart_type = "scatter",  
                                 data = "tab_dat",  
                                 x = "latitude",  
                                 y = "longitude",  
                                 title = "Ebola Scatter Plot")  
  
plot(scatter_chart)
```



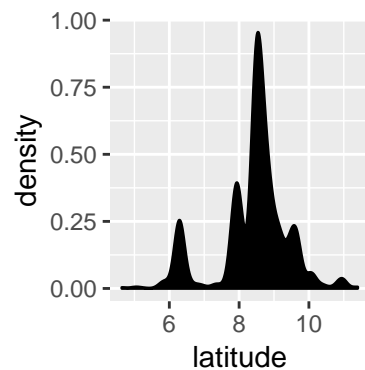
```
# Histogram:
histogram_chart <- specify_single(chart_type = "histogram",
                                   data = "tab_dat",
                                   x = "latitude")

plot(histogram_chart)
```



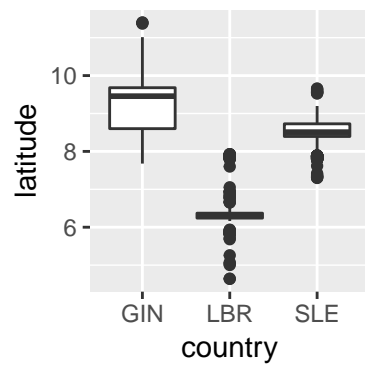
```
# Probability Density Function (PDF) plot:
pdf_chart <- specify_single(chart_type = "pdf",
                             data = "tab_dat",
                             x = "latitude")

plot(pdf_chart)
```



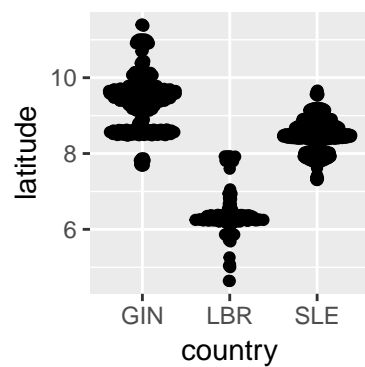
```
# Boxplot
boxplot_chart <- specify_single(chart_type = "boxplot",
                                 data="tab_dat",
                                 x = "country",
                                 y = "latitude")

plot(boxplot_chart)
```



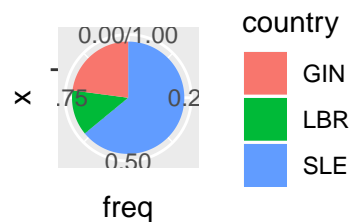
```
# Swarm plot
beeswarm_chart <- specify_single(chart_type = "swarmplot",
                                data="tab_dat",
                                x = "country",
                                y = "latitude")

plot(beeswarm_chart)
```



```
# We'll even let you make a pie chart
pie_chart <- specify_single(chart_type = "pie", data = "tab_dat",
                             x = "country")

plot(pie_chart)
```



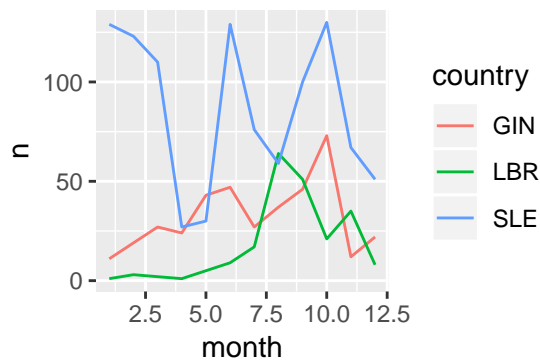
minCombinR can also work with data frames so you can perform some

analyses and go ahead and plot the data. Here's an example using a line chart:

```
# Let's use our tabular data:
ebov <- tab_dat@data[[1]]
ebov_case_counts <- ebov %>%
  group_by(country, month) %>%
  count()

# Now let's specify and plot a line chart:
line_chart <- specify_single(chart_type = "line",
                             data = "ebov_case_counts",
                             x = "month", y = "n",
                             group = "country")

plot(line_chart)
```



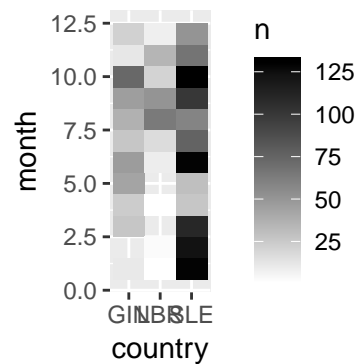
D.1.2 Colour Charts

Color charts are common statistical charts that fundamentally require color to communicate their results. By comparison, adding color to a common statistical chart can be seen as an enhancement (a nice to have, not a need to have).

```
# Get our data:
ebov <- tab_dat@data[[1]]
ebov_heat_data <- ebov %>%
  group_by(country, month) %>%
  count()
```

```
# Specify and plot a line chart:
heatmap_chart <- specify_single(chart_type = "heatmap",
                                data = "ebov_heat_data",
                                x = "country",
                                y = "month",
                                color = "n")

plot(heatmap_chart)
```



D.1.3 Relational Charts

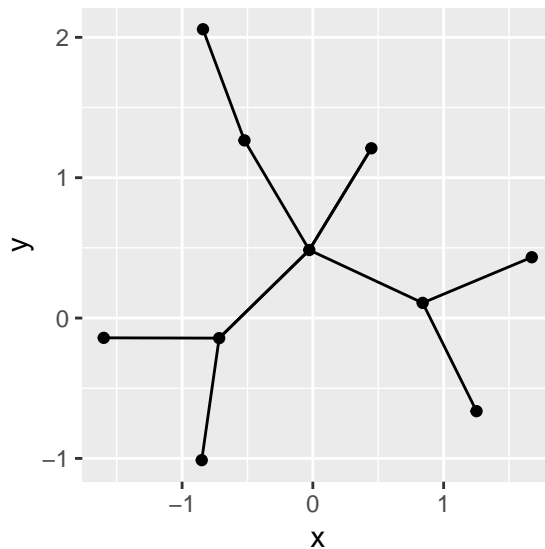
These data do not have network data, so we'll make up so data to use.

```
# Data was found in r-graph-gallery:
links_data <- data.frame(
  source = c("A", "A", "A", "A", "A", "J",
             "B", "B", "C", "C", "D", "I"),
  target = c("B", "B", "C", "D", "J", "A",
             "E", "F", "G", "H", "I", "I")
)

highschool_dat <- ggraph::highschool

node_link <- specify_single(chart_type = "node-link",
                             data = "links_data")

# And plot!
plot(node_link)
```



D.1.4 Spatial Charts

The associated data with the spatial charts is always a bit tricky because there are a lot of ways that a user could bring such data in and it's not possible for the system to catch them all. So, some extra attention needs to be paid for this chart type.

minCombinR does some work for you when you join spatial datasets, so we can use that to add information. At a bare minimum, if you don't just want to draw the polygons of the shape file, but you want to add some information to them, you need to make sure that the data you've loaded in actually contains usable information. A lot of data does not, and it's not possible for minCombinR to fill in those gaps.

```
meta_tmp <- shape_dat@data$metadata
geo <- shape_dat@data$geometry
```

There are two ways to view a geographic map. First, working from shape files:

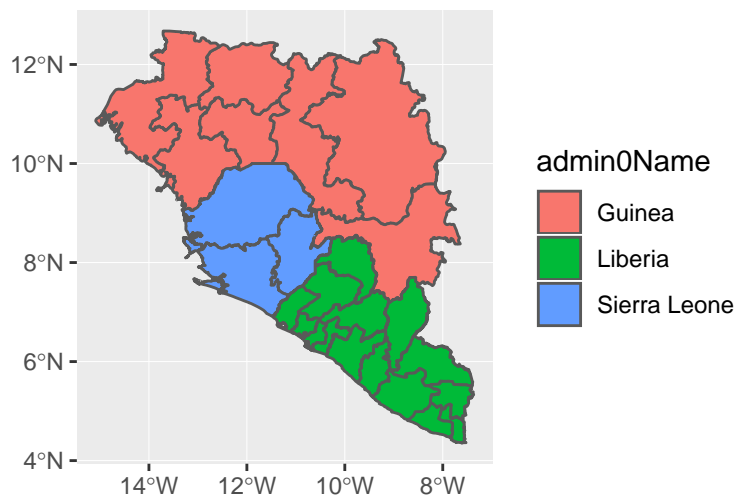
First, let's specify and plot all of the regions in the map together


```

# metadata from the shape files is incomplete
# this is something you need to know and not something
# that the systems resolves for you. So, we'll work to clean
# up the metadata a bit more before we
spatial_chart <- specify_single(chart_type = "choropleth",
                                data = "shape_dat",
                                color = "admin0Name")

plot(spatial_chart)

```



```

# metadata from the shape files is incomplete
# this is something you need to know and not something
# that the systems resolves for you so, we'll work to clean
# up the metadata a bit more before

shape_meta<-shape_dat@data$metadata

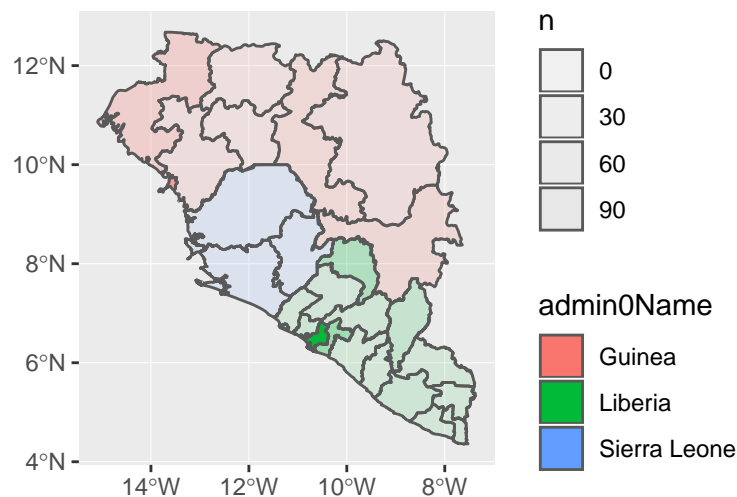
#a little wrangling to put things under the same
#column header since they have slightly different names
tmp<-shape_meta$admin1Name
tmp[is.na(tmp)]<-as.character(shape_meta$admin1name[is.na(tmp)])
shape_meta$admin1Name<-tmp

#now add this with the sample data
tab<-tab_dat@data[[1]] %>% group_by(location) %>% tally()
tab<-left_join(shape_meta,tab,by=c("admin1Name"="location"))
tab[is.na(tab$N),]$N<-0

```

```
#visualize
spatial_chart<-specify_single(chart_type="choropleth",
                              data="shape_dat",
                              metadata="tab",
                              color="admin0Name",
                              alpha="n")

plot(spatial_chart)
```



Also, it possible just to see one individual map too, not just everything together

Here's an example where we compute some new data and specify and plot the "Guinea" region

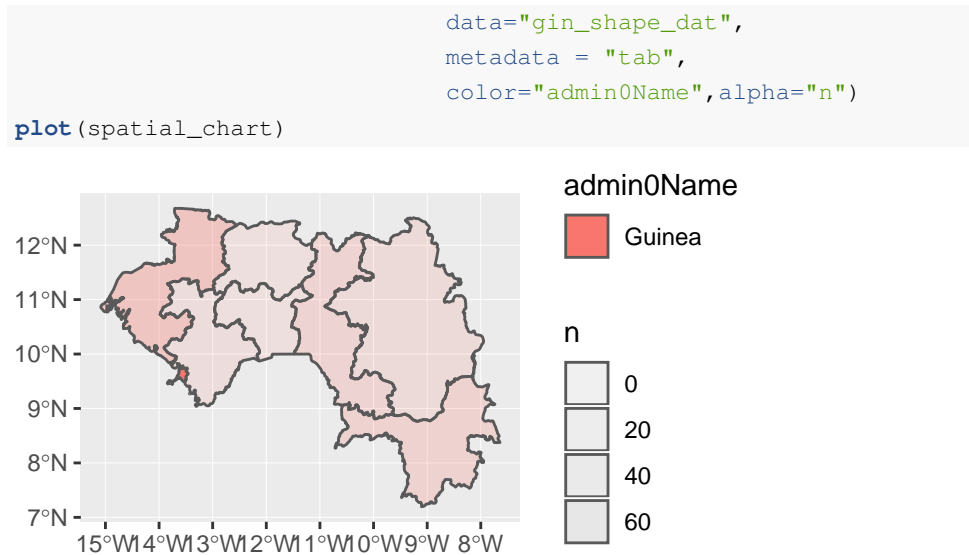
```
# Let's change things up a little bit and plot the incidence

# You can change the metadata
gin_meta <- gin_shape_dat@data$metadata

tab<-tab_dat@data[[1]] %>% group_by(location) %>% tally()

tab<-left_join(gin_meta,tab,by=c("admin1Name"="location"))
tab[is.na(tab$n),]$n<-0

spatial_chart<-specify_single(chart_type="choropleth",
```

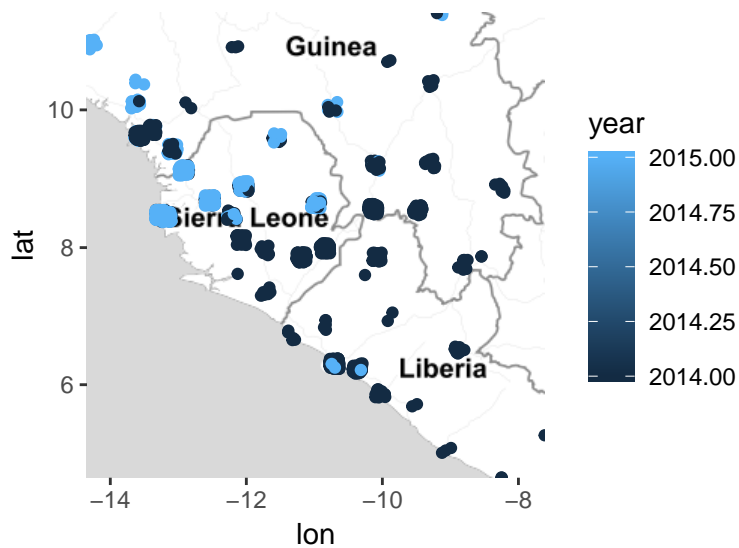


Second, is working from a co-ordinates from a tabular data file. The background raster image is supplied by the ‘Openstreetmaps’ package (and the broader project). They do great work, consider supporting them if you like this feature too.

```

# Geographic Map
map_chart <- specify_single("geographic map",
  data = "tab_dat",
  lat = "latitude",
  long = "longitude",
  color = "year")
plot(map_chart)

```



D.1.5 Tree Charts

```
# Phylogenetic Tree
phyloTree_chart <- specify_single(chart_type = "phylogenetic tree",
                                   data = "tree_dat")
plot(phyloTree_chart)
```

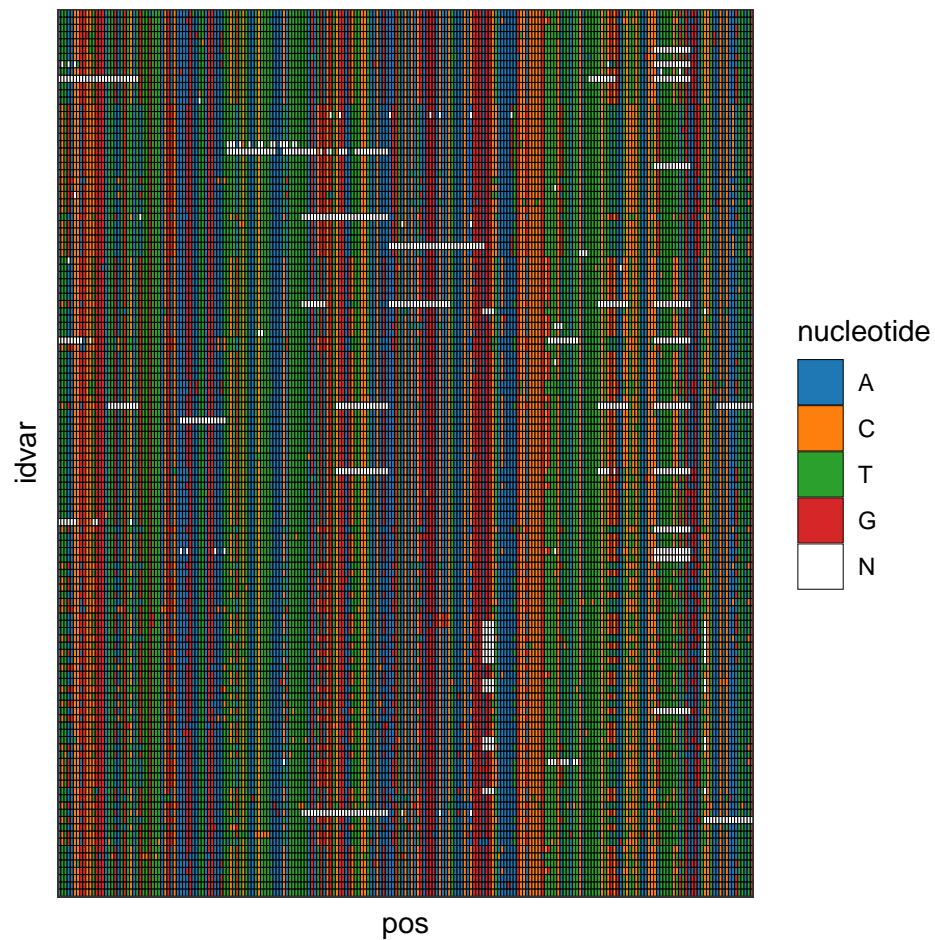


D.1.6 Genomic charts

A very standard alignment chart. This tends to still work if you've got very long sequences. There are some options to simplify the output too.

It's possible to pass a **diff_only** parameter, which will only show variant positions.

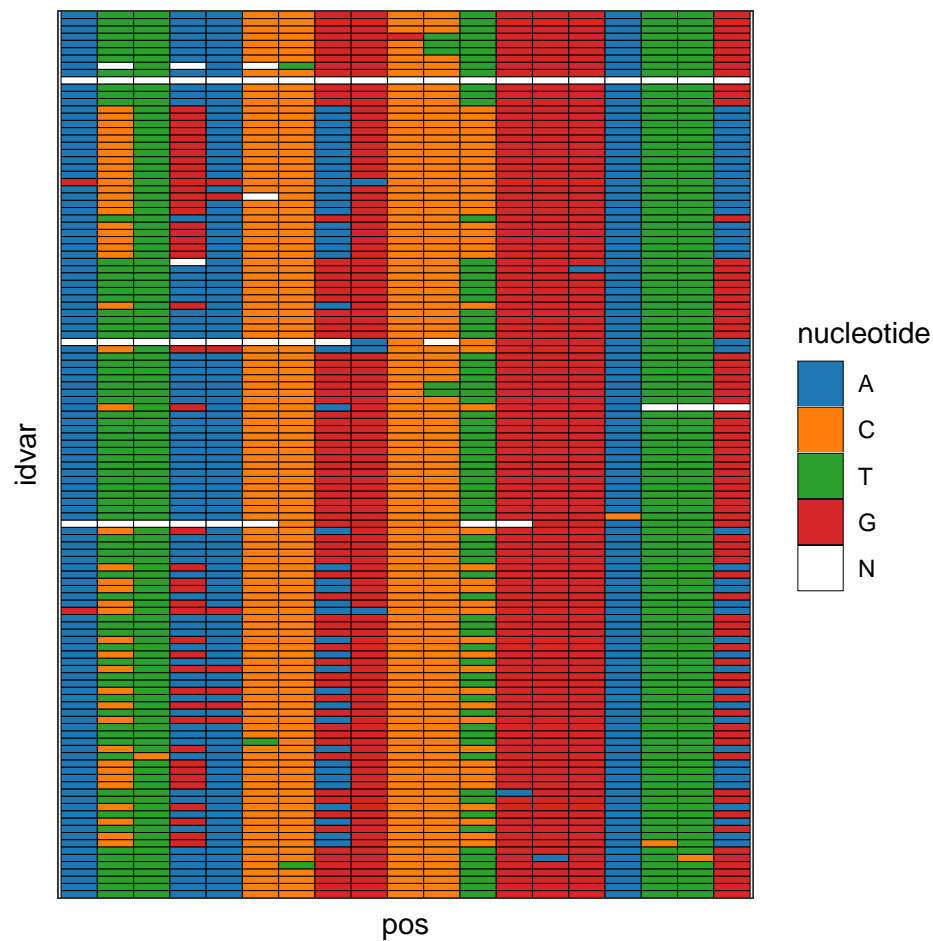
```
# Alignment
genome_chart <- specify_single(data = "genomic_dat",
                              chart_type = "alignment")
plot(genome_chart)
```



We can also just look at the subset of the data, to make life a little bit easier.

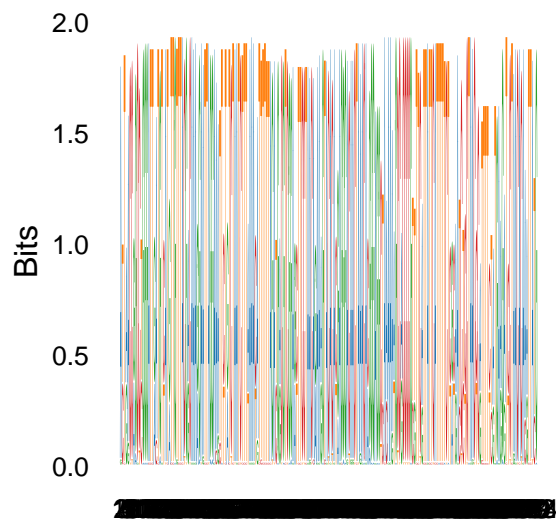
```
# Alignment
diff_seq <- get_diff_pos(genomic_dat)
genome_chart <- specify_single(data = "genomic_dat",
                              chart_type = "alignment",
```

```
show_pos=diff_seq[1:20])
plot(genome_chart)
```



Sequence logo plots are also supported. These work well when there are a small number of variant positions, but when there are many, it's actually not a very nice visual. The user will receive a prompt in those circumstances.

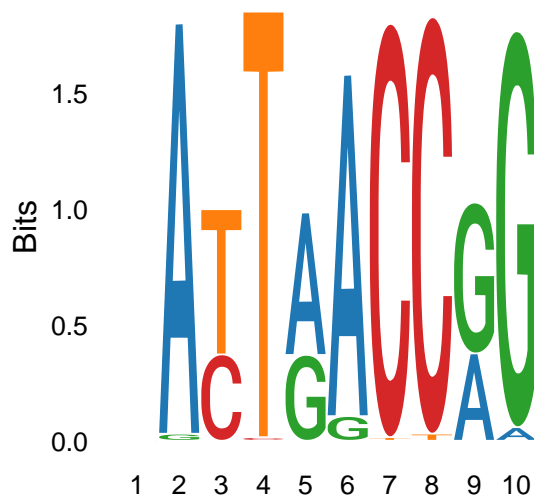
```
seqlogo_chart <- specify_single(data = "genomic_dat",
                                chart_type = "sequence logo")
plot(seqlogo_chart)
```



See? Too many. Instead, it might be better to select a few specific regions to show. In the first case, there will be a warning that nothing is very different and the plot won't show:

```
# A little helper function that extracts the similar sequences
diff_seq <- get_diff_pos(genomic_dat)

# Sequence Logo
seqlogo_chart <- specify_single(data = "genomic_dat",
                                chart_type = "sequence logo",
                                show_pos = diff_seq[1:10])
plot(seqlogo_chart)
```



D.1.7 Temporal Charts

Sometimes data contains start and end dates and its desirable to show these as different types of temporal charts.

There are two times of timelines that can be created:

1. A “gantt chart” type timeline that will show both ranges and single events.
2. A standard epidemic curve, which is essentially a very special case of a histogram or bar chart, depending upon the user.

```
# Using the existing tabular data:
tmp <- tab_dat@data[[1]]
tmp$collection_date <- as.Date(tmp$collection_date)

# Let's add some end dates to keep it interesting:
tmp$collection_date_end <- tmp$collection_date +
  sample(10:30, nrow(tmp), replace = TRUE)

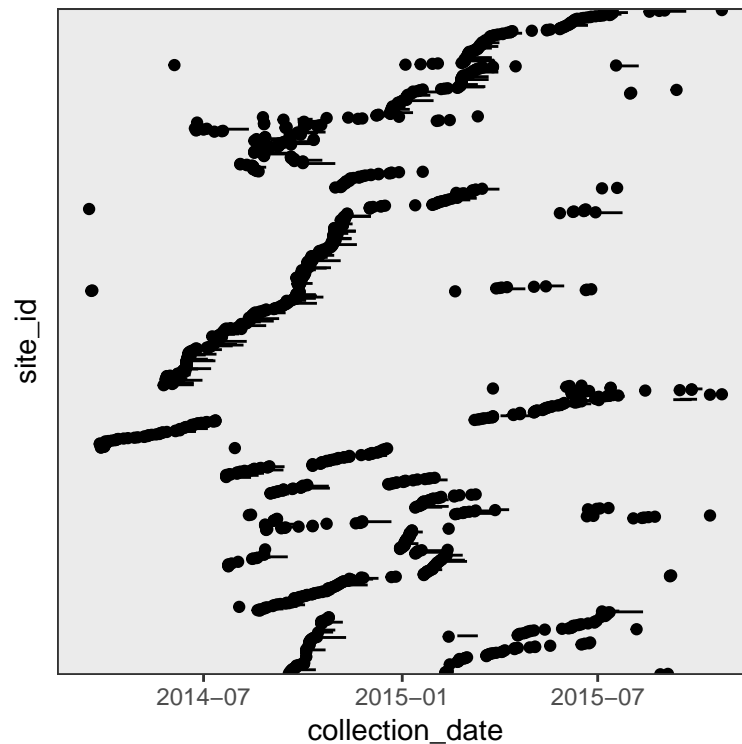
tmp$collection_date_end <- sapply(
  as.character(tmp$collection_date_end),
  function(x) {
    if(runif(1) > 0.9)
      return(x)
    return(NA)
  })
```



```

timeline_chart <- specify_single(chart_type = "timeline",
                                data = "tmp",
                                start = "collection_date",
                                end = "collection_date_end",
                                y = "site_id")
plot(timeline_chart)

```

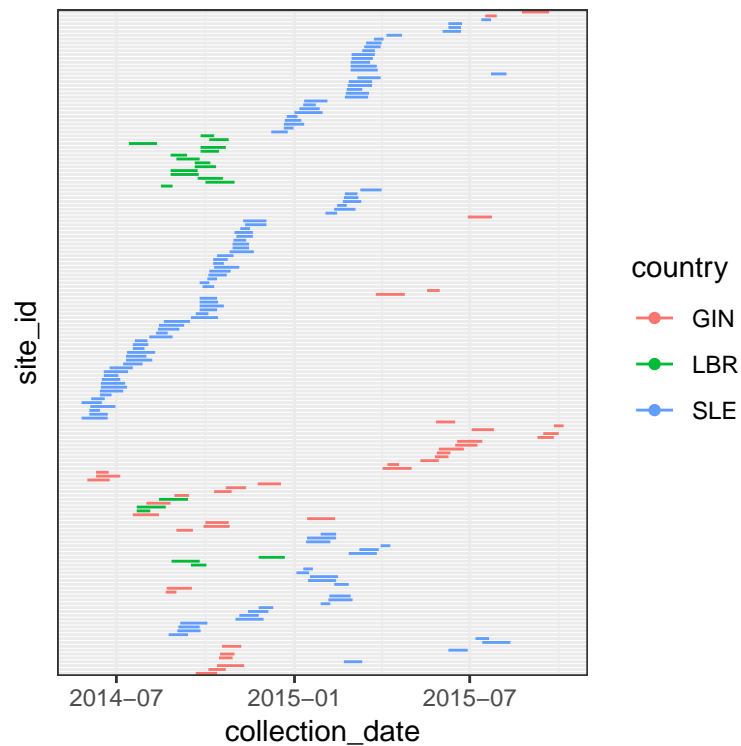


Since there's a lot going on, let's subset this to get a better view. So here, only plotting those items that are ranges.

```

tmp_sub <- dplyr::filter(tmp, !is.na(collection_date_end))
timeline_chart <- specify_single(chart_type = "timeline",
                                data = "tmp_sub",
                                start = "collection_date",
                                end = "collection_date_end",
                                y = "site_id",
                                color = "country")
plot(timeline_chart)

```



D.1.8 Images

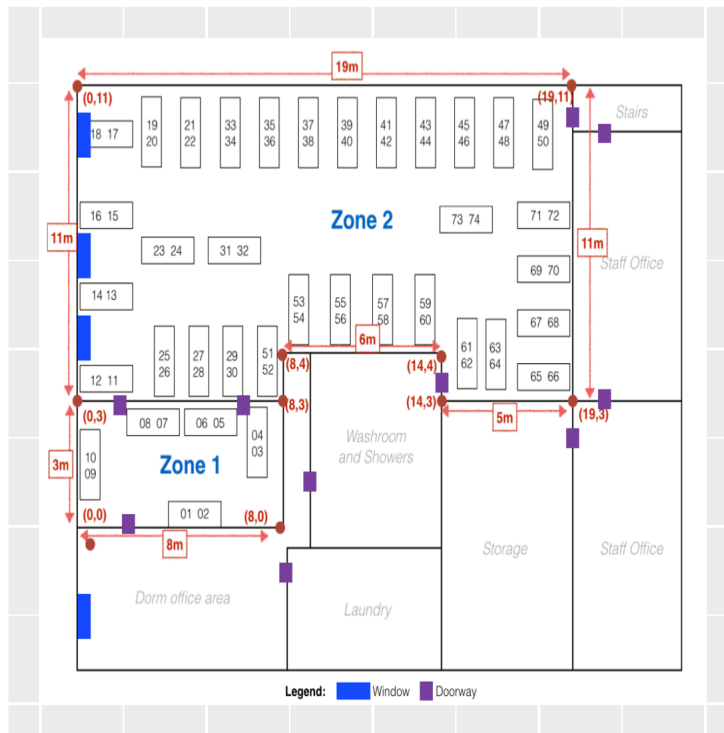
Sometimes it's desirable to use image data, these can be from interior maps or from gel images. Getting image data to work is a little bit more involved because it requires some annotation data that allows you to link pixel space to something useful. There's a small application embedded within minCombinR that lets you do that.

Here's a workflow with a few different types of images:

```
#Load in interior map image data into minCombinR's input_data fcn
interior_img <- input_data(file =
  system.file("extdata",
    "random_interior_map.tiff", package =
    "mincombinr"),
  dataType = "image")
```

It's still possible to simply draw a plot that has no metadata

```
img_chart <- specify_single(chart_type = "image",
                             data = "interior_img")
plot(img_chart)
```



But it's nice to do something more useful with a picture. The error you saw when loading the data is to remind the user that there needs to be some image file loaded. To add metadata, we'll run the special app. Note that in building this markdown file, this block of code is not run, however, **should** be done by the user if they would like to add some metadata.

```
interior_img <- annotate_image(interior_img)

# The annotations are automatically there now:
metadata <- interior_img@data$metadata

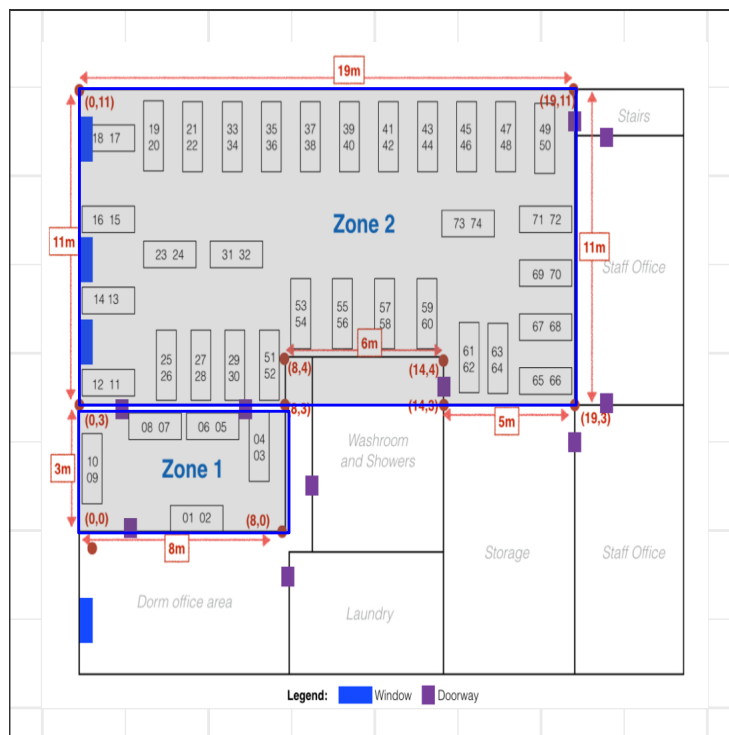
save(file = "../inst/extdata/img_meta.rds", metadata)
```

```

# We'll load an already annotated image to keep things going
# Normally, the user doesn't do this, but we have to do so because
# the previous line of code was not run in the markdown notebook
load(file = system.file("extdata", "img_meta.rds",
                        package = "mincombinr"))
interior_img@data$metadata<-metadata

# Now specify and plot the interior map
img_chart <- specify_single(chart_type = "image",
                           data = "interior_img")
plot(img_chart)

```



Let's do something more interesting and color these regions:

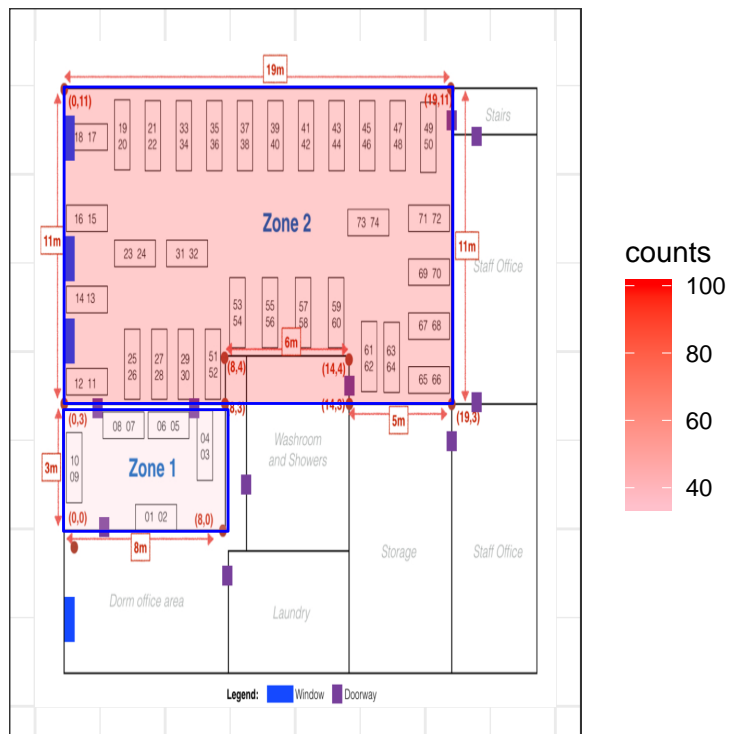
```

# Let's add some arbitrary case counts to the room
interior_img@data$metadata$counts <- c(100,35)

# Now specify and plot the interior map
img_chart <- specify_single(chart_type = "image",

```

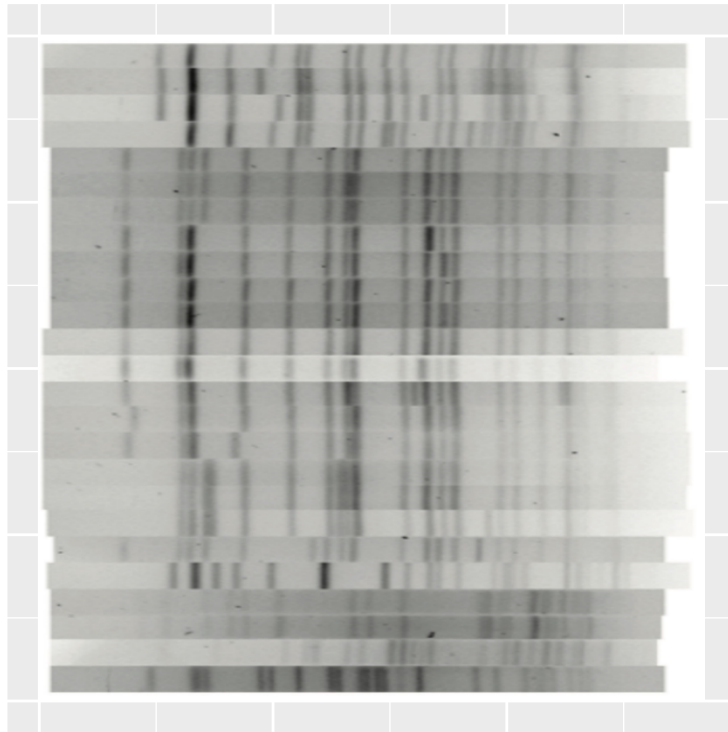
```
data = "interior_img",
color = "counts")
plot(img_chart)
```



The other most common type of image is a gel image - this too can now be part of the analysis fun.

```
gel_img <- input_data(file = system.file("extdata",
                                           "gel_image.tiff",
                                           package = "mincombinr"),
                      dataType = "image")

gel_img_chart <- specify_single(chart_type = "image",
                                data = "gel_img")
plot(gel_img_chart)
```



Just like the interior map, this block of code is not run by the notebook but should be used to annotate an image as a user.

```
gel_img <- annotate_image(gel_img)
metadata <- gel_img@data$metadata
metadata$element_name <- paste("item", 1:nrow(metadata), sep = "_")

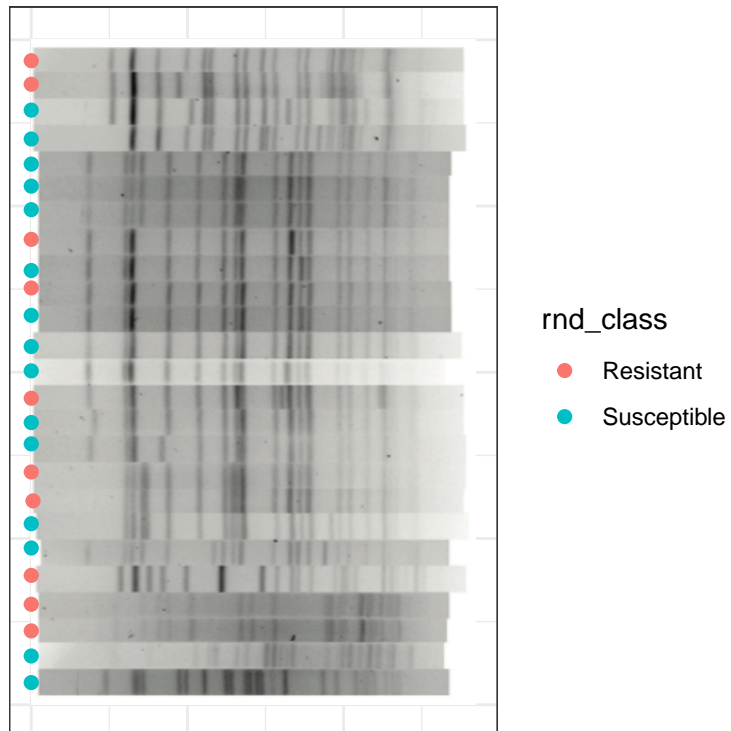
save(metadata, file = "../inst/extdata/gel_img_meta.rds")

# Just like the interior map, we will load in an
# annotated image for the sake of demonstration:
load(file = system.file("extdata",
  "gel_img_meta_shipped.rds",
  package = "mincombinr"))
gel_img@data$metadata <- metadata

gel_img@data$metadata$rnd_class <- sample(c("Resistant",
  "Susceptible"),
  replace = TRUE, size = nrow(metadata))
```

```
#Now we specify and plot the gel image
gel_chart <- specify_single(data = "gel_img",
                             chart_type = "image",
                             color = "rnd_class")

plot(gel_chart)
```



D.2 Generating Combinations of Charts with minCombinR

First, let's store the specifications for a few charts that we want to combine together

```
# Bar chart
bar_chart <- specify_single(chart_type = "bar",
                             data = "tab_dat",
                             x = "country")
```

```

# Phylogenetic Tree
phyloTree_chart <- specify_single(chart_type = "phylogenetic tree",
                                  data = "tree_dat")

# Scatter Plot
scatter_chart <- specify_single(chart_type = "scatter",
                                 data = "tab_dat",
                                 x = "latitude",
                                 y = "longitude",
                                 title = "Ebola Scatter Plot")

# Geographic Map
map_chart <- specify_single("geographic map",
                             data = "tab_dat",
                             lat = "latitude",
                             long = "longitude")

```

D.2.1 Unaligned

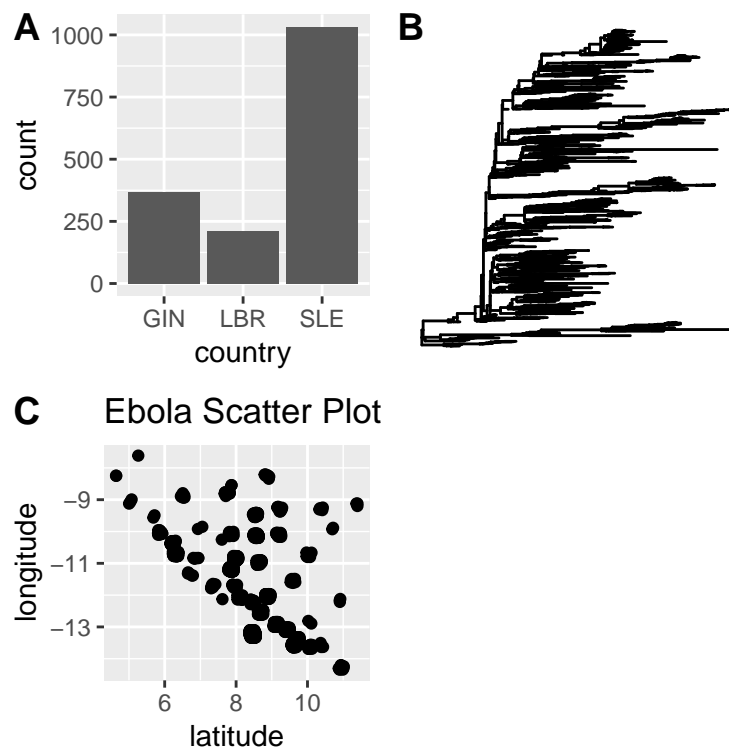
Unaligned combinations can be used when you just want to put a bunch of plots together and there are no spatial or visual linkages between the plots themselves.

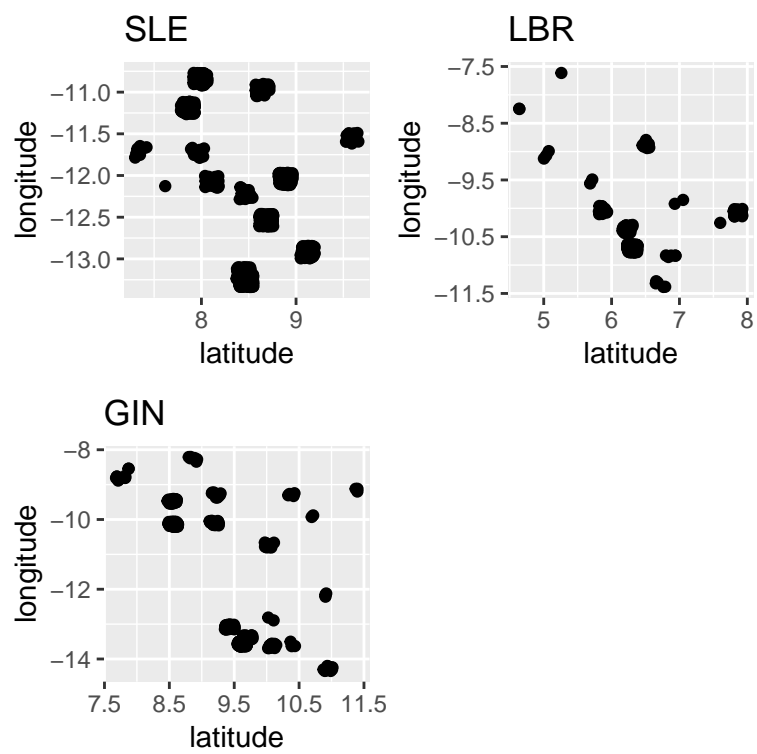
```

# Specify that you want to combine the bar_chart,
# phyloTree_chart and scatter_chart
mg_combo <- specify_combination(combo_type = "unaligned",
                                base_charts = c("bar_chart",
                                                  "phyloTree_chart",
                                                  "scatter_chart"))

# Now plot it!
plot(mg_combo)

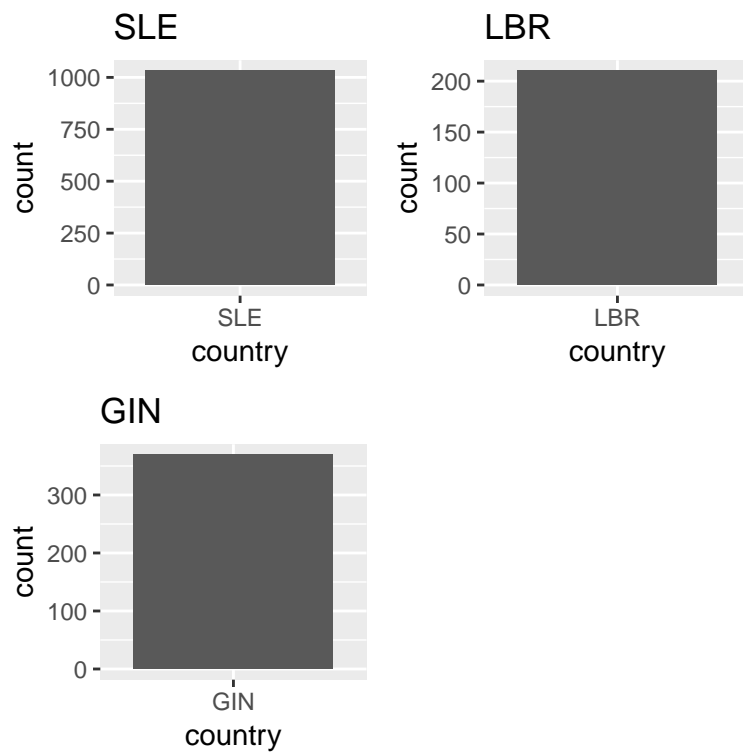
```



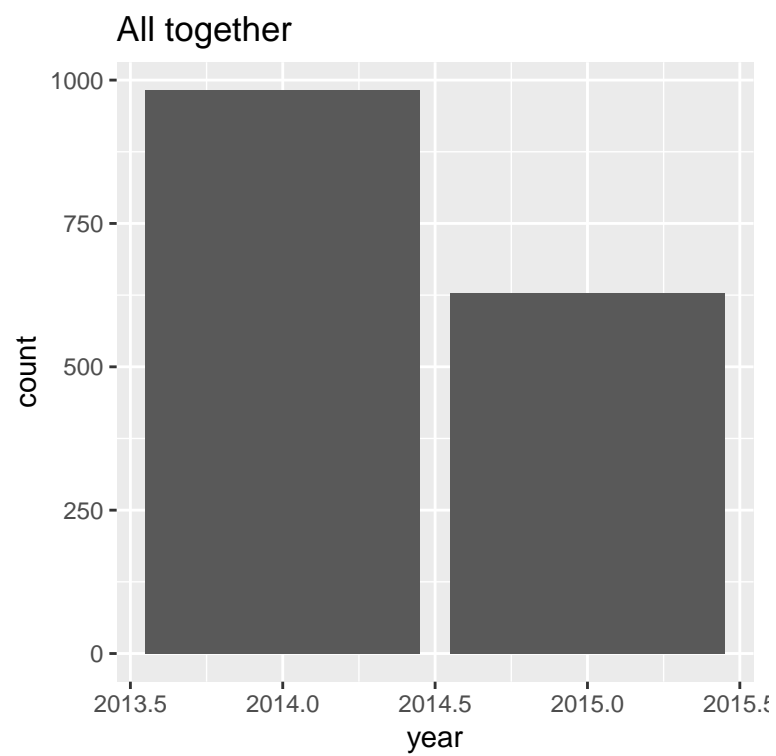
```
sm_combo_bar <- specify_combination(combo_type = "small_multiple",
                                     base_charts = "bar_chart",
                                     facet_by = "country")

plot(sm_combo_bar)
```

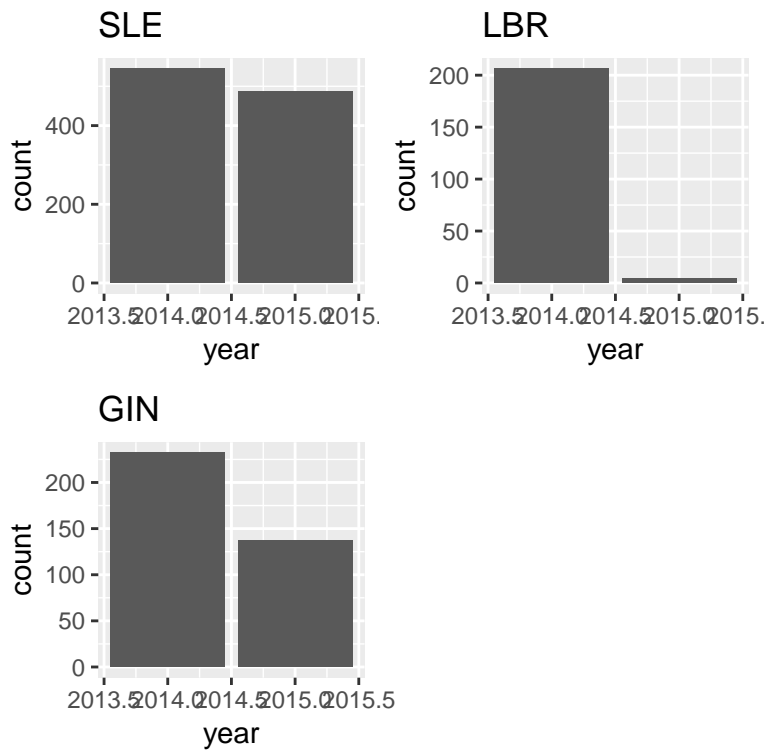


```
# Let's try a more interesting bar_chart small multiple
bar_chart_alt <- specify_single(chart_type = "bar",
                                data = "tab_dat", x = "year",
                                title = "All together")

sm_combo_bar_alt<- specify_combination(
  combo_type= "small_multiple",
  base_charts= "bar_chart_alt",
  facet_by= "country")
plot(bar_chart_alt)
```



```
plot(sm_combo_bar_alt)
```



Other chart types cannot be easily subsetted.

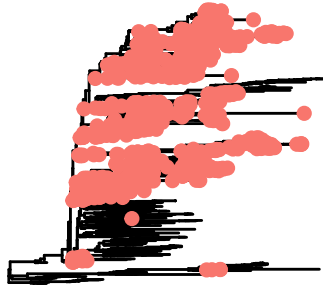
For example, it would not be meaningful to make a small multiple of a phylogenetic tree while only showing a subset of the tree. The same is generally true for maps, and networks.

Although there are ways to truly subset them, it's messy and the whole underlying structure matters, so minCombinR will give you the whole network structure.

In the current implementation of minCombinR, there needs to be some tabular data associated with non-tabular data in order to understand what should be visualized in the first place.

```
# Tree data
tree_dat <- input_data(file = system.file("extdata",
                                           "ebov_tree.nwk",
                                           package = "mincombinr"),
                      dataType = "tree")
```


SLE



LBR



GIN



D.2.3 Colour Aligned Combinations

Finally, it could be interesting to link several different chart types together by color.

In the above examples, we may want to link the phylogenetic tree with the timeline by their countries.

For the non-tabular data, it's important to have some associated metadata, otherwise, it is not possible to link information. It is up to the user to establish that two variables are actually linkable by the same variable. Some of the code from the spatial aligned combination is borrowed to see if two datasets are even linkable to help with the color linkage.

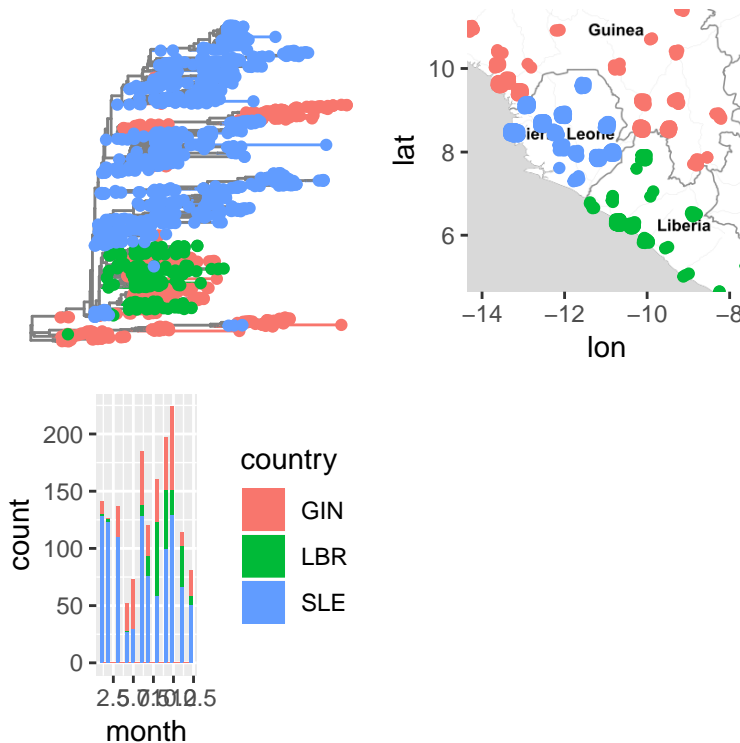
First scenario - no metadata provided for the tree

minCombinR will try to find if there are linkages between the tabular data in one chart and non-tabular data in another chart. If there are EXACT MATCHES, it will link the tabular data to the tree's metadata

```
# Specify the phylogenetic tree and histogram
phyloTree_chart <- specify_single(chart_type = "phylogenetic tree",
                                  data = "tree_dat")
epicurve <- specify_single(chart_type = "histogram",
                            data = "tab_dat",
                            x = "month")

# Specify that you want to combine with color
color_combo <- specify_combination(combo_type = "color_aligned",
                                   base_charts= c("phyloTree_chart",
                                                  "map_chart",
                                                  "epicurve"),
                                   link_by = "country")

# Now plot!
plot(color_combo)
```



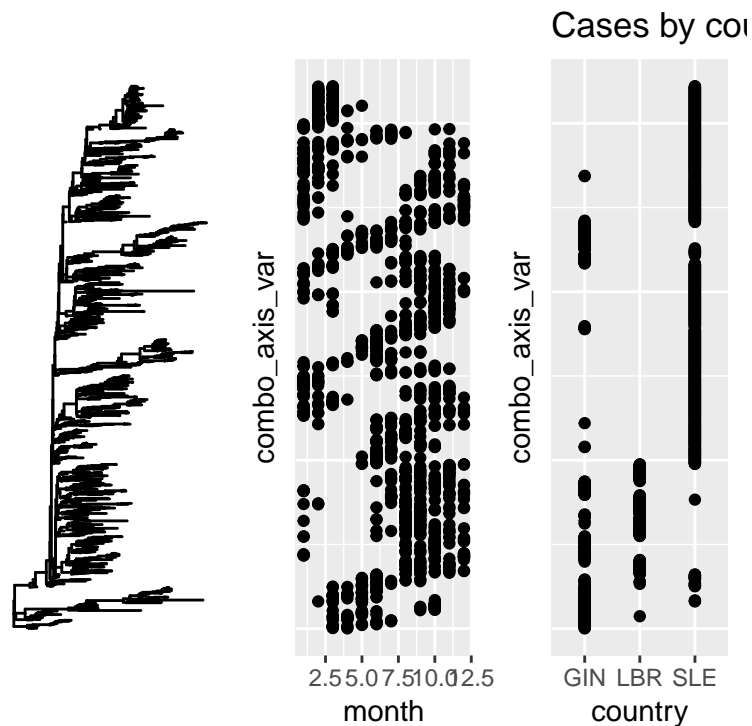
D.2.4 Spatially Aligned Combinations

Spatially aligned combinations line up charts so that they can be read across horizontally and vertically. Critically, this isn't an arbitrary concatenation of charts, but rather, charts are aligned so that the same data are read across.

```
scatter_chart <- specify_single(chart_type = "scatter",
                                data = "tab_dat",
                                x = "month",
                                y = "site_id")
scatter_chart_two <- specify_single(chart_type = "scatter",
                                     data = "tab_dat",
                                     x = "country",
                                     y = "site_id",
                                     title = "Cases by country")

spatial_aligned_combo <- specify_combination(
  combo_type = "spatial_aligned",
  base_charts = c("phyloTree_chart",
                  "scatter_chart",
                  "scatter_chart_two"))

plot(spatial_aligned_combo)
```



This example produce an error (and this is the right thing for it to do) because pie charts cannot be part of spatially aligned combinations. Pie chart will be automatically removed from the specifications with a warning message to the user

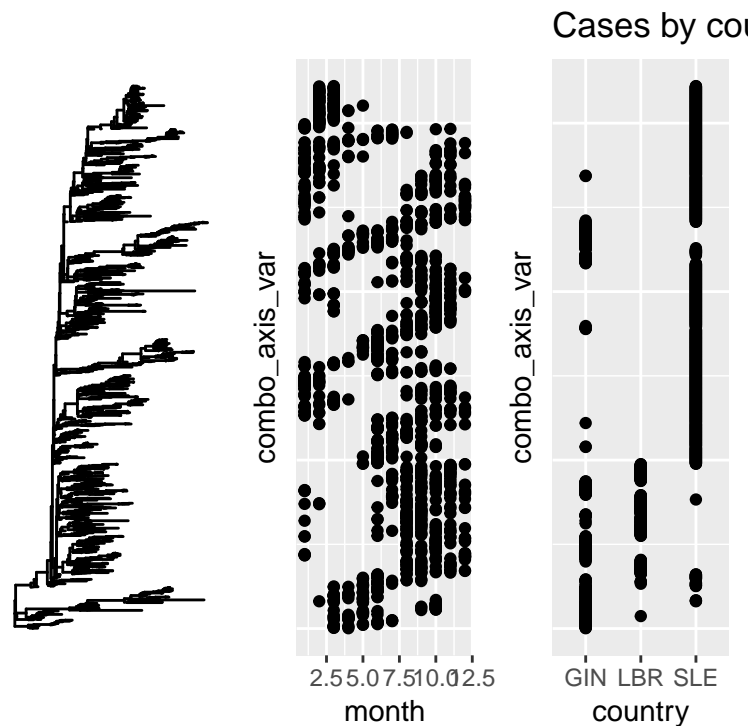
```
# Specifications
pie_chart <- specify_single(chart_type = "pie",
                             data = "tab_dat",
                             x = "country")

spatial_aligned_combo<-specify_combination(
  combo_type = "spatial_aligned",
  base_charts = c("phyloTree_chart",
                  "scatter_chart",
                  "scatter_chart_two",
                  "pie_chart"))
```

```
## [1] "The following chart types cannot form a composite"
```

combination: pie_chart. Composite combination will be formed with the following charts only: phyloTree_chart, scatter_chart, scatter_chart_two"

```
# Plot
plot(spatial_aligned_combo)
```

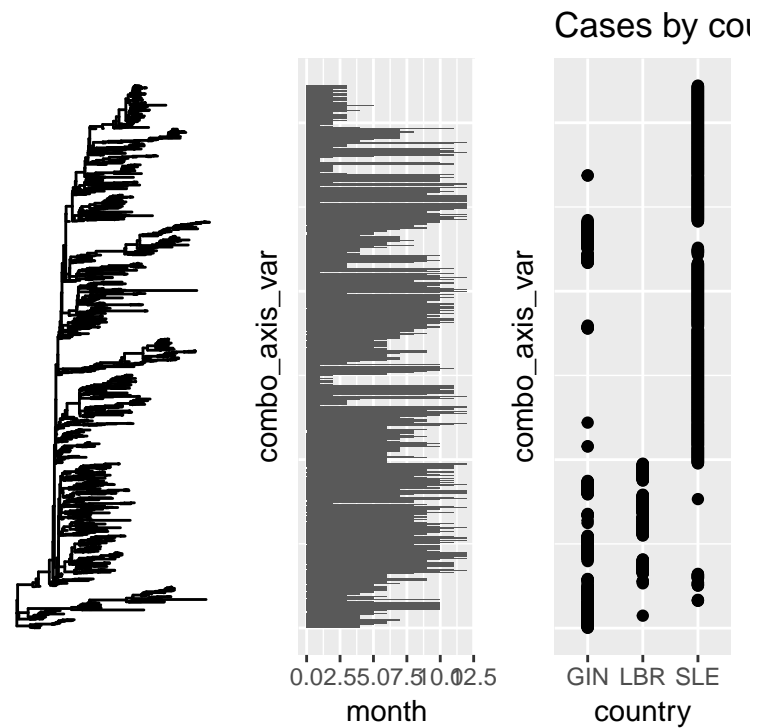


Let's try combining a phylogenetic tree, bar chart, and scatter chart

```
# Specifications
bar_alt <- specify_single(chart_type = "bar", data = "tab_dat",
  x = "site_id",
  y = "month",
  rm_x_label=TRUE)

spatial_aligned_combo <- specify_combination
  (combo_type = "spatial_aligned",
  base_charts = c("phyloTree_chart",
    "bar_alt",
    "scatter_chart_two"))
```

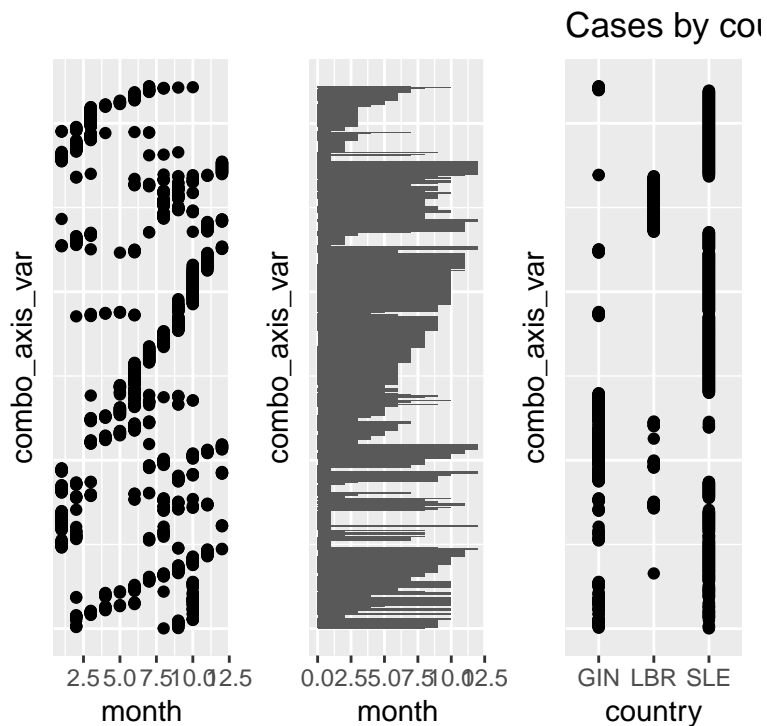
```
# Plot
plot(spatial_aligned_combo)
```



It also works when there isn't a tree involved

```
spatial_aligned_combo<- specify_combination(
  combo_type = "spatial_aligned",
  base_charts = c("bar_alt",
    "scatter_chart",
    "scatter_chart_two"))

plot(spatial_aligned_combo)
```



Here's a combination with a genomic map

```
# Genomic chart
# For illustrative purposes, show fewer positions
diff_seq <- get_diff_pos(genomic_dat)
genome_chart <- specify_single(data = "genomic_dat",
                                chart_type = "alignment",
                                title="Genome Alignment",
                                show_pos=diff_seq[1:20])

# Timeline, with some fake end_dates and using
# the existing tabular data
time_dat <- tab_dat@data[[1]]
time_dat$collection_date <- as.Date(time_dat$collection_date)

# Let's add some end dates to keep it interesting
time_dat$collection_date_end <- time_dat$collection_date +
  sample(10:30, nrow(time_dat), replace = TRUE)
```

```

time_dat$collection_date_end <- sapply(as.character(
  time_dat$collection_date_end
),
function(x){
  if(runif(1)>0.9)
    return(x)
  return(NA)
})

time_dat <- dplyr::filter(time_dat,!is.na(collection_date_end))

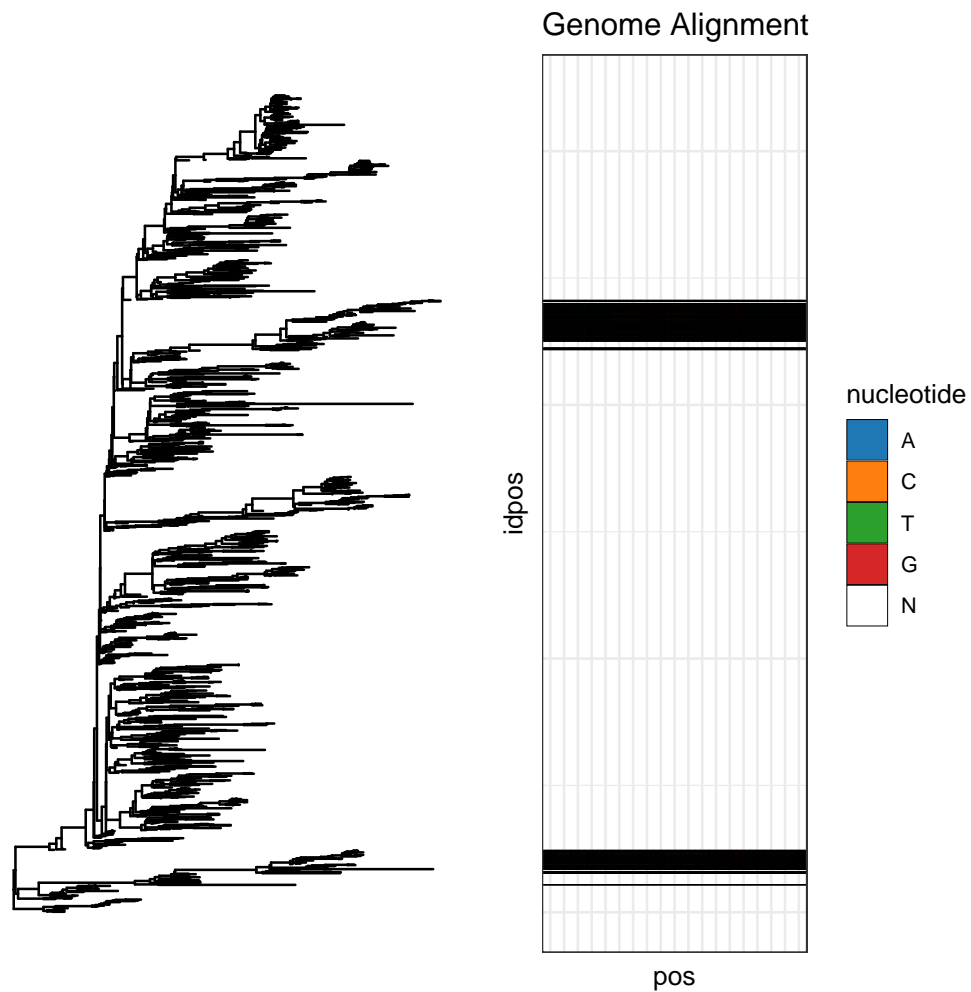
# Specifications
timeline_chart <- specify_single(chart_type = "timeline",
  data="time_dat",
  start = "collection_date",
  end = "collection_date_end",
  y = "site_id",
  title="Timeline")

spatial_aligned_combo <- specify_combination(
  combo_type = "spatial_aligned",
  base_charts = c("phyloTree_chart",
    "scatter_chart_two",
    "timeline_chart",
    "genome_chart"))

spatial_aligned_combo <- specify_combination(
  combo_type= "spatial_aligned",
  base_charts= c("phyloTree_chart",
    "genome_chart"))

plot(spatial_aligned_combo)

```



Note that in the above case, genomic data is not available for all of the items in the phylogenetic tree. In fact, our source data is for Guinea only. The composite algorithm is able to adjust in instances where one dataset is a perfect subset of the other. It is up to the user to ensure this.

Appendix E

GEViTRec Supplemental Materials

```
start_time <- Sys.time()
set.seed(416)

library(dplyr)
library(ggplot2)
library(igraph)
library(gggraph)

# for displaying the resulting and loading the data
library(mincombinr)

#loading gevitrec
devtools::load_all()

#Table data
tab_dat<-input_data(file = system.file("../inst/extdata/",
                                         "ebov_metadata.csv",
                                         package = "gevitRec"),
                    dataType = "table")

#Tree data
```



```

tree_dat<-input_data(file = system.file("./inst/extdata/",
                                         "ebov_tree.nwk",
                                         package = "gevitRec"),
                    dataType = "tree")

#Genomic data
genomic_dat<-input_data(file = system.file("./inst/extdata/",
                                             "ebov_GIN_genomic.fasta",
                                             package = "gevitRec"),
                        dataType = "dna")

#Shape files
#Shape files require that .shp,.shx,and .prj
# files at a minimum to be in the same directory
#to add metadata to the shape file, you can also add .dbf files
gin_file<-"gin_admbnda_adml_ocha_itos.shp"
lbr_file<-"lbr_admbnda_adml_ocha.shp"
sle_file<-"sle_admbnda_adml_lm_gov_ocha_20161017.shp"

gin_shape_dat<-input_data(file =
                          system.file("./inst/extdata/",
                                       gin_file,
                                       package = "gevitRec"),
                          dataType = "spatial")
lbr_shape_dat<-input_data(file =
                          system.file("./inst/extdata/",
                                       lbr_file,
                                       package = "gevitRec"),
                          dataType = "spatial")
sle_shape_dat<-input_data(file =
                          system.file("extdata/",
                                       sle_file,
                                       package = "gevitRec")
                          ,dataType = "spatial")

#clean up the metadata a bit and make it a little more interesting
tmp<-all_spatial@data$metadata
idx_missing<-which(is.na(tmp$admin1Name))

tmp[idx_missing,]$admin1Name<-as.character(tmp[idx_missing,]$admin1name)

```

```

idx_drop<-apply(tmp,2,function(x){all(is.na(x))})

tmp<-tmp[,!idx_drop]

#now add some counts, because I can..
case_counts<-tab_dat@data$table %>%
  dplyr::group_by(country,location) %>%
  tally() %>%
  mutate(case_count = n)

colnames(case_counts) <- c("minID", "admin1Name", "n", "case_count")

tmp<-left_join(tmp, case_counts[,c(1,2,4)])

tmp[is.na(tmp$case_count),]$case_count<-0

tmp<-dplyr::select(tmp, admin1Name,
                  case_count, admin0Name,
                  minPolyID, minID)

all_spatial@data$metadata<-tmp

```

E.1 Data Harmonization

```

harmon_obj<-data_harmonization(tab_dat, tree_dat,
                              genomic_dat, all_spatial)

#plotting the entity graph
view_entity_graph(harmon_obj[["entityGraph"]])

```

E.2 Generate Specifications

```

component_specs<-get_spec_list(harmon_obj)

```

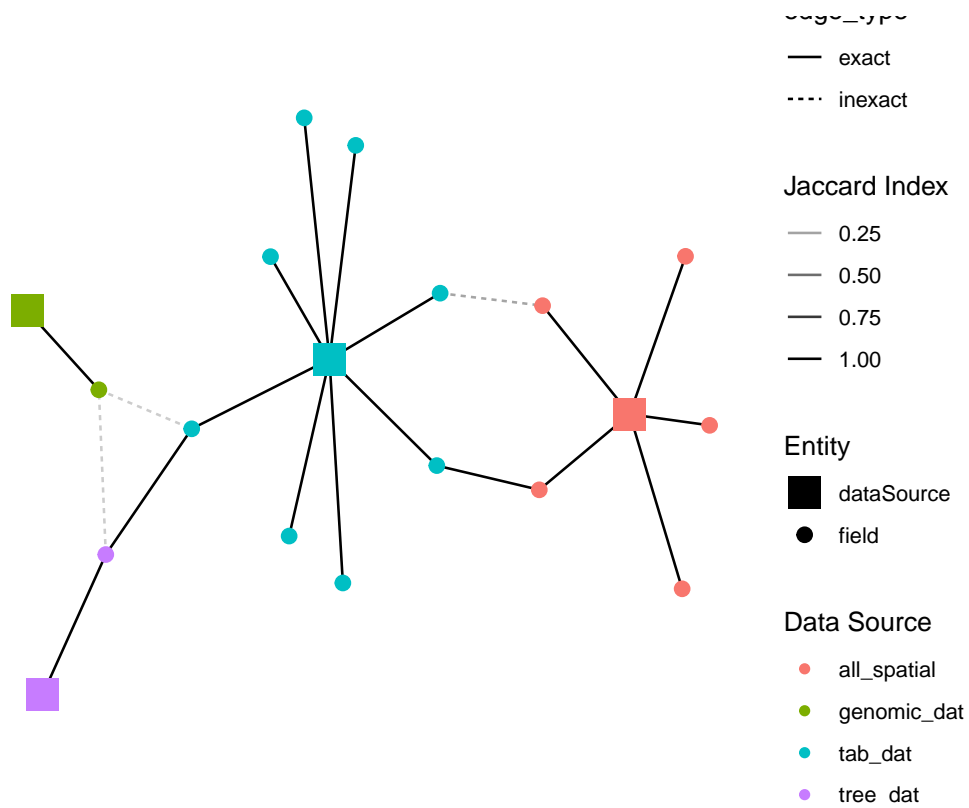


Figure E.1: GEViTRec entity graph

E.3 Generated Views

```
plot_view(component_specs,view_num=1)
```

```
plot_view(component_specs,view_num=2)
```

```
plot_view(component_specs,view_num=3)
```

```
plot_view(component_specs,view_num=4)
```

```
plot_view(component_specs,view_num=5)
```

```
end_time <- Sys.time()
print(end_time - start_time)
```

```
## Time difference of 17.36386 secs
```

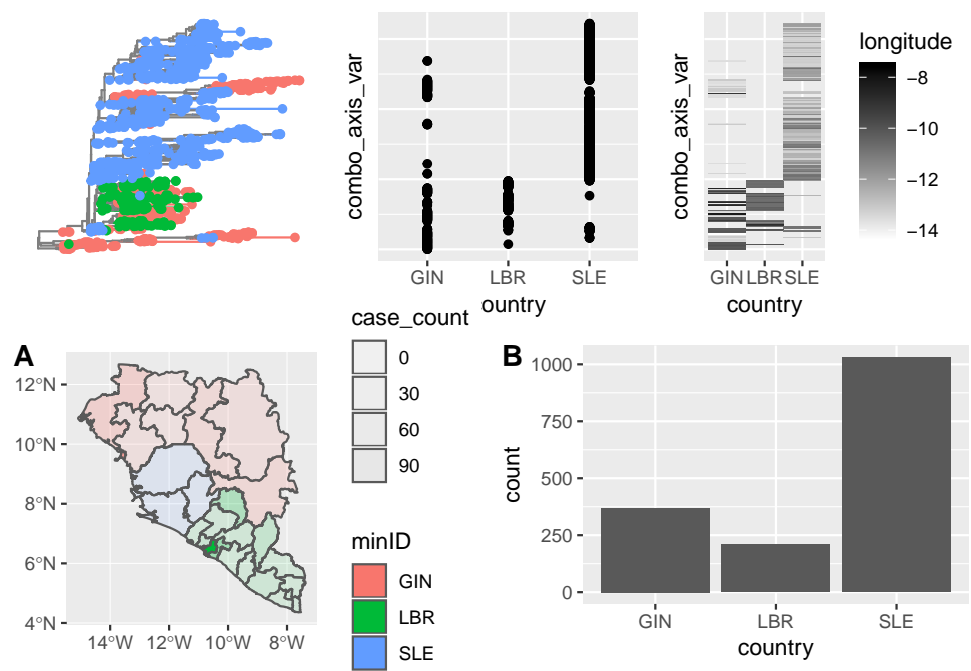


Figure E.2: GEViTRec generated view #1

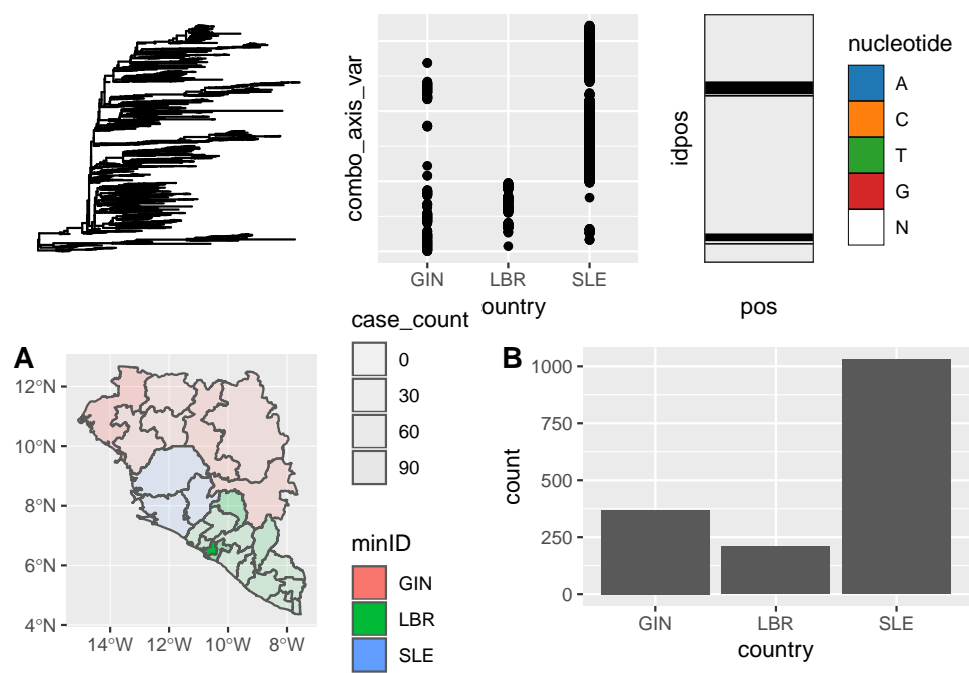


Figure E.3: GEViTRec generated view #2

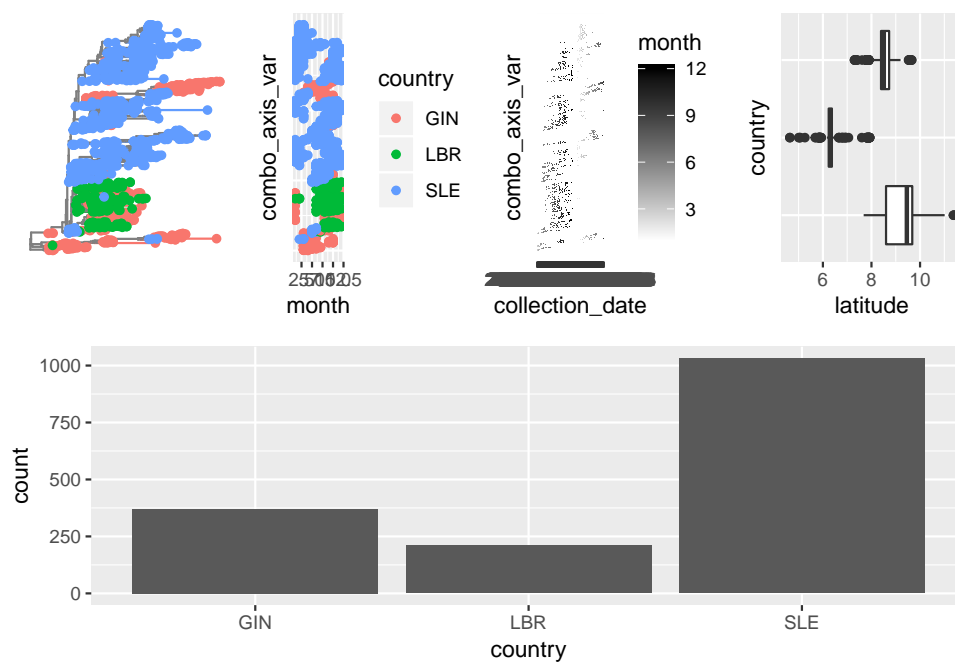


Figure E.4: GEViTRec generated view #3

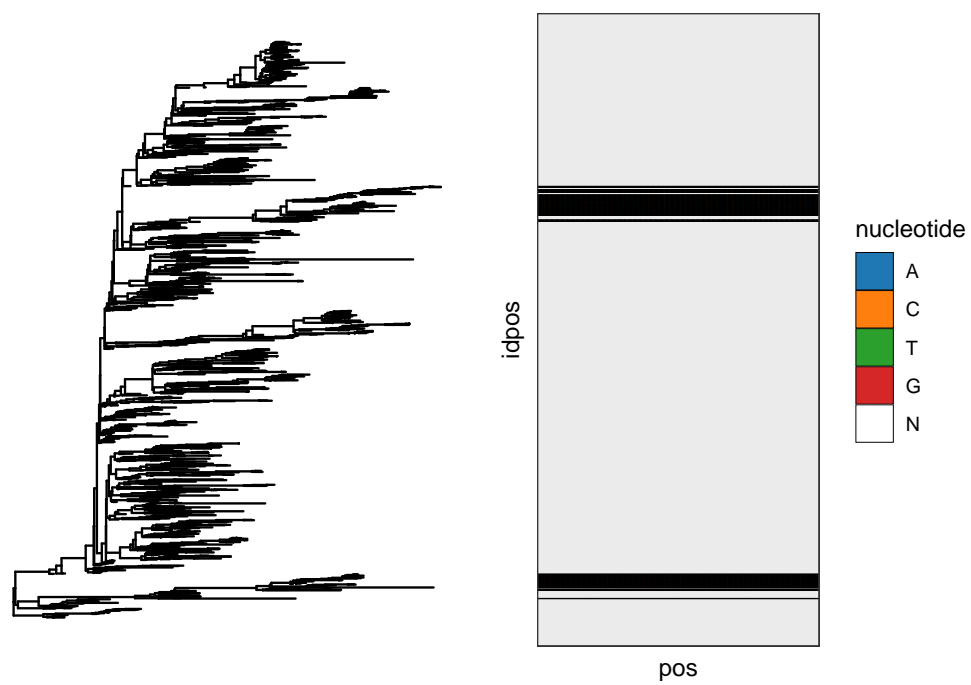


Figure E.5: GEViTRec generated view #4

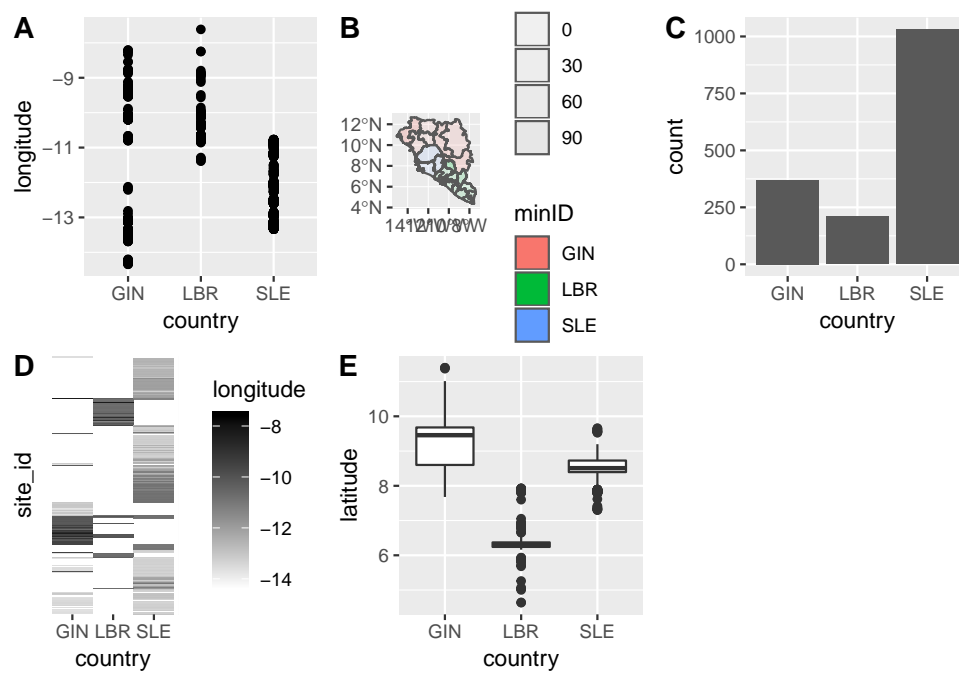


Figure E.6: GEViTRec generated view #5